SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 1–84, January 1994

ON NONSTATIONARY MOTION OF A COMPRESSIBLE BAROTROPIC VISCOUS CAPILLARY FLUID BOUNDED BY A FREE SURFACE*

W. M. ZAJACZKOWSKI†

Abstract. The author considers the motion of a viscous compressible barotropic fluid in $\mathbb{R}^3$, bounded by a free surface that is under surface tension and constant exterior pressure. Assuming the initial density is sufficiently close to a constant, the initial domain is sufficiently close to a ball, the initial velocity is sufficiently small, and the external force vanishes, the existence of a global-in-time solution is proven, which satisfies, at any moment of time, the properties prescribed at the initial moment.

Key words. free boundary, compressible barotropic viscous fluid, global existence, surface tension, anisotropic Sobolev spaces

AMS subject classifications. 35R35, 35A05, 76N10

1. Introduction.

We consider the motion of a viscous compressible barotropic fluid in a bounded domain $\Omega_t \subset \mathbb{R}^3$, which depends on time $t \in \mathbb{R}^1_+$. The free boundary S_t of Ω_t is governed by the surface tension. Let $v = v(x,t)$ be the velocity of the fluid, $\rho = \rho(x,t)$ the density, $f = f(x,t)$ the external force field per unit mass, $p = p(\rho)$ the pressure, μ and ν the viscosity coefficients, σ the surface tension coefficient, and p_0 the external (constant) pressure. Then the problem is described by the following system (see [7, Chaps. 1, 2, 7]):

$$\rho(v_t + v\cdot\nabla v) + \nabla p(\rho) - \mu\Delta v - \nu\nabla \mathrm{div} v = \rho f \quad \text{in } \tilde{\Omega}^T, \tag{1.1a}$$

$$\rho_t + \mathrm{div}(\rho v) = 0 \quad \text{in } \tilde{\Omega}^T, \tag{1.1b}$$

$$\rho|_{t=0} = \rho_0, \qquad v|_{t=0} = v_0 \quad \text{in } \Omega, \tag{1.1c}$$

$$\mathbb{T}\bar{n} - \sigma H\bar{n} = -p_0\bar{n} \quad \text{on } \tilde{S}^T, \tag{1.1d}$$

$$v\cdot\bar{n} = -\phi_t/|\nabla\phi| \quad \text{on } \tilde{S}^T, \tag{1.1e}$$

where $\phi(x,t) = 0$ describes S_t, $\tilde{\Omega}^T = \bigcup_{t\in(0,T)} \Omega_t \times \{t\}$, Ω_t is the domain of the drop at time t, $\Omega_0 = \Omega$ is its initial domain, $\tilde{S}^T = \bigcup_{t\in(0,T)} S_t \times \{t\}$, $\bar{n}$ is the unit outward vector normal to the boundary ($\bar{n} = \nabla\phi/|\nabla\phi|$), and μ, ν, σ are constant coefficients. Moreover, thermodynamic considerations imply $\nu \geq 1/3\mu > 0$, $\sigma > 0$. The last condition (1.1e) means that the free boundary S_t is built of moving fluid particles. Finally, $\mathbb{T} = \mathbb{T}(v,p)$ denotes the stress tensor of the form

$$T_{ij} = -p\delta_{ij} + \mu\left(\partial_{x^i}v^j + \partial_{x^j}v^i\right) + (\nu - \mu)\delta_{ij}\mathrm{div} v \equiv -p\delta_{ij} + D_{ij}(v), \tag{1.2}$$

where $i, j = 1,2,3$, $\mathbb{D} = \mathbb{D}(v)$ is the deformation tensor and H is the double mean curvature of S_t, which is negative for convex domains and can be expressed in the form

$$H\bar{n} = \Delta_{S_t}(t)x, \qquad x = (x^1, x^2, x^3), \tag{1.3}$$

where $\Delta_{S_t}(t)$ is the Laplace–Beltrami operator on S_t. Let S_t be determined by $x = x(s1, s2, t)$, $(s1, s2) \in U \subset \mathbb{R}^2$, where U is an open set. Then we have

$$\Delta_{S_t}(t) = g^{-1/2}\partial_{s^\alpha}g^{-1/2}\hat{g}_{\alpha\beta}\partial_{s^\beta} = g^{-1/2}\partial_{s^\alpha}g^{1/2}g^{\alpha\beta}\partial_{s^\beta}, \qquad \alpha,\beta = 1,2, \tag{1.4}$$

* Received by the editors November 1, 1989; accepted for publication (in revised form) April 1, 1993.

† Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, 00-950 Warsaw, Poland.

where the convention summation over the repeated indices is assumed, $g = \det\{g_{\alpha\beta}\}_{\alpha,\beta=1,2}$, $g_{\alpha\beta} = x_\alpha \cdot x_\beta$, where $x_\alpha = \partial_{s^\alpha} x$, $\{g^{\alpha\beta}\}$ is the inverse matrix to $\{g_{\alpha\beta}\}$ and $\{\hat{g}_{\alpha\beta}\}$ is the matrix of algebraic complements for $\{g_{\alpha\beta}\}$.

Let the domain Ω be prescribed. Then, by (1.1e), $\Omega_t = \{x \in \mathbb{R}^3 : x = x(\xi, t), \xi \in \Omega\}$, where $x = x(\xi, t)$ is the solution of the Cauchy problem

$$\frac{\partial x}{\partial t} = v(x,t), \qquad x|_{t=0} = \xi \in \Omega, \quad \xi = (\xi^1, \xi^2, \xi^3). \tag{1.5}$$

Therefore the transformation $x = x(\xi, t)$ connects the Eulerian x and the Lagrangian ξ coordinates of the same fluid particle. Hence

$$x = \xi + \int_0^t u(\xi, s)ds \equiv X_u(\xi, t) \equiv x(\xi, t), \tag{1.6}$$

where $u(\xi, t) = v(X_u(\xi, t), t)$. Moreover, the kinematic boundary condition (1.1e) implies that the boundary S_t is a material surface; so if $\xi \in S = S_0$, then $X_u(\xi, t) \in S_t$ and $S_t = \{x : x = X_u(\xi, t), \xi \in S\}$.

By the continuity equation (1.1b) and the kinematic condition (1.1e) the total mass M is conserved and

$$\int_{\Omega_t} \rho(x,t)dx = M, \tag{1.7}$$

which is a relation between ρ and Ω_t.

Let us define an equilibrium state to be a solution of (1.1a–e) such that $v = 0$, $\Omega_t = \Omega_e$ is a ball for all $t \in \mathbb{R}^1$ and $f = 0$. Then, in view of (1.1a, d), $\rho = \rho_e = \text{const}$, and $p(\rho_e) = (2\sigma/R_e) + p_0$, where $R_e = ((3/4\pi)|\Omega_e|)^{1/3}$, and by (1.7), $\rho_e = M/((4\pi/3)R_e^3)$. By summarizing we have the following.

DEFINITION 1.1. Let M, σ, p_0 and a functional dependence $p = p(\rho)$ be given. Let $f = 0$. Then by the equilibrium state we mean a solution of (1.1a–e) such that

$$v = 0, \quad \rho = \rho_e, \quad \Omega_t = \Omega_e, \quad t \in \mathbb{R}^1_+, \tag{1.8}$$

is a ball of radius R_e that is a solution of the equation

$$p(M/((4\pi/3)R_e^3)) = 2\sigma/R_e + p_0, \tag{1.9}$$

which is also a relation between the total mass M and volume $|\Omega_e|$.

In this paper the existence of such global solutions is proved when the velocity v is small, the density ρ is close to a constant, the domain Ω_t is close to a ball, and the external force vanishes. Hence, we show the stability of the equilibrium state, which means that any motion described by (1.1a–e), which starts from a state sufficiently close to the equilibrium state, remains close to it for all time. However, we are not able to show that it converges to the equilibrium state as t is passing to infinity. To prove the global existence of solutions of (1.1a–e) close to the equilibrium state, the surface tension is important because it controls a shape of the boundary and implies that it is close to a ball. The case without surface tension is considered in [35].

Finally, to prove local existence (see Lemma 5.1) and then to prove global existence (see Theorems 5.5 and 5.6) we need the following compatibility conditions:

$$D_s^\alpha \partial_t^i (\mathbb{T}(v,p)\bar{n} - \sigma H\bar{n} + p_0\bar{n})|_{t=0,s} = 0, \qquad |\alpha| + i \leq 2,$$

where times derivatives of $v, p, \rho, \bar{n}$ at $t = 0$ are calculated from (1.1 a, b) and D_s means tangent derivatives only (see Remark 5.7).

As far as we know this paper and [35] are the first papers that treat global existence of solutions to free boundary problems for compressible viscous fluids in three dimensions. In the one-dimensional case there is a result of Matsumura and Nishida [8], who additionally takes gravitation into account.

Since 1976 Solonnikov has been working in free boundary problems for equations of incompressible viscous fluid [20], [21], [23], [25]–[29]. In a series of papers he showed the existence of global motion of a viscous incompressible fluid bounded by a free boundary, both with surface tension (see [23] and [25]) and without it (see [26]). The latter case is proved for solutions of incompressible Navier–Stokes equations by using the Korn inequality. To prove the existence of solutions of the incompressible version of problem (1.1a–e) with surface tension the existence of solutions of the initial-boundary value problem for the Stokes system with a corresponding boundary condition of type (1.1d) with surface tension has to be shown (linear problem). By using potential theory techniques, this was also accomplished by Solonnikov (see [21]). It should be emphasized that the existence of solutions of the latter problem was shown in anisotropic Sobolev–Slobodetskii spaces $W_2^{l,l/2}$ with noninteger positive l (see definitions at the end of this section). The boundary condition (1.1d) with surface tension contains both first- and second-order derivatives of solutions, so the considered problem is noncoercive. Solonnikov used the spaces $W_2^{l,l/2}$ with noninteger l to prove the existence of solutions of the incompressible version of the problem (1.1a–e) in Sobolev spaces as low as possible to omit compatibility conditions. In the case of compressible fluid we have had to also prove the existence for the linear problem (3.3) which, because of the boundary condition (1.1d) (see also (3.3)), does not follow from the general theory of initial-boundary value problems for Douglis–Nirenberg parabolic systems (see [24]). The existence of solutions of that problem is shown in the same anisotropic Sobolev–Slobodetskii spaces as in the incompressible case (see [34] and [36]). This implies that in this paper the technique of spaces $W_2^{l,l/2}$ has to be used in §3, where the local existence for (1.1a–e) is considered (see [34] and [36]). In these considerations it is convenient to use noninteger l to reduce a number of coefficients of type T^{-a}, $a > 0$, where T is the time of local existence, in the norms used in its proof (see [36]). Such coefficients may imply difficulties in the proof of local existence. However, to prove global existence we need local solvability of (1.1a–e) in such classes that $v \in W_2^{4,2}(\Omega^T)$, $\rho \in W_2^{3,3/2}(\Omega^T) \cap C(0,T;\Gamma_0^{3,3/2}(\Omega))$ (see Remark 3.2 and [37], where existence of local solutions in these classes is shown). Then global existence is proved in class $\mathcal{M}(t), t \in \mathbb{R}_+$ (see definition of $\mathcal{M}(t)$ at the beginning of §5), which is implied by differential inequality (4.195) (see Lemma 5.1).

Now we make some comments on the literature concerning free boundary problems for the nonstationary incompressible Navier–Stokes system. Local existence of solutions in the case without surface tension is proved in Hölder and Sobolev anisotropic space by Solonnikov in [26] and [27] (see also [20]). Potential theory techniques are used to prove the existence of solutions of the corresponding linear problems in Hölder and in Sobolev spaces (see [28] and [29]). In all papers of Solonnikov the Lagrangian coordinates are used. Global existence is also proved by Beale [3], [4], where the free boundary is infinite and gravitation is taken into account. Local existence with surface tension is considered by Allain [2].

Local existence of solutions for compressible fluids without surface tension is proved by Secchi and Valli (see [16]), but with surface tension by Solonnikov and

Tani (see [31]).

Lately, Secchi showed existence and uniqueness of solutions of equations describing the motion of gaseous stars (see [17]–[19]).

References to the literature concerning stationary free boundary problems can be found in [15]. Moreover, in [15] Pileckas and Zajaczkowski proved the existence of stationary motion of viscous, compressible, barotropic fluid bounded by a free surface governed by surface tension. In the proof, one has to assume that the domain and the external force satisfy some extra symmetry conditions. In the present paper the global existence is proved for $f = 0$ and without any symmetry conditions, so there is no connection with the result in [15]. However, similar to the case treated in [15], to prove global existence the necessary a priori estimate is found by the energy method that was also used in papers [8]–[12], [15], [32], [33], and [35].

This paper relies heavily on the following main points. First we study inequality (2.33), which guarantees that variations of the volume of Ω_t (denoted by $|\Omega_t|$) and the surface area of S_t (denoted by $|S_t|$) are as small as we need for all time. This inequality follows from conservation laws for (1.1a–e) (see Lemma 2.1) and a special choice of the parameters of the problem (1.1a–e) $(\mu, \nu, \sigma, p_0, p = p(\rho), |\Omega|, |S|, v_0, \rho_0, f = 0)$ (see Lemma 2.2), which implies that the right-hand side of (2.20) is sufficiently small. This result is crucial for the proof of global existence in the compressible case only, because in the incompressible case $|\Omega_t|$ is constant. We have to underline that the result is shown under the assumption that the considered fluid is barotropic, so $p = A\rho^\kappa$, where A, κ are constants and $\kappa > 1$. Moreover, (2.65) is essential because it ensures that Ω_t is close to a ball. Second, we prove the local-in-time existence of solutions to (1.1 a–e) by employing the Lagrangian coordinates, and we find a suitable a priori estimate (see Theorem 3.1—the proof is shown in [36]). The existence is proved in such classes that $v \in W_2^{l+2,l/2+1}$, $\rho \in W_2^{l+1,l/2+1/2}$, $l/2 - 3/4 - \kappa \in \mathbb{N} \cup \{0\}$, $\kappa \in (0, 1/4)$. However, to prove global existence we need a local solution such that $v \in W_2^{4,2}(\Omega^T)$, $\rho \in W_2^{3,3/2}(\Omega^T) \cap C([0,T]; \Gamma_0^{3,3/2}(\Omega))$ (see [37] and Remark 3.2).

Moreover, to prove global existence we also need a bound for variations of the solution near the equilibrium state in terms of the appropriate norms of v_0, $p(\rho_0) - p_0 - 2\sigma/R_0$, $H(x,0)) + 2/R_0$ (see Remark 3.2, where (3.6) is such an inequality), where the last expression measures the deviation of the initial domain from a ball. From this inequality and imbedding theorems we have the minimum and maximum of the density, which are necessary to find the inequality (2.65). This fact and the use of Lagrangian coordinates, which also exist only locally suggests that the proof of global existence should be done step by step by employing local existence. However, this procedure needs a special a priori estimates because (3.6) is not sufficiently strong. The inequalities are shown in §4 (see (4.195), Theorems 4.13 and 4.14, and (4.197)), whose proofs constitute the most technically difficult part of this paper. It requires the technique of an energy method, which is very close to the methods used in [33] (see also [8]–[12], [15], and [32]). To prove the inequality we need much more regularity of solutions than is needed to show the local existence. This is achieved in Lemma 5.1 under appropriate assumptions on initial data.

The global existence is proved in the case $f = 0$. The main result is formulated in Theorem 5.5. In the case of the external pressure $p_0 = 0$ the proof of Lemma 2.2 essentially simplifies and the necessary relations between parameters of (1.1a–e) may be explicitly formulated (see Lemma 2.8). In this case the global existence result is stated in Theorem 5.6.

In this paper we use the anisotropic Sobolev–Slobodetskii spaces $W_2^{l,l/2}(\Omega^T)$,

$l \in \mathbb{R}_+$ (see [5, Chap. 18]) of functions defined in $\Omega^T = \Omega \times (0,T)$. In fact, $W_2^{l,l/2}$ are Besov spaces for $l \notin \mathbb{Z}$; the equivalence between $W_2^{l,l/2}$, $l \notin \mathbb{Z}$, and Besov spaces follows from considerations in [1, Chap. 7]. In the case of noninteger l we have the norms ($\Omega \subset \mathbb{R}^3$)

(1.10)

$$\|u\|_{W_2^{l,0}(\Omega^T)} = \left(\int_0^T \|u\|^2_{W_2^l(\Omega)} dt \right)^{1/2},$$

$$\|u\|_{W_2^{0,l/2}(\Omega^T)} = \left(\int_\Omega \|u\|^2_{W_2^{l/2}((0,T))} dx \right)^{1/2},$$

$$\|u\|^2_{W_2^l(\Omega)} = \sum_{|\alpha| \le [l]} \|D_x^\alpha u\|^2_{L_2(\Omega)} + \sum_{|\alpha| = [l]} \int_\Omega \int_\Omega \frac{|D_x^\alpha u(x,t) - D_y^\alpha u(y,t)|^2}{|x-y|^{3+2(l-[l])}} dxdy,$$

$$\|u\|^2_{W_2^{l/2}((0,T))} = \sum_{j=0}^{[l/2]} \|D_t^j u\|^2_{L_2((0,T))} + \int_0^T \int_0^T \frac{|D_t^{[l/2]} u(x,t) - D_\tau^{[l/2]} u(x,\tau)|^2}{|t-\tau|^{1+2(l/2-[l/2])}} dtd\tau,$$

$$\|u\|^2_{W_2^{l,l/2}(\Omega^T)} = \sum_{|\alpha|+2\alpha_0 \le [l]} \left(\|D_x^\alpha \partial_t^{\alpha_0} u\|^2_{L_2(\Omega^T)} \right) + \int_0^T dt \int_\Omega \int_\Omega dxdy \frac{|D_x^\alpha \partial_t^{\alpha_0} u(x,t) - D_y^\alpha \partial_t^{\alpha_0} u(y,t)|^2}{|x-y|^{3+2(l-[l])}} + \int_\Omega dx \int_0^T \int_0^T dtdt' \frac{|D_x^\alpha \partial_t^{\alpha_0} u(x,t) - D_x^\alpha \partial_{t'}^{\alpha_0} u(x,t')|^2}{|t-t'|^{1+2(l/2-[l/2])}},$$

where $D_x^\alpha = \partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n}$, $\partial_x = \partial/\partial x$, $\partial_t = \partial/\partial t$, and we use generalized (Sobolev) derivatives. Similarly, by using local coordinates and a partition of unity we introduce the norm in the space $W_2^{l,l/2}(S^T)$ of functions defined on $S^T = S \times (0,T)$, where $S = \partial\Omega$. We also use $W_2^l(\Omega)$ with the third norm of (1.10) for functions defined in Ω. We do not distinguish norms of scalar and vector-valued functions. To simplify notation we write

$$\|u\|_{l,Q} = \|u\|_{W_2^{l,l/2}(Q)} \quad \text{if } Q = \Omega^T \quad \text{or} \quad Q = S^T, l \ge 0, \|u\|_{l,Q} = \|u\|_{W_2^l(Q)} \quad \text{if } Q = \Omega$$

or $Q = (0,T), l \ge 0$, and $W_2^{0,0}(Q) = W_2^0(Q) = L_2(Q)$. Moreover, $\|u\|_{L_p(Q)} = |u|_{p,Q} 1 \le p \le \infty$ and $\|u\|_{l,p,\Omega^T} = \|u\|_{L_p(0,T;W_2^l(\Omega))}$.

For the proof of global existence we also need the following notation: $H^l(Q) = W_2^l(Q)$, where Q is one of the following expressions: Ω, S, Ω^T, and S^T. We introduce the space $\Gamma_0^{l,l/2}(\Omega)$ with the norm

$$\|u\|_{\Gamma_0^{l,l/2}(\Omega)} = \sum_{0 \le 2i+|\alpha| \le 1} \|\partial_t^i D_x^\alpha u\|_{0,\Omega} \equiv |u|_{l,0,\Omega},$$

and $\Gamma_k^l(\Omega)$ with the norm

$$\|u\|_{\Gamma_k^l(\Omega)} = \sum_{0 \le i \le l-\kappa} \|\partial_t^i u\|_{l-i,\Omega} \equiv |u|_{l,k,\Omega},$$

where $l \geq k$, $k \in \mathbb{Z}_+ \cup \{0\}$, $0 \leq l \in \mathbb{R}$.

We define $L_p(0,T;\Gamma_0^{l,l/2}(\Omega))$ with

$$\|u\|_{L_p(0,T;\Gamma_0^{l,l/2}(\Omega))} \equiv |u|_{l,0,p,\Omega^T},$$

and $C([0,T];\Gamma_0^{l,l/2}(\Omega))$ with the norm

$$\|u\|_{C([0,T];\Gamma_0^{l,l/2}(\Omega))} = \sup_{t\in[0,T]} |u|_{l,0,\Omega}.$$

We also need

$$|u|_{l,k} = \sum_{0\leq i\leq l-\kappa} \sum_{|\alpha|=l-1} |D_x^\alpha \partial_t^i u|,$$

where $||$ is the Euclidean norm either of a vector or a matrix.

Moreover, we shall use the imbedding (see [5] and [13])

$$W_r^\delta(\Omega) \subset L_p^\alpha(\Omega), \qquad \Omega \subset \mathbb{R}^3, \alpha + 3/r - 3/p \leq \delta, \tag{1.11}$$

where $\|u\|_{L_p^\alpha(\Omega)} = |D_x^\alpha u|_{p,\Omega}$ and the corresponding interpolation inequality holds

$$|D_x^\alpha u|_{p,\Omega} \leq \varepsilon^{1-\kappa} |D_x^\delta u|_{r,\Omega} + c\varepsilon^{-\kappa} |u|_{r,\Omega}, \tag{1.12}$$

where $\kappa = \alpha/\delta + (3/\delta)(1/r - 1/p) < 1, \varepsilon \in (0,1)$.

2. Global estimates and relations. We start with conservation laws for problem (1.1a–e).

LEMMA 2.1. *Sufficiently smooth solutions to problem* (1.1a–e) *satisfy*

$$\begin{aligned} &\frac{d}{dt}\left[\int_{\Omega_t} \left(\frac{1}{2}\rho v^2 + \rho h(\rho)\right) dx + p_0|\Omega_t| + \sigma|S_t|\right] \\ &\quad + \frac{\mu}{2} E(v,v) + (\nu - \mu)\|\operatorname{div} v\|_{0,\Omega_t}^2 = \int_{\Omega_t} \rho f \cdot v \, dx, \end{aligned} \tag{2.1}$$

where $|\Omega_t| = \operatorname{vol}\Omega_t$, $|S_t|$ *is the surface area of* S_t, $h(\rho) = \int (p(\rho)/\rho^2) d\rho$,

$$E(v,u) = \int_{\Omega_t} \left(v^i_{x^j} + v^j_{x^i}\right)\left(u^i_{x^j} + u^j_{x^i}\right) dx.$$

Moreover,

$$\frac{d}{dt}\int_{\Omega_t} \rho v \cdot \eta \, dx = \int_{\Omega_t} \rho f \cdot \eta \, dx, \tag{2.2}$$

where $\eta = a + b \times x, a, b$ *are arbitrary constant vectors, and*

$$\frac{d}{dt}\int_{\Omega_t} \rho x \, dx = \int_{\Omega_t} \rho v \, dx. \tag{2.3}$$

Proof. To prove (2.1) we multiply (1.1a) by v, (1.1b) by $v^2/2$, add the results, and integrate over Ω_t to get

$$\int_{\Omega_t} \frac{1}{2}\left[\partial_t(\rho v^2) + \nabla\cdot(\rho v v^2)\right] dx - \int_{S_t} T_{ij} n^i v^j d\tau$$
$$+ \int_{\Omega_t} \left[-p\delta_{ij} + \mu\left(v^i_{x^j} + v^j_{x^i}\right) + (\nu-\mu)\delta_{ij}\ \text{div}\ v\right] v^i_j\, dx = \int_{\Omega_t} \rho f\cdot v\, dx.$$

Using (1.1d) we obtain

$$\frac{d}{dt}\int_{\Omega_t}\frac{1}{2}\rho v^2\, dx - \int_{S_t}(\sigma H - p_0)\bar{n}\cdot v d\tau + \frac{\mu}{2}E(v,v) + (\nu-\mu)\|\text{div}\ v\|^2_{0,\Omega_t}$$
$$- \int_{\Omega_t} p\ \text{div}\ v\, dx = \int_{\Omega_t}\rho f\cdot v\, dx. \tag{2.4}$$

From (1.3) and (1.4), as in [23], one obtains

$$-\int_{S_t} H\bar{n}\cdot v d\tau = -\int_{S_t}\frac{1}{\sqrt{g}}\partial_{s^\alpha}(g^{\alpha\beta}\sqrt{g}x_\beta)x_t d\tau = \int_U g^{\alpha\beta}\sqrt{g}x_\beta x_{\alpha t} ds^1 ds^2$$
$$= \frac{1}{2}\int_U g^{\alpha\beta}\sqrt{g}\partial_t g_{\alpha\beta} ds^1 ds^2 = \int_U \partial_t\sqrt{g} ds^1 ds^2 = \frac{d}{dt}\int_U \sqrt{g} ds^1 ds^2 = \frac{d}{dt}|S_t|. \tag{2.5}$$

By (2.4) and (2.5) we have

$$\frac{d}{dt}\left(\int_{\Omega_t}\frac{1}{2}\rho v^2\, dx + |S_t|\right) + \frac{\mu}{2}E(v,v) + (\nu-\mu)\|\text{div}\ v\|^2_{0,\Omega_t}$$
$$- \int_{\Omega_t}(p-p_0)\text{div}\ v\, dx = \int_{\Omega_t}\rho f\cdot v\, dx. \tag{2.6}$$

Now we consider the term involving the pressure. Using the equation of continuity we have

$$-\int_{\Omega_t}(p-p_0)\text{div}\ v\, dx = \int_{\Omega_t}(p-p_0)(\rho\rho_t + \rho v\cdot\nabla\rho)/\rho^2\, dx$$
$$= \int_{\Omega_t}\left[\rho(h+p_0 h_1)_{,t} + \rho v\cdot\nabla(h+p_0 h_1)\right] dx,$$

where $h_1(\rho) = -\int(d\rho/\rho^2) = 1/\rho$. Since

$$\int_{\Omega_t}\left[\rho_t + \text{div}(\rho v)\right](h+p_0 h_1)\, dx = 0,$$

we get

$$-\int_{\Omega_t}(p-p_0)\text{div}\ v\, dx = \int\left[(\rho(h+p_0 h_1))_{,t} + \text{div}(\rho v(h+p_0 h_1))\right] dx$$
$$= \frac{d}{dt}\int_{\Omega_t}\rho(h+p_0 h_1)\, dx. \tag{2.7}$$

Inserting (2.7) in (2.6) and applying $\rho h_1 = 1$ gives (2.1).

To show (2.2) we pass to Lagrangian coordinates, so we have

$$\frac{d}{dt}\int_{\Omega_t} \rho v \cdot \eta \, dx = \frac{d}{dt}\int_{\Omega} \rho_0(\xi) v(X_u(\xi,t),t) \cdot \eta(X_u(\xi,t)) d\xi,$$

where we used the fact that $\rho = \rho_0/J$ and J is the Jacobian of the transformation (1.6).

Now differentiating the integrand, using (1.1b), and proceeding exactly in the same way as in [23] we get (2.2). Finally,

$$\frac{d}{dt}\int_{\Omega_t} \rho x \, dx = \frac{d}{dt}\int_{\Omega} \rho_0(\xi) X_u(\xi,t) d\xi = \int_{\Omega} \rho_0(\xi) v(X_u(\xi,t),t) d\xi = \int \rho v \, dx,$$

so (2.3) is satisfied. This concludes the proof.

Now we find restrictions on parameters of problem (1.1a–e) which imply that the variation of $|\Omega_t|$ is small for all time. Let us assume that solutions of (1.1a–e) are sufficiently regular. Let

$$\nu - \mu/3 \geq 0. \tag{2.8}$$

Then

$$\begin{aligned}
&\frac{\mu}{2} E(v,v) + (\nu-\mu)\|\operatorname{div} v\|^2_{0,\Omega_t} \\
&\quad = \frac{\mu}{2}\int_{\Omega_t} \left(v^i_{x^j} + v^j_{x^i}\right)^2 dx + (\nu-\mu)\int_{\Omega_t} (\operatorname{div} v)^2 \, dx \\
&\quad \geq \sum_{i=j} \frac{\mu}{2}\int_{\Omega_t} \left(v^i_{x^j} + v^j_{x^i}\right)^2 dx + (\nu-\mu)\int_{\Omega_t} (\operatorname{div} v)^2 \, dx \\
&\quad = 2\mu \sum_i \int_{\Omega_t} \left(v^i_{x^i}\right)^2 dx + (\nu-\mu)\int_{\Omega_t} (\operatorname{div} v)^2 \, dx.
\end{aligned}$$

Since $(\xi_1+\xi_2+\xi_3)^2 \leq 3(\xi_1^2+\xi_2^2+\xi_3^2)$ the above expression is not less than $(\nu - \mu/3)\|\operatorname{div} v\|^2_{0,\Omega_t}$, which by (2.8) is nonnegative.

Hence, assuming $f = 0$, (2.1) implies

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega_t} \rho v^2 \, dx + \int_{\Omega_t} \varphi(\rho) \, dx + p_0|\Omega_t| + \sigma|S_t| \\
&\quad \leq \frac{1}{2}\int_{\Omega} \rho_0 v_0^2 \, dx + \int_{\Omega} \varphi(\rho_0) \, dx + p_0|\Omega| + \sigma|S| \equiv d,
\end{aligned} \tag{2.9}$$

where $p(\rho) = A\rho^\kappa$, so $\varphi(\rho) = (A/(\kappa-1))\rho^\kappa, \kappa > 1$.

Using (1.7) and the Hölder inequality one gets

$$M = \int_{\Omega_t} \rho \, dx \leq |\Omega_t|^{1-1/\kappa} \left(\int_{\Omega_t} \rho^\kappa \, dx\right)^{1/\kappa}.$$

Now from (2.9) one obtains

$$\left(\frac{A}{\kappa-1}\frac{M^\kappa}{d}\right)^{1/(\kappa-1)} \leq |\Omega_t| \leq \frac{d}{p_0}. \tag{2.10}$$

Introducing the mean density $\bar{\rho}_t = M/|\Omega_t|$, one gets

$$\frac{Mp_0}{d} \le \bar{\rho}_t \le \left(\frac{(\kappa-1)d}{AM}\right)^{1/(\kappa-1)}. \tag{2.11}$$

We also obtain

$$\left(\frac{p_0}{d}\right)^{\kappa-1} M^\kappa \le \int_{\Omega_t} \rho^\kappa \, dx \le \frac{(\kappa-1)d}{A}. \tag{2.12}$$

In this way we have shown that $|\Omega_t|$ and $\psi_t = \int_{\Omega_t} \varphi(\rho(x))dx$ are bounded from below and from above.

Multiplying equation (2.9) by $|\Omega_t|^{\kappa-1}$, using the Hölder inequality $(\int_{\Omega_t} \rho \, dx)^\kappa \le |\Omega_t|^{\kappa-1} \int_{\Omega_t} \rho^\kappa \, dx$ and (1.7), we obtain

$$\begin{aligned} y(|\Omega_t|) + |\Omega_t|^{\kappa-1} \frac{1}{2} \int_{\Omega_t} \rho v^2 \, dx + \sigma(|S_t| - 4\pi R_t^2)|\Omega_t|^{\kappa-1} \\ + \frac{A}{\kappa-1} \left(|\Omega_t|^{\kappa-1} \int_{\Omega_t} \rho^\kappa \, dx - \left(\int_{\Omega_t} \rho \, dx \right)^\kappa \right) \le 0, \end{aligned} \tag{2.13}$$

where

$$y(x) = p_0 x^\kappa + c_0 \sigma x^{\kappa-1/3} - dx^{\kappa-1} + \frac{AM^\kappa}{\kappa-1}, \quad \kappa > 1, \quad x \equiv |\Omega_t|, \tag{2.14}$$

and $c_0 = (36\pi)^{1/3}, (4\pi/3)R_t^3 = |\Omega_t|$.

Our aim is to find restrictions on the coefficients of (2.13) $(p_0, \sigma, d, A, \kappa, M)$ that lead to small changes of $|\Omega_t|$ for all $t \ge 0$. For this purpose we have to find a minimum of the left-hand side of (2.13) and to show that for some relations among $p_0, \sigma, d, A, \kappa, M$ (2.13) holds only for small changes of $|\Omega_t|$ near the minimum point. Since the last three terms on the left-hand side of (2.13) are positive, we have

$$y(x) \equiv y(|\Omega_t|) \le 0. \tag{2.15}$$

This inequality holds for all physical drop volumes, and so for all real physical motions governed by (1.1a–e).

Now our aim is to determine minimum points $\{x_0\}$ of $y = y(x)$. In view of (2.15) we must look for such minimum points x_0 that $x_0 > 0, y(x_0) < 0, y'(x_0) = 0, y''(x_0) > 0$. Moreover, the coefficients in $y = y(x)$ must be chosen in such a way that $|x - x_0|$ is small for $x \in \{x : y(x) \le 0\}$. Examining (2.15) instead of (2.13) is justified since $|x - x_0| \le \varepsilon$ for $x \in \{x : y(x) \le 0\}$ (where ε is sufficiently small) implies that $|x - x_0| \le \varepsilon$ also for $x \in \{x : y(x) + a \le 0\}$, where a denotes the sum of the last three terms on the left-hand side of (2.13).

Minimum points of $y = y(x)$ are determined by the equation

$$y'(x) \equiv [p_0 \kappa x + c_0 \sigma(\kappa - 1/3)x^{2/3} - d(\kappa - 1)]x^{\kappa-2} = 0. \tag{2.16}$$

Viete's formulas imply that the solutions x_1, x_2, x_3 of (2.16) satisfy

$$x_1 \cdot x_2 \cdot x_3 = \left(\frac{d(\kappa-1)}{p_0\kappa}\right)^{1/3} > 0, \qquad x_1^{1/3} + x_2^{1/3} + x_3^{1/3} = -\frac{c_0\sigma(\kappa - 1/3)}{p_0\kappa} < 0,$$

so there exists only one positive root of (2.16). Denote it also by x_0. To calculate it we consider the equation which follows from (2.16):

$$\omega^3 + 3p\omega + 2q = 0, \tag{2.17}$$

where $\omega = u + \mu_0$, $u = x^{1/3}$, $p = -\mu_0^2$, $q = \mu_0^3 - \nu_0$, $\mu_0 = (c_0\sigma(\kappa - 1/3))/3p_0\kappa$, $\nu_0 = d(\kappa - 1)/2p_0\kappa$. At a point x_0 of a minimum $y = y(x)$, we have

$$y''(x)\Big|_{x=x_0} = p_0^\kappa x_0^{\kappa-3}\left[-\mu_0 x_0^{2/3} + 2\nu_0\right] = p_0^\kappa x_0^{\kappa-3}\left[x_0 + 2\mu_0 x_0^{2/3}\right] > 0\,, \tag{2.18}$$

because $x_0 > 0$ and the last equality follows from (2.16) expressed in the form $x_0 + 3\mu_0 x_0^{2/3} - 2\nu_0 = 0$.

Now we find restrictions on the variations of $|\Omega_t|$ near the minimum point x_0. Since $\lim_{x\to 0} y(x) = AM^\kappa/(\kappa - 1) > 0$ and $\lim_{x\to\infty} y(x) = \infty$, if we assume that the parameters p_0, σ, κ, and d vary in a bounded set, we find that $x_0 = x_0(p_0, \sigma, \kappa, d)$ belongs to a compact set separated from zero by a positive number, and $y(x) = 0$ has exactly two positive solutions, which also belong to a compact set. Denote them by w_1, w_2, and $0 < w_1 < w_2$. Inequality (2.15) implies that $w_1 \le |\Omega_t| \le w_2$. Expanding $y(x)$ in a Taylor series in a neighborhood of x_0 up to the second order, we obtain

$$y(x) = y(x_0) + \tfrac{1}{2}y''(x_0 + \theta h)h^2,$$

where $\theta \in (0,1)$, $h = x - x_0$. Assume that $(x_0 - h, x_0 + h) \subset (w_1, w_2)$. Since $y(x) \le 0$ for $x \in (x_1, x_2)$, we have

$$h \le \left(\frac{-2y(x_0)}{y''(x_0 + \theta h)}\right)^{1/2}. \tag{2.19}$$

Moreover,

$$0 < y(x) - y(x_0) = \tfrac{1}{2}y''(x_0 + \theta h)h^2;$$

hence for $h > h_* > 0$ we have $y''(x_0 + \theta h) \ge y_* > 0$. For $x_0 + \theta h$ close to x_0 we use the fact that $y''_\alpha = y''_\alpha(x)$, where $\alpha = (p_0, \sigma, \kappa, d) \in B$, which is assumed to be a bounded set, is a continuous function of x. Then, $y''_\alpha(x_0)$, $\alpha \in B$, being separated from zero, so is $y''_\alpha(x)$ for x from a sufficiently small neighborhood of x_0.

By (2.19), $-y(x_0)$ small yields h small. Then (2.13) implies

$$\begin{aligned} &|\Omega_t|^{\kappa-1}\int_{\Omega_t}\frac{1}{2}\rho v^2\,dx + \sigma(|S_t| - 4\pi R_t^2)|\Omega_t|^{\kappa-1} \\ &\quad + \frac{A}{\kappa - 1}\left(|\Omega_t|^{\kappa-1}\int_{\Omega_t}\rho^\kappa\,dx - \left(\int_{\Omega_t}\rho\,dx\right)^\kappa\right) \le -y(x_0). \end{aligned} \tag{2.20}$$

To get $|h|$ small we can also assume that the denominator on the right-hand side of (2.19) is large and $-y(x_0)$ is bounded. In this case the left-hand side of (2.20) is not small. However, it is difficult to find conditions guaranteeing that $y''(x)$ is large.

Now we find the minimum of $y = y(x)$, so we look for solutions to (2.17). Let

$$D = q^2 + p^3 = \nu_0(\nu_0 - \mu_0^3). \tag{2.21}$$

Then we have the following possibilities:

(2.22)

$$D > 0, \nu_0 > 2\mu_0^3, \quad \cosh\varphi := \nu_0/\mu_0^3 - 1 > 1, \quad u = \mu_0(2z - 1) > 0, \quad z = \cosh\varphi/3.$$

(2.23)

$$D \le 0 \quad \text{and } \nu_0 \in (\mu_0^3, 2\mu_0^3], \quad \cos\varphi := \nu_0/\mu_0^3 - 1, \quad u = \mu_0(2z-1) > 0, \quad z = \cos\varphi/3.$$

(2.24)

$$D \le 0 \quad \text{and } \nu_0 \in (0, \mu_0^3], \quad \cos\varphi := 1 - \nu_0/\mu_0^3, \quad u = \mu_0(2z-1), \quad z = \cos(\pi/3 - \varphi/3).$$

Moreover, $x_0 = u^3$.

Using the above notation we have

$$\begin{aligned} y(x_0) = {} & p_0\mu_0^{3(\kappa-1)}(2z-1)^{3(\kappa-1)}[(\kappa-1/3)^{-1}\mu_0^3(2z-1)^2 - 2\nu_0/(\kappa-1)] \\ & + AM^\kappa/(\kappa-1) \end{aligned} \tag{2.25}$$

and $y''(x)_{|x=x_0|} > 0$ yields

$$\mu_0^3(2z-1)^2 - 2\nu_0 < 0. \tag{2.26}$$

Equation (2.26) implies that the first term on the right-hand side of (2.25) is negative. This enables us to make $y(x_0)$ arbitrarily small.

Let (2.22) be satisfied. Then (2.26) is satisfied and (2.25) has the form

$$\begin{aligned} y(x_0) = {} & -(\kappa-1)^{-1}p_0\mu_0^{3\kappa}(2\cosh(\varphi/3) - 1)^{3(\kappa-1)} \\ & \cdot [2(\cosh\varphi + 1) - (\kappa-1)(\kappa-1/3)^{-1}(2\cosh(\varphi/3) - 1)^2\,] \\ & + AM^\kappa/(\kappa-1) \equiv -\Phi_1(\mu_0, \varphi, p_0, \kappa, A, M). \end{aligned} \tag{2.27}$$

The first expression in (2.27) is a decreasing continuous function of $z = \cosh(\varphi/3)$ and μ_0, which vanishes for $z = \frac{1}{2}$. Thus there exists a set of parameters of $\mu_0, \nu_0, p_0, \kappa, A, M$, such that $y(x_0)$ is very small negative.

Let (2.23) be satisfied. Then (2.26) is satisfied and (2.25) yields

(2.28)

$$\begin{aligned} y(x_0) = {} & -(\kappa-1)^{-1}p_0\mu_0^{3\kappa}(2\cos(\varphi/3) - 1)^{3(\kappa-1)}\,[2(\cos\varphi + 1) \\ & -(\kappa-1)(\kappa-1/3)^{-1}(2\cos(\varphi/3) - 1)^2] \\ & + AM^\kappa/(\kappa-1) \equiv -\Phi_2(\mu_0, \varphi, p_0, \kappa, A, M), \end{aligned}$$

where the first expression achieves its minimum at $\varphi = 0$ and increases with φ. Therefore, to guarantee that $y(x_0)$ is negative for some φ we have to assume that

$$(3\kappa - 1/3)(\kappa - 1/3)^{-1}p_0\mu_0^{3\kappa} > AM^\kappa. \tag{2.29}$$

Under this assumption there exists a set of parameters such that $y(x_0)$ given by (2.28) is very small and negative.

Finally, assume that (2.24) is satisfied. Then (2.26) is satisfied for $\varphi \in (0, \pi]$. Moreover, (2.25) gives

(2.30)

$$\begin{aligned} y(x_0) = {} & -(\kappa-1)^{-1}p_0\mu_0^{3\kappa}(2\cos(\pi/3 - \varphi/3) - 1)^{3(\kappa-1)} \\ & \cdot [2(1 - \cos\varphi) - (\kappa-1)(\kappa-1/3)^{-1}(2\cos(\pi/3 - \varphi/3) - 1)^2] + AM^\kappa/(\kappa-1) \\ \equiv {} & -(\kappa-1)^{-1}p_0\mu_0^{3\kappa}\Gamma(\varphi) + AM^\kappa/(\kappa-1) \equiv -\Phi_3(\mu_0, \varphi, p_0, \kappa, A, M), \end{aligned}$$

where $\varphi \in (0,\pi]$, $\Gamma(0)=0$, $\Gamma(\pi/2)=(\sqrt{3}-1)^{3(\kappa-1)}[2-(\kappa-1)(\kappa-1/3)^{-1}(\sqrt{3}-1)^2]$ and $d\Gamma/d\varphi > 0$. Therefore, as in the above case we have to assume that

$$p_0\mu_0^3\Gamma(\varphi) > AM^\kappa, \qquad \varphi \in (0,\pi]. \tag{2.31}$$

The above considerations imply that there exist sets of parameters μ_0, ν_0, p_0, κ, A, M such that conditions (2.22), (2.23), and (2.24) can be satisfied and $y = y(x_0)$ can be determined by (2.27), (2.28), and (2.30), respectively, such that $y(x_0) < 0$ and

$$|y(x_0)| \le \varepsilon, \tag{2.32}$$

where ε may be made arbitrarily small. Then by (2.19) it follows that

$$\sup_t \operatorname{var}|\Omega_t| \le c_1\varepsilon, \qquad t \in \mathbb{R}_+. \tag{2.33}$$

From (2.33) we have $|\int_{\Omega_t}\rho^\kappa\,dx - \int_{\Omega_{t'}}\rho^\kappa\,dx| = |\int_{\Omega_t\setminus\Omega_{t'}}\rho^\kappa\,dx| \le c_2\varepsilon$ (where for convenience we assume that $\Omega_{t'} \subset \Omega_t$) because $\int_{\Omega_t}\rho^\kappa\,dx$ is bounded. Hence

$$\sup \operatorname{var}|\psi_t| \le c_2\varepsilon, \qquad t \in \mathbb{R}_+, \tag{2.34}$$

where $\psi_t = \int_{\Omega_t}\varphi(\rho)\,dx$.

Thus we have proved the following.

LEMMA 2.2. *Let $\varepsilon > 0$ be small. Assume that the parameters $\mu_0, \nu_0, p_0, \kappa, A, M$, where $\mu_0 = c_0\sigma(\kappa-1/3)/(3p_0)$, $\nu_0 = d(\kappa-1)/(2p_0)$, satisfy the relation*

$$\nu_0 \in I_i, \qquad 0 < \Phi_i(\mu_0,\varphi_i,p_0,\kappa,A,M) \le \varepsilon, \tag{2.35}$$

where $i = 1,2,3, \Phi_i, i = 1,2,3$, are determined by (2.27), (2.28), *and* (2.30), *respectively, $I_1 = (2\mu_0^3,\infty)$, $I_2 = (\mu_0^3, 2\mu_0^3]$, $I_3 = (0,\mu_0^3]$, $\cosh\varphi_1 = \nu_0/\mu_0^3 - 1$, $\cos\varphi_2 = \nu_0/\mu_0^3 - 1$, $\cos\varphi_3 = 1 - \nu_0/\mu_0^3$.*

Then there exist constants c_1, c_2 independent of ε (they can depend on the parameters) such that (2.33) *and* (2.34) *hold. Moreover, in the case* (2.35) *we have*

$$\||\Omega_t| - Q_i\| \le c_3\varepsilon \quad \forall t \in \mathbb{R}_+, \tag{2.36}$$

where $i = 1,2,3, Q_1 = \mu_0^3(2\cosh(\varphi_1/3)-1)^3$, $Q_2 = \mu_0^3(2\cos(\varphi_2/3)-1)^3$, and $Q_3 = \mu_0^3(2\cos(\pi/3-\varphi_3/3)-1)^2$.

The condition (2.36) means that for parameters satisfying (2.35) the volume $|\Omega_t|$ of the considered drop does not differ much from the constant value $Q_i, i = 1,2,3$. This means that the initial volume $|\Omega|$ must also be close to $Q_i, i = 1,2,3$.

On the other hand, from physical reasons the drop volume $|\Omega_t|$ should also be close to the volume $|\Omega_e|$ of the drop in the equilibrium state (see Definition 1.1).

For this purpose we show the following.

LEMMA 2.3. *Let the assumptions of Lemma* 2.2 *be satisfied. Then there exists a constant c_4 such that*

$$|Q_i - |\Omega_e|| \le c_4\varepsilon, \qquad i = 1,2,3, \tag{2.37}$$

where ε is sufficiently small.

Proof. Consider the case Q_i, where $i \in \{1,2,3\}$. Recall that $x_0 = Q_i$. Then (2.16) determining x_0 has the form

$$p_0x_0^\kappa + (\kappa-1/3)\kappa^{-1}c_0\sigma x_0^{\kappa-1/3} + (\kappa-1)\kappa^{-1}\,dx_0^{\kappa-1} = 0. \tag{2.38}$$

Moreover, we have

$$0 \le -y(x_0) = -(p_0 x_0^\kappa + c_0 \sigma x_0^{\kappa - 1/3} - d x_0^{\kappa-1} + AM^\kappa/(\kappa-1)) \le \varepsilon^2. \tag{2.39}$$

In the case of the barotropic fluid (1.9) takes the form

$$p_0 |\Omega_e|^\kappa + 2(4\pi/3)^{1/3} \sigma |\Omega_e|^{\kappa - 1/3} - AM^\kappa = 0. \tag{2.40}$$

Employing (2.38) in (2.39) yields

$$0 \le p_0 x_0^\kappa + (2/3) c_0 \sigma x_0^{\kappa - 1/3} - AM^\kappa \le (\kappa - 1)\varepsilon^2. \tag{2.41}$$

Calculating the coefficient in the second term in (2.40) we have

$$2(4\pi/3)^{1/3}\sigma = (2/3)c_0\sigma, \quad \text{because } c_0 = (36\pi)^{1/3}.$$

Hence (2.40) takes the form

$$p_0 |\Omega_e|^\kappa + (2/3) c_0 \sigma |\Omega_e|^{\kappa - 1/3} - AM^\kappa = 0. \tag{2.42}$$

Comparing (2.41) with (2.42) implies (2.37). This concludes the proof.

Considering variations of $|\Omega_t|$ near the volume $|\Omega_e|$ we choose a constant c_5 such that

$$\| |\Omega_e| - |\Omega_t| \| \le c_5 \varepsilon \quad \forall t \in \mathbb{R}_+, \tag{2.43}$$

where comparing (2.33) and (2.36)–(2.38) we have $c_3 + c_4 + c_5 \le c_1$.

Finally, (2.35) and (2.20) imply that for $t \ge 0$ the left-hand side of (2.20) must be smaller than $c_1\varepsilon^2$, so $1/2 \int_{\Omega_t} \rho v^2 \, dx \le c_1 \varepsilon^2$, the considered drop must be close to a ball, and the last term on the left-hand side of (2.20) that vanishes for a constant density is small, too.

To prove the global existence, we assume that Ω_t is diffeomorphic to a ball, so S_t can be described by

$$|x| \equiv r = R(\omega, t), \qquad \omega \in S^1, \tag{2.44}$$

where S^1 is the unit sphere.

LEMMA 2.4 (see the proof of Theorem 3 in [23]). *Let S_t be determined by* (2.44) *and suppose that the origin of coordinates coincides with the barycentre of Ω_t. Let $\rho(x,t)$ be a density defined for $x \in \Omega_t$, and let $t \in (T_1, T_2)$. Assume that there exists a maximum and a minimum of the density for $t \in (T_1, T_2)$ denoted by*

$$\rho_* = \min_{t \in (T_1, T_2)} \min_{\Omega_t} \rho(x,t), \qquad \rho^* = \max_{t \in (T_1, T_2)} \min_{\Omega_t} \rho(x,t). \tag{2.45}$$

Set $|\Omega^*| = M/\rho^*, |\Omega_*| = M/\rho^*, \bar{\rho}_t = M/|\Omega_t|$.

Then there exists a constant $\delta \in (0, 1/2)$ such that if

$$\sup_{S^1} |R(\omega,t) - R_t| + \sup_{S^1} |\nabla R| \le \delta R_t, \qquad t \in (T_1, T_2), \tag{2.46}$$

where $|\nabla R|^2 = R_\theta^2 + (\sin\theta)^{-2} R_\varphi^2$ in spherical coordinates, $R_t = ((3/4\pi)|\Omega_t|)^{1/3}$, then

$$\begin{aligned} &\int_{S^1} (|R(\omega,t) - R_t|^2 + |\nabla R(\omega,t)|^2) d\omega \\ &\qquad \le c_1(|S_t| - 4\pi R_t^2) + c_2 R_t^2 |\Omega_*|^{-2} (|\Omega_t| - |\Omega_*|)^2, \end{aligned} \tag{2.47}$$

where c_1, c_2 are constants that do not depend on δ and R_t.

Proof. To make the barycentre of Ω_t coincide with the origin of coordinates we must have (see Remark 2.6)

$$\int_{\Omega_t} \rho(x,t)x\,dx = \int_{S^1} d\omega \int_0^{R(\omega,t)} \rho(r,\omega,t)r^3\bar{\nu}(\omega)dr = 0, \tag{2.48}$$

where $\bar{\nu}(\omega) = (\cos\varphi\sin\theta, \sin\varphi\sin\theta, \cos\theta)$. We write (2.48) in the form

$$\frac{1}{\bar{\rho}_t}\int_{S^1} d\omega \int_0^{R(\omega,t)} (\rho(r,\omega,t) - \bar{\rho}_t)r^3\bar{\nu}(\omega)dr + \frac{1}{4}\int_{S^1}(R^4(\omega,t) - R_0^4)\bar{\nu}(\omega)d\omega = 0, \tag{2.49}$$

where the first expression can be estimated by

$$\begin{aligned}(\pi/\bar{\rho}_t)(\rho^* - \bar{\rho}_t)R^4(\omega,t) &\le \pi R^4(|\Omega_t|/M)(M/|\Omega_*| - M/|\Omega_t|)\\ &= (\pi R^4/|\Omega_*|)(|\Omega_t| - |\Omega_*|).\end{aligned}$$

From $|\Omega_t| = (4\pi/3)R_t^3$ we have

$$\int_{S^1}(R^3(\omega,t) - R_t^3)d\omega = 0. \tag{2.50}$$

From (2.48)–(2.50) we obtain

$$R_t^2\int_{S^1}(R - R_t)d\omega = -R_t\int_{S^1}(R-R_t)^2d\omega - \frac{1}{3}\int_{S^1}(R-R_t)^3d\omega, \tag{2.51}$$

(2.52)

$$\begin{aligned}R_t^3\int_{S^1}(R-R_t)\bar{\nu}(\omega)d\omega = &-\frac{3}{2}R_t^2\int_{S^1}(R-R_t)^2\bar{\nu}(\omega)d\omega - R_t\int_{S^1}(R-R_t)^3\bar{\nu}(\omega)d\omega\\ &-\frac{1}{4}\int_{S^1}(R-R_t)^4\bar{\nu}(\omega)d\omega + \frac{4}{\bar{\rho}_t}\int_{\Omega_t}(\rho - \bar{\rho}_t)x\,dx.\end{aligned}$$

To estimate $|S_t| - 4\pi R_t^2$ from below we use the formula

$$|S_t| - 4\pi R_t^2 = \int_{S^1}\left(R\sqrt{R^2 + |\nabla R|^2} - R_t^2\right)d\omega,$$

and we write the integrand in the form

(2.53)

$$\begin{aligned}&-R_t^2 + R\sqrt{R^2 + |\nabla R|^2}\\ &\quad = 2R_t(R-R_t) + \frac{1}{2}[2(R-R_t)^2 + |\nabla R|^2]\\ &\qquad + \frac{1}{2}\int_0^1 (1-s)^2\frac{d^3}{ds^3}\left[(R_t + s(R-R_t))\sqrt{R_t^2 + s(R-R_t)^2 + s^2|\nabla R|^2}\right]ds.\end{aligned}$$

Therefore,

$$|S_t| - 4\pi R_t^2 = -\int_{S^1}(R-R_t)^2d\omega + \frac{1}{2}\int_{S^1}|\nabla R|^2d\omega - \frac{2}{3R_t}\int_{S^1}(R-R_t)^3d\omega + I, \tag{2.54}$$

where I is the integral of the last term in (2.53).

Using the inequality

$$\int_{S^1} |f(\omega)|^2 d\omega \le \frac{1}{6} \int_{S^1} |\nabla f|^2 d\omega$$

for $f = a + \bar{b}\cdot\bar{\nu} + R - R_t$, where $a = (1/4\pi)\int_{S^1}(R-R_t)d\omega$, $\bar{b} = (3/4\pi)\int_{S^1}(R-R_t)\bar{\nu}d\omega$, we get

$$\|R - R_t\|^2_{0,S^1} \le (1/6)\|\nabla R\|^2_{0,S^1} + 4\pi(|a|^2 + |\bar{b}|^2) + |\bar{b}|^2\|\nabla\bar{\nu}\|^2_{0,S^1}. \tag{2.55}$$

Using (2.51) and (2.52) we obtain

$$\begin{aligned} &|a| \le c_3\delta\|R - R_t\|_{0,S^1}, \\ &|b| \le c_4\delta\|R - R_t\|_{0,S^1} + c_5(R_t/|\Omega_*|)(|\Omega_t| - |\Omega_*|) \end{aligned} \tag{2.56}$$

and

(2.57)

$$\frac{2}{3R_t}\int_{S^1} |R - R_t|^3 d\omega \le \frac{2}{3}\delta\|R - R_t\|^2_{0,S^1}, \qquad |I| \le c_6\delta(\|R - R_t\|^2_{0,S^1} + \|\nabla R\|^2_{0,S^1}).$$

From (2.55)–(2.57) for sufficiently small δ one gets (2.47). This concludes the proof.

The double mean curvature of S in spherical coordinates has the form

$$\mathcal{H} = \frac{1}{R\sin\theta}\left(\frac{\partial}{\partial\varphi}\frac{R_\varphi}{\sin\theta\sqrt{R^2 + |\nabla R|^2}} + \frac{\partial}{\partial\theta}\frac{\sin\theta R_\theta}{\sqrt{R^2 + |\nabla R|^2}}\right) - \frac{2}{\sqrt{R^2 + |\nabla R|^2}}. \tag{2.58}$$

Now we consider the equation

$$\mathcal{H}(R) + \frac{2}{R_t} = h(\omega), \tag{2.59}$$

where $h(\omega) = \bar{n}\mathbb{T}(v, p_\sigma)\bar{n}|_S$.

From Theorem 4 in [14] we have the following.

THEOREM 2.5. *Let $R \in H_2^{k+5/2}(S^1)$, $k \in \mathbb{Z}_+ \cup \{0\}$ be a solution to* (2.59) *that satisfies* (2.46) *with sufficiently small δ. If $h \in H_2^{k+1/2}(S^1)$, then*

$$\|R - R_t\|_{2+\mu,S^1} \le C_1\|h\|_{\mu,S^1} + c_2\|R - R_t\|_{0,S^1}, \qquad \mu = k + 1/2. \tag{2.60}$$

To guarantee that the barycentre of Ω_t coincides with the origin of coordinates, we need the following.

Remark 2.6. Assume $f = 0$ and

$$\int_\Omega \rho_0 v_0\, dx = 0, \qquad \int_\Omega \rho_0 \xi d\xi = 0. \tag{2.61}$$

Then (2.2) and (2.61) imply $\int_{\Omega_t} \rho v\, dx = 0$, and then (2.3) gives (2.48).

Now we formulate the main result, which is necessary in the proof of global existence.

Remark 2.7. Let the assumptions of Lemma 2.2 be satisfied. Let $|\Omega^{**}| = \max_t |\Omega_t|$, $|\Omega_{**}| = \min_t |\Omega_t|$, $\psi^{**} = \max_t \psi_t$, and $\psi_{**} = \min_t \psi_t$. Let $|S_{**}| = 4\pi R^2_{**}$,

where R_{**} is determined by $(4\pi/3)R_{**}^3 = |\Omega_{**}|$. Then $|S_t| - |S_{**}| \geq 0$. Moreover, by Lemma 2.2 we have $|\Omega^{**}| - |\Omega_{**}| \leq \varepsilon$, $\psi^{**} - \psi_{**} \leq \varepsilon$. Furthermore, by writing (2.1) in the form

$$\frac{d}{dt}\left(\int_{\Omega_t} \frac{1}{2}\rho v^2 dx + \psi_t + p_0|\Omega_t| + \sigma|S_t|\right) \leq 0, \tag{2.62}$$

we obtain

$$\begin{aligned} &\frac{1}{2}\int_{\Omega_t} \rho v^2 dx + \psi_t - \psi_{**} + p_0(|\Omega_t| - |\Omega_{**}|) + \sigma(|S_t| - |S_{**}|) \\ &\quad \leq \frac{1}{2}\int_{\Omega_t} \rho_0 v_0^2 dx + \psi - \psi_{**} + p_0(|\Omega| - |\Omega_{**}|) + \sigma(|S| - |S_{**}|) \leq \kappa_0 \varepsilon_0, \end{aligned} \tag{2.63}$$

where $\psi = \psi_0, \varepsilon_0 = \varepsilon_0(\varepsilon)$ will be made arbitrarily small.

Since the minima of $|\Omega_t|, |S_t|, \psi_t$, for all $t \in \mathbb{R}_+$, exist, we can always obtain (2.63) with an arbitrarily small right-hand side because we can choose ε in Lemma 2.2 so small that the right-hand side of (2.63) is as small as we please. To ensure the last inequality we have to assume

$$\frac{1}{2}\int_{\Omega} \rho_0 v_0^2 dx \leq \varepsilon. \tag{2.64}$$

Moreover, the coefficient κ_0 in (2.63) will be chosen in such a way that Lemma 2.4 yields

$$\frac{1}{2}\int_{\Omega_t} \rho v^2 dx + \|R(\omega, t) - R_t\|_{1,S^1}^2 \leq \varepsilon_0. \tag{2.65}$$

Finally, we consider the case $p_0 = 0$. Then, instead of (2.9), we obtain

$$\begin{aligned} &\frac{1}{2}\int_{\Omega_t} \rho v^2 dx + \int_{\Omega_t} \varphi(\rho) dx + \sigma|S_t| \\ &\quad \leq \frac{1}{2}\int_{\Omega} \rho_0 v_0^2 dx + \int_{\Omega} \varphi(\rho_0) dx + \sigma|S| \equiv d_0. \end{aligned} \tag{2.66}$$

Hence instead of (2.13) we have

$$\begin{aligned} &y_1(|\Omega_t|) + |\Omega_t|^{\kappa-1}\int_{\Omega_t} \frac{1}{2}\rho v^2 dx + \sigma(|S_t| - 4\pi R_t^2)|\Omega_t|^{\kappa-1} \\ &\quad + \frac{A}{\kappa - 1}\left(|\Omega_t|^{\kappa-1}\int_{\Omega_t} \rho^\kappa dx - \left(\int_{\Omega_t} \rho dx\right)^\kappa\right) \leq 0, \end{aligned} \tag{2.67}$$

where

$$y_1(x) = c_0 \sigma x^{\kappa - 1/3} - d_0 x^{\kappa-1} + \frac{A}{\kappa - 1} M^\kappa, \qquad x := |\Omega_t|. \tag{2.68}$$

Inequality (2.67) holds for physical volumes $|\Omega_t|$. Since the last three terms in (2.67) are positive, we have

$$y_1(|\Omega_t|) \leq 0. \tag{2.69}$$

An extremum point x_0 of the function $y_1 = y_1(x)$ calculated from the relation

$$y_1'(x) \equiv c_0\sigma(\kappa - 1/3)x^{\kappa-4/3} - d_0(\kappa-1)x^{\kappa-2} = 0 \tag{2.70}$$

is

$$x_0 = \left(\frac{d_0(\kappa-1}{c_0\sigma(\kappa-1/3)}\right)^{3/2}. \tag{2.71}$$

This is a positive number. From the form of $y_1 = y_1(x)$ we have $y_1(0) = AM^\kappa/(\kappa - 1) > 0$, $y_1(\infty) = \infty$, $y_1'(x)|_{x<x_0} < 0$, $y_1'(x)|_{x>x_0} > 0$, so x_0 is a minimum point. We also have

$$y_1''(x_0) = (2/3)d_0(\kappa-1)x_0^{\kappa-3} > 0. \tag{2.72}$$

From (2.69) it follows that there are two strictly positive solutions of the equation $y_1(x) = 0$. Denote them by w_1, w_2. We have $0 < w_1 < x_0 < w_2 < \infty$.

To guarantee (2.69) we have to require that

$$y_1(x_0) = -\frac{2}{3}\left(\frac{d_0}{\kappa-1/3}\right)^{(3\kappa-1)/2}\left(\frac{\kappa-1}{c_0\sigma}\right)^{3(\kappa-1)/2} + AM^\kappa/(\kappa-1) < 0. \tag{2.73}$$

Using the form of d_0 from (2.66) in (2.73) yields

$$\begin{aligned} y_1(x_0) = &-\frac{2}{3}(\kappa-1)^{3(\kappa-1)/2}(\kappa-1/3)^{(3\kappa-1)/2}(c_0\sigma)^{-3(\kappa-1)/2} \\ &\cdot\left(\frac{1}{2}\int_\Omega \rho_0 v_0^2 dx + \frac{A}{\kappa-1}\int_\Omega \rho_0^\kappa dx + \sigma|S|\right)^{(3\kappa-1)/2} + AM^\kappa/(\kappa-1) < 0. \end{aligned} \tag{2.74}$$

Since $M^\kappa \le |\Omega|^{\kappa-1}\int_\Omega \rho_0^\kappa dx$, we see that (2.74) may hold.

Now we show that for a given $\varepsilon > 0$ there exists a set of parameters κ, d_0, A, M such that meas $X \equiv |X| \le \varepsilon$, where $X = \{x \in \mathbb{R}_+ : y_1(x) < 0, \text{or } w_1 < x < w_2\}$. Moreover, $x_0 \in X$. To calculate meas X in the case of small ε we estimate solutions to the equation

$$y_1(x) = c_0\sigma x^{\kappa-1/3} - d_0 x^{\kappa-1} + AM^\kappa/(\kappa-1) = 0. \tag{2.75}$$

We expand the functions $x^{\kappa-1}$ and $x^{2/3}$ near the minimum point x_0 in the form

$$\begin{aligned} x^{\kappa-1} &= x_0^{\kappa-1} + (\kappa-1)x_0^{\kappa-2}h + (1/2)(\kappa-1)(\kappa-2)x^{\kappa-3}(\tilde{x}_1)h^2, \\ x^{2/3} &= x_0^{2/3} + (2/3)x_0^{-1/3}h - (1/9)\tilde{x}_2^{-4/3}h^2, \end{aligned} \tag{2.76}$$

where $(x_0 - h, x_0 + h) \subset (w_1, w_2)$ and $\tilde{x}_i \in (x_0 - h, x_0 + h)$, $i = 1, 2$.

Assume that h is small. Put (2.76) into (2.75) and take into account terms up to the second-order with respect to h only. Then we obtain

$$\left[c_0\sigma\left(\frac{\kappa^2}{2} - \frac{5}{6}\kappa + \frac{2}{9}\right)x_0 - d_0\frac{(\kappa-1)(\kappa-2)}{2}\right]x_0^{\kappa-3}h^2 \cong -y(x_0). \tag{2.77}$$

Using (2.71) in (2.77) yields

$$|h| \le (-3y(x_0)/(d_0(\kappa-1)x_0^{\kappa-1}))^{1/2}. \tag{2.78}$$

Now we impose restrictions on the parameters $\kappa, \rho_0, v_0, \Omega, A, \sigma, S, M$ such that for a given $\varepsilon > 0$ we have $y_1(x_0) < 0$ and $-y_1(x_0) < \varepsilon^2$. Hence by (2.78) we have $|h| \le (3/(d_0(\kappa-1)x_0^{\kappa-1}))^{1/2}\varepsilon \equiv c_1\varepsilon$, where ε may be assumed to be as small as we need. Thus we have shown the following.

LEMMA 2.8. *Let $\varepsilon > 0$. Then there exists a set of parameters $\kappa, \rho_0, v_0, \Omega, S, \sigma, A, M$ such that $y_1(x_0) < 0, |y_1(x_0)| \le \varepsilon$ and*

$$\sup_t \operatorname{var} |\Omega_t| \le c_1\varepsilon, \qquad t > 0, \tag{2.79}$$

and there exists a constant c_2 such that

$$\sup_t \operatorname{var} \psi_t \le c_2\varepsilon, \qquad t > 0. \tag{2.80}$$

3. Local existence. To prove the local existence of solutions to (1.1a–e) we write it in the Lagrangian coordinates introduced by (1.5) and (1.6):

$$\begin{aligned}
\eta u_t - \mu\nabla_u^2 u - \nu\nabla_u\nabla_u\cdot u + \nabla_u q &= \eta g && \text{in } \Omega^T,\\
\eta_t + \eta\nabla_u\cdot u &= 0 && \text{in } \Omega^T,\\
\mathbb{T}_u(u,q)\bar{n} - \sigma\Delta_{S_t}(t)X_u(\xi,t) &= -p_0\bar{n} && \text{on } S^T,\\
u|_{t=0} &= v_0(\xi) && \text{in } \Omega,\\
\eta|_{t=0} &= \rho_0(\xi) && \text{in } \Omega,
\end{aligned} \tag{3.1}$$

where $u(\xi,t) = v(X_u(\xi,t),t)$, $\eta(\xi,t) = \rho(X_u(\xi,t),t)$, $q(\xi,t) = p(X_u(\xi,t),t)$, $g(\xi,t) = f(X_u(\xi,t),t)$,

$$\nabla_u = \partial_x\xi^i\nabla_{\xi^i}, \quad \nabla_{\xi^i} = \partial_{\xi^i}, \quad \mathbb{T}_u(u,q) = -qI + \mathbb{D}_u(u) \quad \text{and}$$
$$\mathbb{D}_u(u) = \left\{\mu\left(\partial_{x^i}\xi^k\nabla_{\xi^k}u^j + \partial_{x^j}\xi^k\nabla_{\xi^k}u^i\right) + (\nu-\mu)\delta_{ij}\nabla_u\cdot u\right\},$$

$\nabla_u\cdot u = \partial_{x^i}\xi^k\nabla_{\xi^k}u^i$. Let A be the Jacobi matrix of the transformation $x = x(\xi,t) = X_u(\xi,t)$ with elements $a_{ij} = \delta_{ij} + \int_0^t \partial_{\xi^j}u^i(\xi,\tau)d\tau$. Assuming $|\nabla_\xi u|_{\infty,\Omega^T} \le M$, we obtain

$$0 < c_1(1-Mt)^3 \le \det\left\{\partial_{\xi^j}x^i\right\} \le c_2(1+Mt)^3, \qquad t \le T, \tag{3.2}$$

where c_1, c_2 are constants and T is sufficiently small. Moreover, $\det A = \exp(\int_0^t \nabla_u\cdot u\,d\tau) = \rho_0/\eta$.

Let S_t be determined (at least locally) by the equation $\phi(x,t) = 0$. Then S is described by $\phi(x(\xi,t),t)|_{t=0} \equiv \tilde{\phi}(\xi) = 0$. Moreover, we have

$$\bar{n} = \bar{n}(x(\xi,t),t) = \frac{\nabla_x\phi(x,t)}{|\nabla_x\phi(x,t)|}|_{x=x(\xi,t)} \quad \text{and} \quad \bar{n}_0 = \bar{n}_0(\xi) = \frac{\nabla_\xi\tilde{\phi}(\xi)}{|\nabla_\xi\tilde{\phi}(\xi)|}.$$

THEOREM 3.1. *Let $v_0 \in W_2^{l+1}(\Omega)$, $\rho_0 \in W_2^{l+1}(\Omega)$, $f \in C^{l+1}(\mathbb{R}^3\times(0,T))$, $S \in W_2^{l+5/2}$, $l \ge \frac{3}{2}$ and $\frac{l}{2} - \frac{3}{4} - \kappa \in \mathbb{N} \subset \{0\}$, $\kappa \in (0,\frac{1}{4})$. Let $G = G(t, t\tilde{A}, \alpha, \beta, \gamma)$ be an increasing positive function of its arguments (its definition is given by (3.55) in [36]), where $\alpha = \|\rho_0\|_{l+1,\Omega}, \beta = \|f\|_{C^{l+1}(\mathbb{R}^3\times(0,T))}$, $\gamma = |u(0)|_{l+1,0,\Omega}$, which is such*

that $G(0,0,\alpha,\beta,\gamma) > 0$. *Suppose that* $\tilde{A} > G(0,0,\alpha,\beta,\gamma)$. *Let* $|v(0)|_{l+1,0,\Omega} \le \tilde{A}$. *Let* δ_1 *be sufficiently small (see the proof of Lemma* 3.3 *in* [36]). *Let* T_* *be so small that*

$$T_*^a \tilde{A}\varphi_1(T_*, T_*^a\tilde{A}, \tilde{A}, \tilde{A}) \le \delta_1,$$
$$0 < c_1(1-\tilde{A}T_*)^3 \le \det\left\{\frac{\partial x}{\partial \xi}\right\} \le c_2(1+\tilde{A}T_*)^3,$$

where φ_1 *is an increasing positive function which is defined in the assumptions of Lemma* 3.3 *from* [36], $x(\xi,t) = \xi + \int_0^t v_0(\xi,\tau)d\tau$, $t \le T_*$, $G(T_*, T_*^a\tilde{A}, \alpha, \beta, \gamma) \le \tilde{A}$, $a > 0$ *and* $\tilde{v}_0(\xi,t)$ *is defined in the proof of Theorem* 3.6 *from* [36].

Then there exists $T_{**}, 0 < T_{**} \le T_*$ *such that for* $T \le T_{**}$ *there exists a unique solution to problem* (3.1) *such that* $u \in W_2^{l+2,l/2+1}(\Omega^T)$, $\eta \in W_2^{l+1,l/2+1/2}(\Omega^T) \cap C([0,T]; \Gamma_0^{l+1,(l+1)/2}(\Omega))$ *and*

$$\|u\|_{l+2,\Omega^T} \le \tilde{A}, \qquad \|\eta\|_{l+1,\Omega^T} + \|\eta\|_{l+1,\infty,\Omega^T} + \|1/\eta\|_{l+1,\Omega^T} \le (\|\rho_0\|_{l+1,\Omega} + \|1/\rho_0\|_{l+1,\Omega})\,\varphi_2(T, T^a\tilde{A}), \tag{3.3}$$

where φ_2 *is an increasing positive function defined in the theses of Lemma* 3.5 *in* [36].

Having shown the local existence of solutions to (3.1), we find a more appropriate estimate that will be useful in the proof of global existence. Let us recall that $R_t = \left(\frac{3}{4}\pi|\Omega_t|\right)^{1/3}$, $t \ge 0$. In view of Definition 1.1, we shall look for motions of (1.1a–e) which are close to the equilibrium state. Assuming that the initial motion is sufficiently close to the equilibrium state, we introduce the quantity $q_\sigma = q - p_0 - q_0$, where $q_0 = 2\sigma/R_0$. The quantity describes changes of the pressure near the sum of external pressure p_0 and the initial pressure of surface tension in the case when, at the initial moment, our drop is a ball (q_0). Therefore, we consider

$$\begin{aligned}
&\eta u_t - \mu\nabla_u^2 u - \nu\nabla_u\nabla_u\cdot u = -\nabla_u q_\sigma + \eta g,\\
&\Pi_0\Pi\mathbb{D}_u(u)\bar{n} = 0,\\
&\bar{n}_0\mathbb{D}_u(u)\bar{n} - \sigma\Delta_{S_t}(t)\int_0^t u(\tau)d\tau\cdot\bar{n}_0 = \bar{n}_0\cdot\bar{n}q_\sigma + \sigma\bar{n}_0\cdot(\Delta_{S_t}(t) - \Delta_S(0))\,\xi\\
&\qquad\qquad + \sigma\left(H(\xi,0) + \frac{2}{R_0}\right),\\
&u|_{t=0} = v_0,
\end{aligned} \tag{3.4}$$

where $\Pi_0 g = g - g\cdot\bar{n}_0\bar{n}_0$, $\Pi g = g - g\cdot\bar{n}\bar{n}$, and

$$\begin{aligned}
&q_{\sigma t} = -q_\sigma\Psi(\eta)\mathrm{div}_u u - (p_0+q_0)\mathrm{div}_u u,\\
&q_\sigma|_{t=0} = p(\rho_0) - p_0 - q_0,
\end{aligned} \tag{3.5}$$

where $\Psi(\eta) = p_\eta(\eta)\eta/p(\eta)$, $p_\eta = \partial_\eta p$.

By Theorem 3.1 we have the existence of solutions to (3.4) and (3.5). Moreover, we obtain the following.

Remark 3.2. Let u, η be solutions of problem (1.1a–e). Then from (3.4) and (3.5)

for sufficiently small T we obtain the estimate

$$
\begin{aligned}
&\|u\|_{l+2,\Omega^T} + \|q_\sigma\|_{l+1,\Omega^T} + |q_\sigma|_{l+1,0,\infty,\Omega^T} \\
&\quad \le \varphi_3\left(T, \|v_0\|_{l+1,\Omega}, \|\rho_0\|_{l+1,\Omega}, \|f\|_{C^{l+1}(\mathbb{R}^3\times(0,T))}, \|S\|_{W_2^{l+5/2}}\right) \\
&\qquad \left[\|f\|_{C^{l+1}(\mathbb{R}^3\times(0,T))} + \|v_0\|_{l+1,\Omega}\right. \\
&\qquad \left. + \|p(\rho_0) - p_0 - q_0\|_{l+1,\Omega} + \|H(\xi,0) + 2/R_0\|_{l+1/2,S}\right],
\end{aligned} \tag{3.6}
$$

where $1 > \frac{3}{2}$. The existence of solution v, p of (1.1a–e) such that $u \in W_2^{4,2}(\Omega^T)$, $q_\sigma \in W_2^{3,3/2}(\Omega^T) \cap C(0,T;\Gamma_0^{3,3/2}(\Omega))$ and estimate (3.6) for $l = 2$ are proved in [37].

Proof. Applying Lemma 3.3 from [36] to (3.4) yields

$$
\begin{aligned}
\|u\|_{l+2,\Omega^T} \le c\left(T, \tilde{A}, \alpha, \beta, \gamma, \|S\|_{W_2^{l+5/2}}\right) &\left[\|q_\sigma\|_{l+1,\Omega^T} + \|g\|_{l,\Omega^T}\right. \\
&\left. + \|H(\xi,0) + 2/R_0\|_{l+1/2,S}\right].
\end{aligned} \tag{3.7}
$$

Integrating (3.5) implies

$$
\begin{aligned}
q_\sigma(\xi,t) = &-\exp\left[-\int_0^t \Psi(\eta)\mathrm{div}_u u dt'\right] \\
&\cdot \left[\int_0^t \left[(p_0+q_0)\mathrm{div}_u u \exp\int_0^{t'} \Psi(\eta)\mathrm{div}_u u dt''\right] dt' + p(\rho_0) - p_0 - q_0\right].
\end{aligned} \tag{3.8}
$$

From (3.8) we have

$$
\begin{aligned}
&\|q_\sigma\|_{l+1,\Omega^T} + |q_\sigma|_{l+1,0,\infty,\Omega^T} \le c(T, \tilde{A}, \alpha, \beta, \gamma, \|S\|_{W_2^{l+5/2}}) \\
&\quad \cdot (T^a\|u\|_{l+2,\Omega^T} + \|p(\rho_0) - p_0 - q_0\|_{l+1,\Omega}), \qquad a > 0.
\end{aligned} \tag{3.9}
$$

From (3.7) and (3.9) for sufficiently small T we have (3.6). This concludes the proof.

4. Global differential inequality. Assume that we have proved the existence of a sufficiently smooth local solution. First we find a special differential inequality that enables us to prove the existence of a solution by energy estimates and then to prove global existence.

To show it we consider the motion near the equilibrium state $v_e = 0$, $p_e = p_0 + 2\sigma/R_0$, $\rho_e = M/((4\pi/3)R_0^3)$; R_0 is a solution of the equation $p_0(M/((4\pi/3)R_0^3)) = p_0 + 2\sigma/R_0$, $q_0 = 2\sigma/R_0$ and $p_\sigma = p - p_0 - q_0$. Therefore, we examine the following system:

$$
\rho(v_t^i + v\cdot\nabla v^i) - \partial_{x^j} T_{ij}(v, p_\sigma) = \rho f^i \text{ in } \Omega_t, t \in [0,T], \tag{4.1a}
$$

$$
\rho_t + \mathrm{div}(\rho v) = 0 \text{ in } \Omega_t, t \in [0,T], \tag{4.1b}
$$

$$
\mathbb{T}(v, p_\sigma)\bar{n} = \sigma\Delta_{S_t} x\cdot\bar{n}\bar{n} + q_0\bar{n} \text{ on } S_t, t \in [0,T], \tag{4.1c}
$$

where $\mathbb{T} = \{T_{ij}\} = \{\mu(\partial_{x^j} v^i + \partial_{x^i} v^j) + (\nu - \mu)\delta_{ij}\mathrm{div}\, v - p_\sigma\delta_{ij}\}$.

Using the barotropic law $p = p(\rho)$ we write (4.1b) in the form

$$
p_{\sigma t} + v\cdot\nabla p_\sigma + p\Psi(\rho)\mathrm{div}\, v = 0, \tag{4.2}
$$

where $\Psi(\rho) = p_\rho\rho/p$.

Set $\rho_* = \min_{\tilde{\Omega}^T} \rho(x,t)$, $\rho^* = \max_{\tilde{\Omega}^T} \rho(x,t)$.

Now we point out the following facts concerning the estimates in Lemmas 4.1–4.12 and Theorems 4.13 and 4.14.

(1) The numbers δ_i are assumed to be small and are separately numbered in each lemma.

(2) We distinguish absolute constants, denoted by c, which may depend on such parameters of the problem as $\mu, \nu, \kappa, \sigma, A$ and which are coefficients in those terms in the right-hand sides of the inequalities that contain the highest derivatives only, and are finally balanced by the left-hand side main terms after appropriate summing.

(3) We distinguish the coefficients by the lower-order terms, nonlinear terms, and also by the force terms which depend on ρ_*, ρ^*, T,

$$a \equiv \int_0^T \|v\|_{3,\Omega_{t'}} dt', \quad a_0(t) \equiv \left| \int_0^t v_x d\tau \right|_{\infty,\Omega}, \quad b \equiv \|S\|_{4+1/2}, M, p_0,$$

on the parameters that guarantee the existence of the inverse transformation to $x = x(\xi,t)$, and also on the constants of imbedding theorems considered over Ω_t. Generally, the coefficients are increasing functions of the parameters. In the statements of the lemmas, we denote such coefficients by $P_1, P_2, \ldots,$ (common numbering for all lemmas) and independently in each lemma by $a_1, a_2, \ldots$. Moreover, P_i, a_i are positive and increasing functions of a and b.

(4) We have to underline that all estimates in this section are obtained under the assumption that there exists a local solution of (1.1a–e) so that all the quantities ρ_*, ρ^*, T, a, b are estimated by the data functions. Moreover, the local solution guarantees the existence of the inverse transformation to $x = x(\xi,t)$. Generally, the quantities $\rho_*, \rho^*, a, b, M, p_0$ might be large.

LEMMA 4.1. *Let v, p_σ be a sufficiently smooth solution of* (4.1). *Then*

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v^2 + \frac{1}{p\Psi(\rho)}p_\sigma^2\right)dx + \frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}\bar{n}\cdot\int_0^t v_{s^\alpha}d\tau\bar{n}\cdot\int_0^t v_{s^\beta}d\tau\,ds \\ &\quad + \frac{\mu}{2}\|v\|^2_{1,\Omega_t} + (\nu-\mu)\|\operatorname{div} v\|^2_{0,\Omega_t} \\ &\le \varepsilon_1\left(\|p_\sigma\|^2_{0,\Omega_t} + \left\|\int_0^t v_s d\tau\right\|^2_{0,S_t} + \left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{0,S^1}\right) \\ &\quad + P_1\left(\|v\|^2_{0,\Omega_t} + \|f\|^2_{0,\Omega_t}\right) + P_2 X_1 Y_1, \end{aligned} \tag{4.3}$$

where $\varepsilon_1 \in (0,1)$, $P_i = P_i^*(\rho_*, \rho^*, a_0(t))$, $i = 1,2$, and

$$\begin{aligned} X_1 &= \|v\|^2_{2,\Omega_t} + \|p_\sigma\|^2_{1,\Omega_t}, \\ Y_1 &= X_1 + \left\|\int_0^t v d\tau\right\|^2_{2,S_t}. \end{aligned} \tag{4.4}$$

Proof. Multiplying (4.1a) by v, integrating over Ω_t, and using (4.1b, c) implies

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\rho v^2 dx + \frac{\mu}{2}E_{\Omega_t}(v) + (\nu-\mu)\|\operatorname{div} v\|^2_{0,\Omega_t} - \int_{\Omega_t} p_\sigma \operatorname{div} v\,dx \\ &\quad -\sigma\int_{S_t}\left(\Delta_{S_t}x\cdot\bar{n} + \frac{2}{R_0}\right)\bar{n}\cdot v\,ds = \int_{\Omega_t}\rho f\cdot v\,dx. \end{aligned} \tag{4.5}$$

Equation (4.2) yields

$$-\int_{\Omega_t} p_\sigma \operatorname{div} v\, dx = \int_{\Omega_t} \frac{1}{p\Psi(\rho)}(\partial_t + v\cdot\nabla)\frac{p_\sigma^2}{2}\, dx$$

and

$$[F_t + \operatorname{div}(Fv) + (F_\rho \rho - F)\operatorname{div} v]\frac{p_\sigma^2}{2} = 0, \tag{4.6}$$

where $F = 1/(p(\rho)\Psi(\rho))$; so

$$\begin{aligned}\int_{\Omega_t} p_\sigma \operatorname{div} v\, dx &= -\frac{d}{dt}\int_{\Omega_t}\frac{1}{p\Psi(\rho)}\frac{p_\sigma^2}{2}dx + \int_{\Omega_t}(F - F_\rho\rho)\operatorname{div} v\frac{p_\sigma^2}{2}dx \\ &\le -\frac{d}{dt}\int_{\Omega_t}\frac{1}{p\Psi(\rho)}\frac{p_\sigma^2}{2}dx + \delta_1\|p_\sigma\|^2_{0,\Omega_t} \\ &\quad + a_1(\rho_*,\rho^*,\delta_1)\|p_\sigma\|^2_{1,\Omega_t}\|v\|^2_{2,\Omega_t}.\end{aligned} \tag{4.7}$$

In view of Lemma 5.2 from [35] and the relation $p(\rho) - p_0 = p - p(\rho_e) = p'(\tilde\rho)(\rho - \rho_e)$, $\tilde\rho \in [\rho, \rho_e]$, we have

$$\|v\|^2_{1,\Omega_t} \le c_2(\rho^*)(E_{\Omega_t}(v) + \|p_\sigma\|^2_{0,\Omega_t}\|v\|^2_{2,\Omega_t}). \tag{4.8}$$

By the Hölder and Young inequalities the right-hand side of (4.5) is estimated by

$$\delta_2\|v\|^2_{0,\Omega_t} + c(\delta_2)\rho_*^2\|f\|^2_{0,\Omega_t}. \tag{4.9}$$

Now we consider the boundary term in (4.5). By exploiting the Lagrangian coordinates we express S_t as follows: $x(s^1,s^2,t) = \xi(s^1,s^2) + \int_\sigma^t u(\xi(s^1,s^2),\tau)d\tau$, where $\{s^1,s^2\} \in U \subset \mathbb{R}^2$, so the boundary term takes the form

(4.10)

$$\begin{aligned}-\int_{S_t}\left(\Delta_{S_t}x\cdot\bar n + \frac{2}{R_0}\right)v\cdot\bar n ds &= -\int_U\left(\partial_{s^\alpha}\left(g^{\alpha\beta}\sqrt{g}x_\beta\right)\cdot\bar n + \frac{2}{R_0}\sqrt{g}\right)v\cdot\bar n\, ds^1ds^2 \\ &= -\int_U\left(\partial_{s^\alpha}\left(g^{\alpha\beta}\sqrt{g}\xi_{s^\beta}\right)\cdot\bar n + \frac{2}{R_0}\sqrt{g}\right)u\cdot\bar n\, ds^1ds^2 \\ &\quad -\int_U\partial_{s^\alpha}\left(g^{\alpha\beta}\sqrt{g}\int_0^t u_{s^\beta}d\tau\right)\cdot\bar n u\cdot\bar n\, ds^1ds^2.\end{aligned}$$

The first term we write is the following:

$$\begin{aligned}-\int_U\Big[&(g^{\alpha\beta}\sqrt{g})_{,s^\alpha}\,\xi_{s^\beta}\cdot(\bar n - \bar n_0) + (g^{\alpha\beta} - g^{\alpha\beta}(0))\sqrt{g}\xi_{s^\alpha s^\beta}\cdot\bar n \\ &+ g^{\alpha\beta}(0)\sqrt{g}\xi_{s^\alpha s^\beta}\cdot(\bar n - \bar n_0) + g^{\alpha\beta}(0)\sqrt{g}\xi_{s^\alpha s^\beta}\cdot\bar n_0 + \frac{2}{R_0}\sqrt{g}\Big]u\cdot\bar n\, ds^1ds^2 \\ &= -\int_{S_t}\left(H(0) + \frac{2}{R_0}\right)v\cdot\bar n\, ds + N_1,\end{aligned}$$

where $H(0) = g^{\alpha\beta}(0)\xi_{s^\alpha s^\beta}\cdot\bar n_0$, $\bar n = (x_{s^1}\times x_{s^2})/|x_{s^1}\times x_{s^2}|$, $\bar n_0 = (\xi_{s^1}\times\xi_{s^2})/|\xi_{s^1}\times\xi_{s^2}|$, $g_{\alpha\beta} = x_{s^\alpha}\cdot x_{s^\beta}$, $g_{\alpha\beta}(0) = \xi_{s^\alpha}\cdot\xi_{s^\beta}$ and

$$|N_1| \le \delta_3\|\int_0^t v_s d\tau\|^2_{0,S_t} + \delta_4\|v\|^2_{1,\Omega_t} + a_2\|v\|^2_{0,\Omega_t}.$$

The second term in the right-hand side of (4.10) takes the form

$$\frac{1}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}\bar{n}\cdot\int_0^t v_{s^\alpha}\,d\tau\bar{n}\cdot\int_0^t v_{s^\beta}\,d\tau\,ds + N_2,$$

where

$$|N_2| \le \delta_5\left(\left\|\int_0^t v_s d\tau\right\|^2_{0,S_t} + \|v\|^2_{1,\Omega_t}\right) + a_3\|v\|^2_{0,\Omega_t} + a_4\|\int_0^t v d\tau\|^2_{2,S_t}\|v\|^2_{2,\Omega_t}.$$

Hence, taking $\delta_i, i = 1,\dots,5$, sufficiently small and using (4.7) and (4.10) in (4.5) we obtain (4.3). This concludes the proof.

LEMMA 4.2. *For a sufficiently smooth solution of* (4.1) *we have*

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v_t^2 + \frac{1}{p\Psi(\rho)}p_{\sigma t}^2\right)dx + \frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}v_{s^\alpha}\cdot\bar{n}v_{s^\beta}\cdot\bar{n}\,ds \\
&\qquad + \frac{\mu}{4}\|v_t\|^2_{1,\Omega_t} + (\nu-\mu)\|\operatorname{div} v_t\|^2_{0,\Omega_t} \\
&\le \varepsilon_2\left(\|p_{\sigma t}\|^2_{0,\Omega_t} + \|v_{xx}\|^2_{0,\Omega_t}\right) + P_3(\rho_*,\rho^*,\varepsilon_2)X_2Y_2 \\
&\qquad + P_4\left(\|f\|^2_{1,0,\Omega_t} + \|v\|^2_{0,\Omega_t}\right),
\end{aligned} \tag{4.11}$$

where $\varepsilon_2 \in (0,1)$, $P_3(\varepsilon_2)$ *behaves like* ε_2^{-a}, $a > 0$, *and*

$$Y_2 = X_2 = |p_\sigma|^2_{2,1,\Omega_t} + |v|^2_{2,1,\Omega_t}. \tag{4.12}$$

Proof. Differentiating (4.1a) with respect to t, multiplying by v_t, and integrating over Ω_t yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\rho v_t^2 dx + \frac{\mu}{2}E_{\Omega_t}(v_t) + (\nu-\mu)\|\operatorname{div} v_t\|^2_{0,\Omega_t} - \int_{\Omega_t} p_{\sigma t}\operatorname{div} v_t dx \\
&\qquad - \int_{S_t} T(v,p_\sigma)_{,t}\cdot\bar{n}\cdot v_t ds = N_1,
\end{aligned} \tag{4.13}$$

where

$$N_1 \le \delta_1\|v_t\|^2_{0,\Omega_t} + c(\delta_1)[\|f\|^2_{1,\Omega_t} + X_2^2] + c(\delta_1,\rho^*)\|f_t\|^2_{0,\Omega_t}.$$

From (4.2) and (4.6) with $p_{\sigma t}$ in place of p_σ we obtain

$$\begin{aligned}
\int_{\Omega_t} p_{\sigma t}\operatorname{div} v_t dx \le -\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\frac{1}{p\Psi(\rho)}p_{\sigma t}^2\,dx + \delta_2\|p_{\sigma t}\|^2_{0,\Omega_t} \\
+ c(\rho_*,\rho^*,\delta_2)X_2^2.
\end{aligned} \tag{4.14}$$

The boundary term in (4.13) is equal to

$$-\int_{S_t} T(v,p_\sigma)_{,t}\cdot\bar{n}\cdot v_t\,ds = -\sigma\int_{S_t}\Delta_{S_t}v\cdot\bar{n}\,v_t\cdot\bar{n}\,ds + N_2, \tag{4.15}$$

where

$$|N_2| \le \delta_3(\|v_t\|^2_{1,\Omega_t} + \|v_{xx}\|^2_{0,\Omega_t}) + a_1\|v\|^2_{0,\Omega_t}.$$

Next, using the Lagrangian coordinates we have

$$\begin{aligned} -\sigma\int_{\Delta_{S_t}} \Delta_{s_t} v\cdot\bar n\, v_t\cdot\bar n\, ds &= -\sigma\int_U \partial_{s^\alpha}\left(g^{\alpha\beta}\sqrt{g}\partial_{s^\beta} v\right)\cdot\bar n\, v_t\cdot\bar n\, ds^1 ds^2 \\ &= \frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta} v_{s^\alpha}\cdot\bar n\, v_{s^\beta}\cdot\bar n\, ds + N_3, \end{aligned} \tag{4.16}$$

and

$$|N_3| \le \delta_4\left(\|v_t\|^2_{1,\Omega_t} + \|v_{xx}\|^2_{0,\Omega_t}\right) + a_2\|v\|^2_{0,\Omega_t} + a_3 X_2^2.$$

Finally, from (4.13)–(4.16) and Lemma 5.3 from [35] we obtain (4.11) for sufficiently small δ's. This concludes the proof.

From Lemmas 4.1 and 4.2 we have the following.

LEMMA 4.3.

(4.17)

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left[\rho\left(v^2+v_t^2\right) + \frac{1}{p\Psi(\rho)}\left(p_\sigma^2 + p_{\sigma t}^2\right)\right]dx \\ &\quad + \frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}\left[\bar n\cdot\int_0^t v_{s^\alpha} d\tau\,\bar n\cdot\int_0^t v^{s^\beta} d\tau + \bar n\cdot v_{s^\alpha}\bar n\cdot v_{s^\beta}\right]ds \\ &\quad + \frac{\mu}{2}\left(\|v\|^2_{1,\Omega_t} + \|v_t\|^2_{1,\Omega_t}\right) + (\nu-\mu)\left(\|\operatorname{div} v\|^2_{0,\Omega_t} + \|\operatorname{div} v_t\|^2_{0,\Omega_t}\right) \\ &\le \varepsilon_3\left(\|p_{\sigma t}\|^2_{0,\Omega_t} + \|p_\sigma\|^2_{0,\Omega_t} + \left\|\int_0^t v d\tau\right\|^2_{1,S_t} + \|v_{xx}\|^2_{0,\Omega_t} + \left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{0,S^1}\right) \\ &\quad + P_5(\rho_*,\rho^*,\varepsilon_3)X_3Y_3 + P_6(|f|^2_{1,0,\Omega_t} + \|v\|^2_{0,\Omega_t}), \end{aligned}$$

where $\varepsilon_3 \in (0,1)$ *and*

$$X_3 = X_2, Y_3 = X_2 + \left\|\int_0^t v d\tau\right\|^2_{2,S_t}. \tag{4.18}$$

To obtain an inequality for x-derivatives we write problem (4.1) in Lagrangian coordinates, so we can introduce a partition of unity in the fixed domain Ω. Therefore, we have

$$\begin{aligned} \eta u_t^i - \nabla_{u^j} T_u^{ij}(u,q_\sigma) &= \eta g^i, \\ q_{\sigma t} + q\Psi(\eta)\nabla_u\cdot u &= 0, \\ \mathbb{T}_u(u,q_\sigma)\bar n &= \sigma\Delta_{S_t} x(\xi,t)\cdot\bar n\bar n + q_0\bar n, \end{aligned} \tag{4.19}$$

where $\eta(\xi,t) = \rho(x(\xi,t),t)$, $u(\xi,t) = v(x(\xi,t),t)$, $g(\xi,t) = f(x(\xi,t),t)$, $q(\xi,t) = p(x(\xi,t),t)$, $q_\sigma = q - p_0 - q_0$, $\nabla_{u^j} = \xi^k_{x^j}\partial_{\xi^k}$, $\Psi(\eta) = q_\eta\eta/q$, and

(4.20)

$$\mathbb{T}_u(u,q_\sigma) = \left\{T_u^{ij}(u,q_\sigma)\right\} = \left\{-q_\sigma\delta_{ij} + \mu\left(\nabla_{u^i}u^j + \nabla_{u^j}u^i\right) + (\nu-\mu)\delta_{ij}\nabla_u\cdot u\right\}.$$

Next we introduce a partition of unity $(\{\tilde{\Omega}_i\}, \{\zeta_i\}), \Omega = \cup_i \tilde{\Omega}_i$. Let $\tilde{\Omega}$ be one of the $\tilde{\Omega}_i$, and s and $\zeta(\xi) = \zeta_i(\xi)$ be the corresponding function. If $\tilde{\Omega}$ is an interior subdomain, then let $\tilde{\omega}$ be such that $\bar{\tilde{\omega}} \subset \tilde{\Omega}$ and $\zeta(\xi) = 1$ for $\xi \in \tilde{\omega}$. Otherwise, we assume that $\bar{\tilde{\Omega}} \cap S \neq \phi$, $\bar{\tilde{\omega}} \cap S \neq \phi$, $\bar{\tilde{\omega}} \subset \bar{\tilde{\Omega}}$. Let $\beta \in \bar{\tilde{\omega}} \cap S \subset \bar{\tilde{\Omega}} \cap S$, $\tilde{S} \equiv \tilde{\Omega} \cap S$. Introduce local coordinates $\{y\}$ connected with $\{\xi\}$ by

$$y^k = \alpha^{kl}(\xi^l - \beta^l), \alpha^{3k} = n^k(\beta), k = 1, 2, 3, \tag{4.21}$$

where α^{kl} is a constant orthogonal matrix, such that $\tilde{S}$ is determined by $y^3 = F(y^1, y^2), F \in H^{4+1/2}$, and

$$\tilde{\Omega} = \{y : |y^i| < d, i = 1, 2, F(y') < y^3 < F(y') + d, y' = (y^1, y^2)\}.$$

Next we introduce functions u' and q' by

$$u'^i(y) = \alpha^{ij} u^j(\xi)|_{\xi=\xi(y)}, \qquad q'(y) = q(\xi)|_{\xi=\xi(y)}, \tag{4.22}$$

where $\xi = \xi(y)$ is the inverse transformation to (4.21). Furthermore, we introduce new variables by

$$z^i = y^i, \quad i = 1, 2, \quad z^3 = y^3 - \tilde{F}(y), \quad y \in \tilde{\Omega}, \tag{4.23}$$

which will be denoted by $z = \Phi(y)$, where $\tilde{F}$ is an extension of F to $\tilde{\Omega}$ with $\tilde{F} \in H^5(\tilde{\Omega})$. Let $\hat{\Omega} = \Phi(\tilde{\Omega}) = \{z : |z^i| < d, i = 1, 2, 0 < z^3 < d\}$ and $\hat{S} = \Phi(\tilde{S})$. Define

$$\hat{u}(z) = u'(y)|_{y=\Phi^{-1}(z)}, \qquad \hat{q}(z) = q'(y)|_{y=\Phi^{-1}(z)}. \tag{4.24}$$

We introduce $\hat{\nabla}_k = \xi^l_{x^k}(\xi) z^i_{\xi^l} \nabla_{z^i}|_{\xi=\chi^{-1}(z)}$, where $\chi(\xi) = \Phi(\psi(\xi))$ and $y = \psi(\xi)$ are described by (4.21). We also introduce the following notation:

$$\tilde{u}(\xi) = u(\xi)\zeta(\xi), \quad \tilde{q}_\sigma(\xi) = q_\sigma(\xi)\zeta(\xi), \quad \xi \in \tilde{\Omega}, \quad \tilde{\Omega} \cap S = \phi, \tag{4.25}$$

and

$$\tilde{u}(z) = \hat{u}(z)\hat{\zeta}(z), \quad \tilde{q}_\sigma(z) = \hat{q}_\sigma(z)\hat{\zeta}(z), \quad z \in \hat{\Omega} = \Phi(\tilde{\Omega}), \quad \bar{\hat{\Omega}} \cap S \neq \phi, \tag{4.26}$$

where $\hat{\zeta}(z) = \zeta(\xi)|_{\xi=\chi^{-1}(z)}$.

Under the above notation problem (4.19) has the following form in an interior subdomain:

$$\eta \tilde{u}^i_t - \nabla_{u^j} T^{ij}_u(\tilde{u}, \tilde{q}_\sigma) = \eta \tilde{g}^i - \nabla_{u^j} B^{ij}_u(u, \zeta) - T^{ij}_u(u, q_\sigma)\nabla_{u^j}\zeta \equiv \eta \tilde{g}^i + k^i_1, \tag{4.27a}$$

$$\tilde{q}_{\sigma t} + q\Psi(\eta)\nabla_u \cdot \tilde{u} = q\Psi(\eta) u \cdot \nabla_u \zeta \equiv k_2, \tag{4.27b}$$

and in a boundary subdomain:

$$\hat{\eta}\tilde{u}^i_t - \hat{\nabla}_j \hat{T}^{ij}(\tilde{u}, \tilde{q}_\sigma) = \hat{\eta}\tilde{g}^i - \hat{\nabla}_j \hat{B}^{ij}(\hat{u}, \hat{\zeta}) - \hat{T}^{ij}(\hat{u}, \hat{q}_\sigma)\hat{\nabla}_j\hat{\zeta} \equiv \hat{\eta}\tilde{g}^i + k^i_3, \tag{4.28a}$$

$$\tilde{q}_{\sigma t} + \hat{q}\Psi(\hat{\eta})\hat{\nabla} \cdot \tilde{u} = \hat{q}\Psi(\hat{\eta})\hat{u} \cdot \hat{\nabla}\hat{\zeta} \equiv k_4, \tag{4.28b}$$

$$\hat{\mathbb{T}}(\tilde{u}, \tilde{q}_\sigma)\hat{n} - \sigma\hat{\Delta}_{\hat{S}}\hat{\xi} \cdot \hat{n}\hat{n}\hat{\zeta} - \sigma\hat{\Delta}_{\hat{S}} \int_0^t \tilde{u} d\tau \cdot \hat{n}\hat{n} = \frac{2\sigma}{R_0}\hat{\zeta}\hat{n} + k_5 + k_6, \tag{4.28c}$$

where

$$k^i_5 = \hat{B}^{ij}(\hat{u}, \hat{\zeta})\hat{n}_j, \qquad k_6 = -\sigma\left(2\hat{\nabla}\int_0^t \hat{u} d\tau\, \hat{\nabla}\hat{\zeta} + \int_0^t \hat{u} d\tau\, \hat{\nabla}^2\hat{\zeta}\right) \cdot \hat{n}\hat{n}, \tag{4.29}$$

$$B^{ij}_u(u, \zeta) = \mu(u^i \nabla_{u^j}\zeta + u^j \nabla_{u^i}\zeta) + (\nu - \mu)\delta_{ij} u \cdot \nabla_u \zeta,$$

and $\hat{\mathbb{T}}$, $\hat{B}$ indicate that the operator ∇_u is replaced by $\hat{\nabla}$.

In the next considerations we denote z^1, z^2 by τ and z^3 by n.

LEMMA 4.4. *Let the assumptions of Lemmas* 4.1 *and* 4.2 *be satisfied. Then we have*

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho|v|^2_{1,0}+\frac{1}{p\Psi(\rho)}|p_\sigma|^2_{1,0}\right)dx\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\left[\bar{n}\cdot\int_0^t v_{s^\alpha}d\tau\,\bar{n}\cdot\int_0^t v_{s^\beta}d\tau+\bar{n}\cdot v_{s^\alpha}\bar{n}\cdot v_{s^\beta}\right]ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\frac{1}{2}\tilde{\delta}^{\alpha\beta}\bar{n}\cdot\int_0^t v_{ss^\alpha}d\tau\,\bar{n}\cdot\int_0^t v_{ss^\beta}d\tau\,ds+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\left|\bar{n}\cdot\int_0^t v_{s^1s^2}d\tau\right|^2ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\sum_{i=1}^{2}\left(\frac{1}{2}\bar{n}\cdot\int_0^t v_{s^is^i}d\tau+2\left(H(0)+\frac{2}{R_0}\right)\right)^2ds\\
&\quad+\frac{\mu}{2}|v|^2_{2,1,\Omega_t}+|p_\sigma|^2_{1,0,\Omega_t}\qquad(4.30)\\
&\le\varepsilon_4\left(\|v_t\|^2_{2,\Omega_t}+\|p_{\sigma t}\|^2_{1,\Omega_t}+\left\|\int_0^t v d\tau\right\|^2_{0,\Omega_t}\right.\\
&\qquad\left.+\|H(\cdot,0)+\frac{2}{R_0}\|^2_{0,S^1}+\|R(\cdot,t)-R(\cdot,0)\|^2_{2,S^1}\right)\\
&\quad+P_7\left(\|f\|^2_{0,\Omega_t}+\|v\|^2_{0,\Omega_t}+\|p_\sigma\|^2_{0,\Omega_t}\right)\\
&\quad+P_8\left(X_4Y_4+\|H(\cdot,0)+\frac{2}{R_0}\|^4_{0,S^1}\right),
\end{aligned}
$$

where the summation over the repeated indices $(\alpha,\beta=1,2)$ *and coordinates* $(x, s=(s^1,s^2))$ *is assumed,* P_7 *is a positive increasing function,* $P_7=P_7(a,b)$, *and*

$$
\begin{aligned}
X_4&=X_4(t)=|v|^2_{2,1,\Omega_t}+|p_\sigma|^2_{2,1,\Omega_t},\\
Y_4&=Y_4(t)=X_4(t)+\|v\|^2_{3,\Omega_t}+\int_0^t\|v\|^2_{3,\Omega_\tau}d\tau.
\end{aligned}
\qquad(4.31)
$$

Proof. At first we consider interior subdomains. By differentiating (4.27) with respect to ξ, multiplying the result by $\tilde{u}_\xi A$ (A is the Jacobian of the transformation $x=x(\xi)$), and integrating over $\tilde{\Omega}$, we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}}\eta\tilde{u}^2_\xi Ad\xi+\frac{\mu}{2}\int_{\tilde{\Omega}}(\nabla_{u^i}\tilde{u}^j_\xi+\nabla_{u^j}\tilde{u}^i_\xi)^2Ad\xi\\
&\quad+(\nu-\mu)\|\nabla_u\cdot\tilde{u}_\xi\|^2_{0,\tilde{\Omega}}\\
&\quad-\int_{\tilde{\Omega}}\tilde{q}_{\sigma\xi}\cdot\nabla_u\tilde{u}_\xi Ad\xi\le\delta_1(\|u_{\xi\xi}\|^2_{0,\tilde{\Omega}}+\|q_{\sigma\xi}\|^2_{0,\tilde{\Omega}})\qquad(4.32)\\
&\quad+a_1(\|u\|^2_{1,\tilde{\Omega}}+\|q_\sigma\|^2_{0,\tilde{\Omega}}+\|\tilde{g}\|^2_{0,\tilde{\Omega}})\\
&\quad+a_2\left(\|u\|^2_{2,\tilde{\Omega}}\left\|\int_0^t u d\tau\right\|^2_{3,\tilde{\Omega}}+\|q_\sigma\|^2_{1,\tilde{\Omega}}|u|^2_{2,1,\tilde{\Omega}}\right),
\end{aligned}
$$

where $\|h\|_{0,\tilde{\Omega}}=(\int_{\tilde{\Omega}}|h|^2Ad\xi)^{1/2}$.

By the continuity equation (4.27b), we have

$$-\int_{\tilde{\Omega}} \tilde{q}_{\sigma\xi} \nabla_u \cdot \tilde{u}_\xi A d\xi = \frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \frac{1}{q\Psi(\eta)} \tilde{q}^2_{\sigma\xi} A d\xi + N_1, \tag{4.33}$$

where

$$|N_1| \le \delta_2 \|\tilde{q}_{\sigma\xi}\|^2_{0,\tilde{\Omega}} + a_4 \|u\|^2_{1,\tilde{\Omega}}$$

$$+a_5 \left[|\tilde{q}_\sigma|^2_{2,1,\tilde{\Omega}} (\|u\|^2_{2,\tilde{\Omega}} + |q_\sigma|^2_{2,1,\tilde{\Omega}}) + \|u\|^2_{2,\tilde{\Omega}} \left\| \int_0^t u d\tau \right\|^2_{3,\tilde{\Omega}} \right].$$

Consider the Stokes problem in $\tilde{\Omega}$:

$$\begin{gathered} \mu \nabla^2_u \tilde{u} - \nu \nabla_u \nabla_u \cdot \tilde{u} + \nabla_u \tilde{q}_\sigma = \eta \tilde{g} - \eta \tilde{u}_t + k_1, \\ \nabla_u \cdot \tilde{u} = \nabla_u \cdot \tilde{u}, \\ \tilde{u}|_{\partial\tilde{\Omega}} = 0. \end{gathered} \tag{4.34}$$

Hence, we have

$$\begin{aligned} \|\tilde{u}\|^2_{2,\tilde{\Omega}} + \|\tilde{q}_\sigma\|^2_{1,\tilde{\Omega}} &\le a_6 (\|\tilde{g}\|^2_{0,\tilde{\Omega}} + |u|^2_{1,0,\tilde{\Omega}} + \|q_\sigma\|^2_{0,\tilde{\Omega}}) \\ &+ a_7 (\|u\|^2_{2,\tilde{\Omega}} + \|q_\sigma\|^2_{1,\tilde{\Omega}}) \left\| \int_0^t u d\tau \right\|^2_{3,\tilde{\Omega}} + c \|\nabla_u \cdot \tilde{u}\|^2_{1,\tilde{\Omega}}. \end{aligned} \tag{4.35}$$

Using Lemma 5.1 from [35] in the case $G = \tilde{\Omega}$, $v = \tilde{u}_\xi$, from (4.32), (4.33), and (4.35) for sufficiently small δ_1 and δ_2, we obtain

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \left(\eta \tilde{u}^2_\xi + \frac{1}{q\Psi(\eta)} q^2_{\sigma\xi} \right) A d\xi \\ &\quad + \frac{\mu}{2} \|\tilde{u}_\xi\|^2_{1,\tilde{\Omega}} + (\nu - \mu) \|\nabla_u \cdot \tilde{u}_\xi\|^2_{0,\tilde{\Omega}} + \|\tilde{q}_{\sigma\xi}\|^2_{0,\tilde{\Omega}} \\ &\quad \le \delta_1 (\|u_{\xi\xi}\|^2_{0,\tilde{\Omega}} + \|q_{\sigma\xi}\|^2_{0,\tilde{\Omega}}) \\ &\quad + a_8 (\|\tilde{g}\|^2_{0,\tilde{\Omega}} + |u|^2_{1,0,\tilde{\Omega}} + \|q_\sigma\|^2_{0,\tilde{\Omega}}) + a_9 X_4(\tilde{\Omega}) Y_4(\tilde{\Omega}), \end{aligned} \tag{4.36}$$

where $X_4(\tilde{\Omega}) = |u|^2_{2,1,\tilde{\Omega}} + |q_\sigma|^2_{2,1,\tilde{\Omega}}$, $Y_4(\tilde{\Omega}) = X_2(\tilde{\Omega}) + \|u\|^2_{3,\tilde{\Omega}} + \int_0^t \|u\|^2_{3,\tilde{\Omega}} d\tau$.

Now we consider subdomains near the boundary. Differentiating (4.28a) with respect to τ, multiplying the result by $\tilde{u}_\tau J$, and integrating over $\hat{\Omega}$ yields (J is the Jacobian of the transformation $x = x(z)$)

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}} \hat{\eta} \tilde{u}^2_\tau J dz + \frac{\mu}{2}\int_{\hat{\Omega}} (\hat{\nabla}_i \tilde{u}^j_\tau + \hat{\nabla}_j \tilde{u}^i_\tau)^2 J dz + (\nu - \mu) \|\hat{\nabla} \cdot \tilde{u}_\tau\|^2_{0,\hat{\Omega}} \\ &\quad - \int_{\hat{\Omega}} \tilde{q}_{\sigma\tau} \hat{\nabla} \cdot \tilde{u}_\tau J dz - \int_{\hat{S}} (\hat{n} \hat{\mathbb{T}}(\tilde{u}, \tilde{q}_\sigma))_{,\tau} \tilde{u}_\tau J dz' \\ &\quad \le \delta_3 (\|\hat{u}_{zz}\|^2_{0,\hat{\Omega}} + \|\hat{q}_{\sigma z}\|^2_{0,\hat{\Omega}}) \end{aligned} \tag{4.37}$$

$$+ a_{10}(\|\tilde{g}\|^2_{0,\hat{\Omega}} + \|\hat{u}\|^2_{1,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}})$$
$$+ a_{11}\|\hat{u}\|^2_{2,\hat{\Omega}}\left(\|\hat{u}\|^2_{2,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}} + \left\|\int_0^t u d\tau\right\|^2_{3,\hat{\Omega}}\right),$$

where we have used the inequalities

$$\int_{\hat{\Omega}}[(\hat{\nabla}_j\hat{T}^{ij}(\tilde{u},\tilde{q}_\sigma))_{,\tau} - \hat{\nabla}_j\hat{T}^{ij}(\tilde{u}_\tau,\tilde{q}_{\sigma\tau})]\tilde{u}_\tau J dz' \le \delta_4(\|\tilde{u}_{zz}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z}\|^2_{0,\hat{\Omega}})$$
$$+ a_{12}\|\tilde{u}\|^2_{2,\hat{\Omega}}\left\|\int_0^t u d\tau\right\|^2_{3,\hat{\Omega}} + c(\|\hat{u}\|^2_{1,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}}),$$

and

$$\int_{\hat{S}}[(\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,\tau} - \hat{n}\hat{\mathbb{T}}(\tilde{u}_\tau,\tilde{q}_{\sigma\tau})]\tilde{u}_\tau J dz' \le \delta_5(\|\tilde{u}_{zz}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z}\|^2_{0,\hat{\Omega}})$$
$$+ a_{13}\left(\|\tilde{u}\|^2_{1,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}} + \|\hat{u}\|^2_{2,\hat{\Omega}}\left\|\int_0^t \hat{u} d\tau\right\|^2_{3,\hat{\Omega}}\right),$$

and the fact that ∇F can be expressed in terms of $\int_0^t u_\xi d\tau$. Consider the boundary term in (4.37). Using the boundary condition (4.28c), we obtain

$$-\int_{\hat{S}}(\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma)),_\tau \tilde{u}_\tau J dz' = -\sigma\int_{\hat{S}}\left(\hat{\Delta}_{\hat{s}_t}\hat{\xi}\cdot\hat{n}\hat{n}\hat{\zeta} + \frac{2}{R_0}\hat{n}\hat{\zeta}\right),_\tau \tilde{u}_\tau J dz'$$
$$-\sigma\int_{\hat{S}}\left(\hat{\Delta}_{\hat{S}_t}\int_0^t \tilde{u} d\tau\cdot\hat{n}\hat{n}\right),_\tau \cdot\,\tilde{u}_\tau J dz' + \int_{\hat{S}}(k_5+k_6),_\tau \tilde{u}_\tau J dz'. \tag{4.38}$$

Similarly as in (4.10) the first term on the right-hand side is equal to

$$-\sigma\int_U\left(H(\cdot,0) + \frac{2}{R_0}\right)\bar{n}\zeta\cdot\tilde{u}_{ss}\sqrt{g}ds^1ds^2 + N_2,$$

where

$$|N_2| \le \delta_6\left(\left\|\int_0^t \tilde{u} d\tau\right\|^2_{2,\hat{S}} + \|\hat{u}_{zz}\|^2_{2,\hat{\Omega}} + \left\|\hat{H}(\cdot,0) + \frac{2}{R_0}\right\|^2_{0,\hat{S}} + \|R(\cdot,t) - R(\cdot,0)\|^2_{2,S^1}\right)$$
$$+a_{14}(\|\xi\|_4)\left(\|\tilde{u}\|^2_{0,\hat{\Omega}} + \|\hat{u}\|^2_{2,\hat{\Omega}}\left\|\int_0^t \hat{u} d\tau\right\|^2_{3,\hat{\Omega}}\right).$$

The second term on the right-hand side of (4.38) takes the form

$$\frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}\bar{n}\cdot\int_0^t \tilde{u}_{ss^\alpha} d\tau\bar{n}\cdot\int_0^t \tilde{u}_{ss^\beta} d\tau\, ds + N_3,$$

where

$$|N_3| \le \delta_7\left(\left\|\int_0^t \tilde{u} d\tau\right\|^2_{2,\hat{S}} + \|\tilde{u}_{zz}\|^2_{0,\hat{\Omega}}\right) + a_{15}\left(\|\tilde{u}\|^2_{0,\hat{\Omega}} + \|\hat{u}\|^2_{2,\hat{\Omega}}\left\|\int_0^t \tilde{u} d\tau\right\|^2_{3,\hat{\Omega}}\right).$$

Finally, the last term in (4.38) is bounded by

$$\delta_8\left(\left\|\int_0^t \tilde{u}d\tau\right\|_{2,\hat{S}}^2 + \|\hat{u}_{zz}\|_{0,\hat{\Omega}}^2\right) + a_{16}\|\hat{u}\|_{0,\hat{\Omega}}^2.$$

By summarizing, we have proved

(4.39)

$$\begin{aligned}
&\int_{\hat{S}}\left(\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma)\right),\ {}_\tau\tilde{u}_\tau J dz' \\
&\le -\frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}\bar{n}\cdot\int_0^t \tilde{u}_{ss^\alpha}d\tau\bar{n}\cdot\int_0^t \tilde{u}_{ss^\beta}d\tau\,ds \\
&\quad -\sigma\int_{S_t}\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\tilde{u}_{ss}\cdot\bar{n}ds \\
&\quad +\delta_9\left(\left\|\int_0^t\tilde{u}d\tau\right\|_{2,\hat{S}}^2 + \|\tilde{u}_{zz}\|_{0,\hat{\Omega}}^2 + \left\|\left(\hat{H}(0)+\frac{2}{R_0}\right)\hat{\zeta}\right\|_{0,\hat{S}}^2 + \|R(\cdot,t)-R(\cdot,0)\|_{2,S^1}^2\right) \\
&\quad +a_{17}\left(\|\hat{u}\|_{0,\hat{\Omega}}^2 + \|\hat{u}\|_{2,\hat{\Omega}}^2\left\|\int_0^t\tilde{u}d\tau\right\|_{3,\hat{\Omega}}^2\right).
\end{aligned}$$

From the continuity equation (4.28b) we get

$$-\int_{\hat{\Omega}}\tilde{q}_{\sigma\tau}\nabla_u\cdot\tilde{u}_\tau J dz = \frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}_{\sigma\tau}^2 J dz + N_4, \tag{4.40}$$

where

$$\begin{aligned}
|N_4| &\le \delta_{10}\|\tilde{q}_{\sigma\tau}\|_{0,\hat{\Omega}}^2 + c\|\hat{u}\|_{1,\hat{\Omega}}^2 \\
&\quad + a_{18}\left[|\hat{q}_\sigma|_{2,1,\hat{\Omega}}^2\left(\|\hat{u}\|_{2,\hat{\Omega}}^2 + |\hat{q}_\sigma|_{2,1,\hat{\Omega}}^2\right) + \|\hat{u}\|_{2,\hat{\Omega}}^2\left\|\int_0^t u dt'\right\|_{2,\hat{\Omega}}^2\right],
\end{aligned}$$

and $P = P(|\int_0^t \hat{u}_z dt'|_{0,\hat{\Omega}}, \|h\|_{0,\hat{\Omega}} = (\int_{\hat{\Omega}}|h|^2 J dz)^{1/2})$. From (4.37)–(4.40) and Lemma 5.1 from [35] in the case $G = \hat{\Omega}$, $v = \tilde{u}_\tau$, we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}_\tau^2 + \frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}_{\sigma\tau}^2\right)Jdz \\
&\qquad +\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}} g^{\alpha\beta}\bar{n}\cdot\int_0^t\tilde{u}_{ss^\alpha}d\tau\bar{n}\cdot\int_0^t\tilde{u}_{ss^\beta}d\tau\,ds' \\
&\qquad +\sigma\int_{\hat{S}}\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\tilde{u}_{ss}\cdot\bar{n}ds' \\
&\qquad +\frac{\mu}{2}\|\tilde{u}_\tau\|_{1,\hat{\Omega}}^2 + (\nu-\mu)\|\hat{\nabla}\cdot\tilde{u}_\tau\|_{0,\hat{\Omega}}^2 \qquad (4.41)\\
&\quad \le \delta_{11}\left(\|\hat{u}_{zz}\|_{0,\hat{\Omega}}^2 + \|\hat{q}_{\sigma z}\|_{0,\hat{\Omega}}^2 + \left\|\int_0^t\tilde{u}d\tau\right\|_{2,\hat{S}}^2 + \|R(\cdot,t)-R(\cdot,0)\|_{2,S^1}^2\right) \\
&\qquad +a_{19}\left(\|\hat{u}\|_{1,\hat{\Omega}}^2 + \|\hat{q}_\sigma\|_{0,\hat{\Omega}}^2 + \|\tilde{q}\|_{0,\hat{\Omega}}^2 + \|H(\cdot,0)+\frac{2}{R_0}\|_{0,\hat{S}}^2\right) \\
&\qquad +a_{20}X_4(\hat{\Omega})Y_4(\hat{\Omega}),
\end{aligned}$$

where $X_4(\hat{\Omega}) = |\hat{u}|^2_{2,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}$, $Y_4(\hat{\Omega}) = X_4(\hat{\Omega}) + \|\hat{u}\|^2_{3,\hat{\Omega}} + \int_0^t \|\hat{u}\|^2_{3,\hat{\Omega}} dt'$.

Applying the operator $(\mu+\nu)\hat{\nabla}$ to (4.28b), dividing the result by $\hat{q}\Psi(\hat{\eta})$, and adding to (4.28a) gives

(4.42)
$$\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\hat{\nabla}_i\tilde{q}_{\sigma t} + \hat{\nabla}_i\tilde{q}_\sigma = \mu(\hat{\nabla}^2\tilde{u}^i - \hat{\nabla}_i\hat{\nabla}\cdot\tilde{u}) - \hat{\eta}\tilde{u}^i_t + \hat{\eta}\tilde{g}^i - \frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\hat{\nabla}(\hat{q}\Psi(\hat{\eta}))\hat{\nabla}\cdot\tilde{u} + \frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\hat{\nabla}(\hat{q}\Psi(\hat{\eta})\hat{u}\cdot\hat{\nabla}\hat{\zeta}) + k_3.$$

Multiplying the normal component of (4.42) by $\tilde{q}_{\sigma n}J$ and integrating over $\hat{\Omega}$ implies

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma n}Jdz + \frac{1}{2}\|\tilde{q}_{\sigma n}\|^2_{0,\hat{\Omega}} \\
&\quad\le c\|\tilde{u}_{z\tau}\|^2_{0,\hat{\Omega}} + a_{21}(\|\tilde{u}_t\|^2_{0,\hat{\Omega}} + \|\hat{u}\|^2_{1,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}} + \|\tilde{g}\|^2_{0,\hat{\Omega}}) \\
&\qquad + a_{22}\left(\|\tilde{u}\|^2_{2,\hat{\Omega}}\|\int_0^t \hat{u}dt'\|^2_{3,\hat{\Omega}} + (\delta_7' + cd)(\|\tilde{u}_{zz}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z}\|^2_{0,\hat{\Omega}})\right. \\
&\qquad\quad \left. + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}\left(\|\hat{q}_\sigma\|^2_{2,\hat{\Omega}} + \|\hat{u}\|^2_{2,\hat{\Omega}} + \|\int_0^t \hat{u}dt'\|^2_{2,\hat{\Omega}}\right)\right).
\end{aligned} \tag{4.43}$$

We write (4.28a) in the form

$$\hat{\eta}\tilde{u}^i_t - \mu\Delta\tilde{u}^i - \nu\nabla_i\nabla\cdot\tilde{u} = \hat{\nabla}_i\tilde{q}_\sigma + \eta\tilde{g}^i + k^i_3 - k^i_7, \tag{4.44}$$

where $k_7 = (\mu\Delta\tilde{u}^i + \nu\nabla_i\nabla\cdot\tilde{u}) - (\mu\hat{\nabla}^2\tilde{u}^i + \nu\hat{\nabla}_i\hat{\nabla}\cdot\tilde{u})$.

Multiplying the third component of (4.44) by $\tilde{u}^3_{nn}J$ and integrating over $\hat{\Omega}$ yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}|\tilde{u}^3_n|^2Jdz + \frac{\mu+\nu}{2}\|\tilde{u}^3_{nn}\|^2_{0,\hat{\Omega}} \\
&\quad\le c(\|\tilde{u}_{z\tau}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma n}\|^2_{0,\hat{\Omega}}) \\
&\qquad + a_{23}(\|\hat{u}\|^2_{1,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}} + \|\tilde{g}\|^2_{0,\hat{\Omega}} + \|\tilde{u}_t\|^2_{0,\hat{\Omega}}) \\
&\qquad + \delta_{12}(\|\tilde{u}^3_{nt}\|^2_{0,\hat{\Omega}} + \|\tilde{u}_{zz}\|^2_{0,\hat{\Omega}}) \\
&\qquad + a_{24}\left(|\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}\|\tilde{u}\|^2_{2,\hat{\Omega}} + \|\hat{u}\|^4_{2,\hat{\Omega}} + \|\tilde{u}\|^2_{2,\hat{\Omega}}\left\|\int_0^t \hat{u}dt'\right\|^2_{3,\hat{\Omega}}\right).
\end{aligned} \tag{4.45}$$

To estimate $\tilde{u}^i_{nn}$, $i = 1, 2$, and $\tilde{q}_{\sigma\tau}$ we write (4.44) in the form

$$\begin{aligned}
&-\mu\Delta\tilde{u}^i + \nabla_{z^i}\tilde{q}_\sigma \\
&\quad = \hat{\eta}\tilde{g}^i - \hat{\eta}\tilde{u}^i_t + k^i_3 - k^i_7 + \nabla_{z^i}\tilde{q}_\sigma - \hat{\nabla}_i\tilde{q}_\sigma + \nu\nabla_{z^i}\mathrm{div}\tilde{u} \\
&\quad = \tilde{f}^i + \nu\nabla_{z^i}\mathrm{div}\tilde{u},
\end{aligned} \tag{4.46}$$

and the boundary condition (4.28c) as

$$\frac{\partial\tilde{u}^i}{\partial z^3} = -\frac{\partial\tilde{u}^3}{\partial z^i} + \left(\frac{\partial\tilde{u}^i}{\partial z^3} + \frac{\partial\tilde{u}^3}{\partial z^i} - \frac{1}{\mu}\hat{\tau}_i\hat{\mathbb{T}}\hat{n}\right) + \frac{1}{\mu}k_5\cdot\hat{\tau}_i \equiv \tilde{h}^i, \qquad i = 1, 2, z^3 = 0, \tag{4.47}$$

where we have also used the fact that $\hat{\tau}_i \cdot \hat{n} = 0$, $i = 1, 2$. Considering the problem (4.46) and (4.47) in $\hat{\Omega}$ we have to add the boundary conditions

$$\tilde{u}^i|_{|z'|=d} = 0, \quad \tilde{u}^i|_{z^3=d} = 0, \quad i = 1, 2, \tilde{q}_\sigma|_{|z'|=d} = 0, \quad \tilde{q}_\sigma|_{z^3=d} = 0. \tag{4.48}$$

Multiplying (4.46) by $\tilde{u}^i$, summing over $i = 1, 2$, integrating over $\hat{\Omega}$, and using boundary conditions (4.47) and (4.48) yields

$$\|\nabla \tilde{u}'\|^2_{0,\hat{\Omega}} \le \delta_9 \|\tilde{q}_\sigma\|^2_{0,\hat{\Omega}} + c\left(\|\tilde{f}'\|^2_{0,\hat{\Omega}} + \|\tilde{h}'\|^2_{0,\hat{S}} + \|\operatorname{div} \tilde{u}\|^2_{0,\hat{\Omega}}\right), \tag{4.49}$$

where the prime denotes that only two components ($i = 1, 2$) are taken into account.

We now look for a function $w \in H^1(\hat{\Omega})$ such that

$$\operatorname{div} w = \tilde{q}_\sigma, w^3|_{z^3=0} = \chi(z') \int_{\hat{\Omega}} \tilde{q}_\sigma dz, \quad w|_{\partial\hat{\Omega}\setminus\hat{S}} = 0, \quad w^i|_{z^3=0} = 0, \quad i = 1, 2, \tag{4.50}$$

where $\chi(z')$ is a smooth function such that $\int_{\hat{S}} \chi(z')dz' = 1$, $\chi(z') \ge 0$, $\chi|_{|z'|=d} = 0$. Moreover, $1 \le 4d^2|\chi|_{\infty,\hat{S}}$, so $|\chi|_{\infty,\hat{S}} \ge 1/(4d^2)$. Finally, assuming that χ vanishes only in a neighborhood of the boundary of $\hat{S}$, we require that $\min \chi(z')_{|z'|\le d/2} > 0$. Hence

$$1 = \int_{\hat{S}} \chi(z')dz' \ge \int_{|z'|\le d/2} \chi(z')dz' \ge d^2 \min_{|z'|\le d/2} \chi(z'), \quad \text{so } \min_{|z'|\le d/2} \chi(z') \le 1/d^2.$$

Therefore, we can assume that $\chi(z') \le c/d^2$.

We look for solutions of (4.50) in the form $w = \nabla\varphi + \alpha$, where φ is a solution to the Neumann problem

$$\begin{gathered} \Delta\varphi = \tilde{q}_\sigma, \quad \partial_{z^3}\varphi|_{z^3=0} = \chi(z') \int_{\hat{\Omega}} \tilde{q}_\sigma dz \equiv \varphi_0, \quad \partial_{z^3}\varphi|_{z^3=d} = 0, \\ \partial_{z^i}\varphi|_{|z^i|=d} = 0, \qquad i = 1, 2, \int_{\hat{\Omega}} \varphi dz = 0, \end{gathered} \tag{4.51}$$

and

$$\operatorname{div} \alpha = 0, \quad \alpha|_{\partial\hat{\Omega}\setminus\hat{S}} = -\nabla\varphi|_{\partial\hat{\Omega}\setminus\hat{S}}, \quad \alpha \cdot \bar{n}|_{\hat{S}} = 0, \quad \alpha \cdot \bar{\tau}_i|_{\hat{S}} = -\bar{\tau}_i \cdot \nabla\varphi|_{\hat{S}}, \quad i = 1, 2, \tag{4.52}$$

where $\bar{n}, \bar{\tau}_i$, $i = 1, 2$, are normal and tangent vectors to $\hat{S}$.

Since the compatibility condition for (4.51) is satisfied, there exists a unique solution to (4.51) such that $\varphi \in H^2(\hat{\Omega})$ and

$$\|\varphi\|_{2,\hat{\Omega}} \le c(\|\tilde{q}_\sigma\|_{0,\hat{\Omega}} + \|\varphi_0\|_{0,\hat{S}} + \|\varphi_0\|_{1/2,\hat{S}}) \le c(1 + d^{1/2})\|\tilde{q}_\sigma\|_{0,\hat{\Omega}} \tag{4.53}$$

because

$$\begin{aligned} \|\varphi_0\|_{0,\hat{S}} &\le \left|\int_{\hat{\Omega}} \hat{q}_\sigma dz\right| \left(\int_{\hat{S}} |\chi(z')|^2 dz'\right)^{1/2} \le cd^{1/2}\|\tilde{q}_\sigma\|_{0,\hat{\Omega}}, \\ \|\varphi_0\|_{1/2,\hat{S}} &\le \left|\int_{\hat{\Omega}} \hat{q}_\sigma dz\right| \left(\int_{\hat{S}}\int_{\hat{S}} \frac{|\chi(x') - \chi(y')|^2}{|x' - y'|^3} dx'dy'\right)^{1/2} \\ &\le \frac{c}{d^3} \left|\int_{\hat{\Omega}} \tilde{q}_\sigma dz\right| \left(\int_{\hat{S}}\int_{\hat{S}} |x' - y'|^{-1} dx'dy'\right)^{1/2} \\ &\le cd^{-3/2} \left|\int_{\hat{\Omega}} \tilde{q}_\sigma dz\right| \le c\|\tilde{q}_\sigma\|_{0,\tilde{\Omega}}, \end{aligned}$$

where we have used the fact that $|\nabla\chi| \le c/d^3$.

Similarly, the compatibility condition for (4.52) is satisfied because $\bar{n}\cdot\nabla\varphi|_{\partial\hat{\Omega}\setminus\hat{S}} = 0$. Hence, there exists a solution to (4.52) such that $\alpha \in H^1(\hat{\Omega})$ and

$$\|\alpha\|_{1,\hat{\Omega}} \le c\|\nabla\varphi\|_{1/2,\partial\hat{\Omega}} \le c\|\varphi\|_{2,\hat{\Omega}} \le c\|\tilde{q}_\sigma\|_{0,\hat{\Omega}}. \tag{4.54}$$

By summarizing, there exists a solution of problem (4.50) such that $w \in H^1(\hat{\Omega})$ and

$$\|w\|_{1,\hat{\Omega}} \le c\|\tilde{q}_\sigma\|_{0,\hat{\Omega}}. \tag{4.55}$$

Now we estimate $\|\tilde{q}_\sigma\|_{0,\hat{\Omega}}$. Multiplying (4.46) by w and integrating over $\hat{\Omega}$ yields

$$-\mu\int_{\hat{\Omega}} \Delta\tilde{u}\cdot w dz + \int_{\hat{\Omega}} \nabla\tilde{q}_\sigma \cdot w dz = \int_{\hat{\Omega}} \tilde{f}\cdot w dz + \nu\int_{\hat{\Omega}} \nabla\mathrm{div}\tilde{u}\cdot w dz. \tag{4.56}$$

The boundary term which follows from integration by parts in the first term of (4.56) is estimated in the following way:

$$\left|\mu\int_{\hat{S}} \bar{n}\cdot\nabla\tilde{u}\cdot w dz'\right| \le c\left|\int_{\hat{S}} \tilde{u}^3_{z^3} w^3 dz'\right| \le c\|\tilde{u}^3_{z^3}\|_{-1/2,\hat{S}}\|w^3\|_{1/2,\hat{S}}$$

$$\le c\|\tilde{u}^3_{z^3}\|_{0,\hat{\Omega}}\|w\|_{1,\hat{\Omega}}.$$

The second term on the left-hand side of (4.56) is equal to

$$\int_{\hat{S}} \tilde{q_\sigma}\cdot w^3 dz' - \int_{\hat{\Omega}} \tilde{q}_\sigma \mathrm{div}\, w dz,$$

where

$$\left|\int_{\hat{S}} \tilde{q}_\sigma w^3 dz'\right| = \left|\int_{\hat{\Omega}} \tilde{q}_\sigma dz \int_{\hat{S}} \tilde{q}_\sigma\chi(z')dz'\right| \le c/d^2 \int_{\hat{\Omega}} |\tilde{q}_\sigma| dz \int_{\hat{S}} |\tilde{q}_\sigma| dz'$$

$$\le cd^{-1/2}\|\tilde{q}_\sigma\|_{0,\hat{\Omega}} \int_{\hat{S}} |\tilde{q}_\sigma(z')| dz'$$

$$\le cd^{1/2}\|\tilde{q}_\sigma\|_{0,\hat{\Omega}}\|\tilde{q}_\sigma\|_{0,\hat{S}} \le \delta_{13}^2\|\tilde{q}_{\sigma z}\|^2_{0,\hat{\Omega}} + c(\delta_{13})d\|\tilde{q}_\sigma\|^2_{0,\hat{\Omega}},$$

and

$$\int_{\hat{\Omega}} \tilde{q}_\sigma \mathrm{div}\, w\, dz = \|\tilde{q}_\sigma\|^2_{0,\hat{\Omega}}.$$

Finally, the last term in (4.56) can be expressed in the form

$$\int_{\hat{\Omega}} \nabla\mathrm{div}\, \tilde{u}\cdot w dz = \int_{\hat{S}} \mathrm{div}\, \tilde{u}\, w^3 dz' - \int_{\hat{\Omega}} \mathrm{div}\, \tilde{u}\, \mathrm{div}\, w\, dz,$$

where

$$\left|\int_{\hat{S}} \mathrm{div}\, \tilde{u}\, w^3\, dz'\right| = \left|\int_{\hat{\Omega}} \tilde{q}_\sigma dz \int_{\hat{S}} \mathrm{div}\, \tilde{u}\, \chi(z')dz'\right| \le cd^{-1/2}\|\tilde{q}_\sigma\|_{0,\hat{\Omega}} \int_{\hat{S}} |\mathrm{div}\, \tilde{u}| dz'$$

$$\le cd^{1/2}\|\tilde{q}_\sigma\|_{0,\hat{\Omega}}\|\mathrm{div}\tilde{u}\|_{0,\hat{S}} \le d\|\tilde{q}_\sigma\|^2_{0,\hat{\Omega}} + c\|\mathrm{div}\tilde{u}\|^2_{1,\hat{\Omega}}.$$

By summarizing, we obtain the estimate

$$\|\tilde{q}_\sigma\|^2_{0,\hat{\Omega}} \le \delta_{13}\|\tilde{q}_{\sigma z}\|^2_{0,\hat{\Omega}} + c\left(\|\tilde{f}\|^2_{0,\hat{\Omega}} + \|\tilde{h}'\|^2_{0,\hat{\Omega}} + \|\tilde{u}^3_{z^3}\|^2_{0,\hat{\Omega}} + \|\mathrm{div}\ \tilde{u}\|^2_{1,\hat{\Omega}}\right) \tag{4.57}$$

for sufficiently small d.

Now instead of problems (4.46) and (4.47) we consider the problem

$$\begin{aligned} -\mu\Delta\tilde{u}^i_{z'} + \nabla_{z^i}\tilde{q}_{\sigma z'} &= \tilde{f}^i_{z'} + \nabla_{z^i}\mathrm{div}\ \tilde{u}_{z'}, \qquad i = 1,2,3, \\ \partial_{z^3}\tilde{u}^i_{z'} &= \tilde{h}^i_{z'}, \qquad i = 1,2. \end{aligned} \tag{4.58}$$

Multiplying (4.58) by $\tilde{u}^i_{z'}$, summing over $i = 1,2$, and integrating over $\hat{\Omega}$ yields

$$\|\tilde{u}'_{zz'}\|^2_{0,\hat{\Omega}} \le \delta_{13}\|\tilde{q}_{\sigma z'}\|^2_{0,\hat{\Omega}} + c(\|\tilde{f}'\|^2_{0,\hat{\Omega}} + \|\tilde{h}'_{z'}\|^2_{0,\hat{\Omega}} + \|\mathrm{div}\tilde{u}\|^2_{1,\hat{\Omega}}). \tag{4.59}$$

Finally, let us introduce a function $w_1 \in W^1_2(\hat{\Omega})$ such that

$$\mathrm{div}w_1 = \tilde{q}_{\sigma z'}, \qquad w_1|_{\partial\hat{\Omega}} = 0. \tag{4.60}$$

By $\int_{\hat{\Omega}} \tilde{q}_{\sigma z'}dz = 0$ there exists a solution of (4.60) such that $w_1 \in H^1(\hat{\Omega})$ and

$$\|w_1\|_{1,\hat{\Omega}} \le c\|\tilde{q}_{\sigma z'}\|_{0,\hat{\Omega}}. \tag{4.61}$$

Multiplying the first equation of (4.58) by w_1 and integrating over $\tilde{\Omega}$ gives

$$\|\tilde{q}_{\sigma z'}\|^2_{0,\hat{\Omega}} \le c(\|\tilde{f}\|^2_{0,\hat{\Omega}} + \|\mathrm{div}\ \tilde{u}\|^2_{1,\hat{\Omega}} + \|\tilde{u}_{zz'}\|^2_{0,\hat{\Omega}}). \tag{4.62}$$

From (4.49), (4.57), (4.59), and (4.62) we have

$$\begin{aligned} &\|\tilde{u}'_z\|^2_{0,\hat{\Omega}} + \|\tilde{u}'_{zz'}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_\sigma\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z'}\|^2_{0,\hat{\Omega}} \\ &\quad \le c(\|\tilde{f}'\|^2_{0,\hat{\Omega}} + \|\tilde{h}'\|^2_{1,\hat{\Omega}} + \|\mathrm{div}\ \tilde{u}\|^2_{1,\hat{\Omega}} + \|\tilde{u}\|^2_{1,\hat{\Omega}}) \\ &\qquad +\delta_{13}\|\tilde{q}_{\sigma z^3}\|^2_{0,\hat{\Omega}}. \end{aligned} \tag{4.63}$$

From the form of $\tilde{f}'$ and $\tilde{h}'$ we have

$$\begin{aligned} \|\tilde{f}\|^2_{0,\hat{\Omega}} &\le c(\|\tilde{g}\|^2_{0,\hat{\Omega}} + \|\tilde{u}_t\|^2_{0,\hat{\Omega}} + \|\hat{u}\|^2_{1,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}}) \\ &\quad +c\left(\left\|\int_0^t \hat{u}dt'\right\|^2_{3,\hat{\Omega}} + d\right)(\|\tilde{q}_\sigma\|^2_{2,\hat{\Omega}} + \|\tilde{u}\|^2_{2,\hat{\Omega}}), \\ \|\tilde{h}\|^2_{1,\hat{\Omega}} &\le c\left(\|\tilde{u}^3_{z\tau}\|^2_{0,\hat{\Omega}} + \left(\left\|\int_0^t \hat{u}dt'\right\|^2_{3,\hat{\Omega}} + d\right)(\|\tilde{u}\|^2_{2,\hat{\Omega}} + \|\hat{u}\|^2_{1,\hat{\Omega}})\right). \end{aligned} \tag{4.64}$$

Finally, from (4.46) we obtain

$$\|\tilde{u}'_{nn}\|^2_{0,\hat{\Omega}} \le c(\|\tilde{q}_{\sigma\tau}\|^2_{0,\hat{\Omega}} + \|\tilde{f}\|^2_{0,\hat{\Omega}} + \|\mathrm{div}\ \tilde{u}\|^2_{1,\hat{\Omega}} + \|\tilde{u'}_{zz'}\|^2_{0,\hat{\Omega}}). \tag{4.65}$$

Therefore, equations (4.63)–(4.65) imply

$$
\begin{aligned}
&\|\tilde{u}'_{nn}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma\tau}\|^2_{0,\hat{\Omega}} \\
&\quad \le c(\|\tilde{u}^3_{z\tau}\|^2_{0,\hat{\Omega}} + \|\operatorname{div}\tilde{u}\|^2_{1,\hat{\Omega}}) \\
&\quad\quad + a_{25}(\|\tilde{g}\|^2_{0,\hat{\Omega}} + |\tilde{u}|^2_{1,0,\hat{\Omega}} + \|\tilde{q}_\sigma\|^2_{0,\hat{\Omega}}) \\
&\quad\quad + a_{26}\left\|\int_0^t \hat{u}dt'\right\|^2_{3,\hat{\Omega}} (\|\tilde{q}_\sigma\|^2_{2,\hat{\Omega}} + \|\tilde{u}\|^2_{2,\hat{\Omega}}) + \delta_{14}\|\tilde{q}_{\sigma n}\|^2_{0,\hat{\Omega}}.
\end{aligned} \tag{4.66}
$$

Now equations (4.41), (4.43), (4.45), and (4.66) follow:

(4.67)

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left[\hat{\eta}(\tilde{u}^2_\tau + |\tilde{u}^3_n|^2) + \frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma z}\right]dz \\
&\qquad + \frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}} g^{\alpha\beta}\bar{n}\cdot\int_0^t \tilde{u}_{ss^\alpha}d\tau\,\bar{n}\cdot\int_0^t \tilde{u}_{ss^\beta}d\tau\,ds' \\
&\qquad + \sigma\int_{\hat{S}}\left(H(0) + \frac{2}{R_0}\right)\zeta\tilde{u}_{ss}\cdot\bar{n}\,ds' \\
&\qquad + \frac{\mu}{2}\|\tilde{u}\|^2_{2,\hat{\Omega}} + \|\tilde{q}_{\sigma z}\|^2_{0,\hat{\Omega}} \\
&\quad \le \delta_{15}\left(\left\|\int_0^t \tilde{u}d\tau\right\|^2_{2,\hat{S}} + \|\hat{u}_{zz}\|^2_{0,\hat{\Omega}} + \left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{2,\hat{S}} + \|R(\cdot,t) - R(\cdot,0)\|^2_{2,S^1}\right) \\
&\qquad + a_{27}(\|\tilde{g}\|^2_{0,\hat{\Omega}} + |\hat{u}|^2_{1,0,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}}) \\
&\qquad + a_{28}X_4(\hat{\Omega})Y_4(\hat{\Omega}).
\end{aligned}
$$

We also have

$$
\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}\tilde{u}^2_n J dz \le \delta_{16}\|\tilde{u}_{nt}\|^2_{0,\hat{\Omega}} + c\|\tilde{u}\|^2_{1,\hat{\Omega}} + a_{29}X_4(\hat{\Omega})Y_4(\hat{\Omega}). \tag{4.68}
$$

From (4.67) and (4.68) we have

(4.69)

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left[\hat{\eta}\tilde{u}^2_z + \frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma z}\right]dz \\
&\qquad + \frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}\left[g^{\alpha\beta}\bar{n}\cdot\int_0^t \tilde{u}_{ss^\alpha}d\tau\,\bar{n}\cdot\int_0^t \tilde{u}_{ss^\beta}d\tau + 2\left(H(0) + \frac{2}{R_0}\right)\zeta\bar{n}\cdot\int_0^t \tilde{u}_{ss}d\tau\right]ds' \\
&\qquad + \frac{\mu}{2}\|\tilde{u}\|^2_{2,\hat{\Omega}} + \|\tilde{q}_{\sigma z}\|^2_{0,\hat{\Omega}} \\
&\quad \le \delta_{17}\left(\left\|\int_0^t \tilde{u}d\tau\right\|^2_{2,\hat{S}} + \|\hat{u}_{zz}\|^2_{0,\hat{\Omega}} + \left\|H(0) + \frac{2}{R_0}\right\|^2_{0,\hat{S}} + \|R(\cdot,t) - R(\cdot,0)\|^2_{2,S^1}\right) \\
&\qquad + a_{30}(|\hat{u}|^2_{1,0,\hat{\Omega}} + \|\hat{q}_\sigma\|^2_{0,\hat{\Omega}} + \|\tilde{g}\|^2_{0,\hat{\Omega}}) \\
&\qquad + a_{31}X_4(\hat{\Omega})Y_4(\hat{\Omega}).
\end{aligned}
$$

We examine the second term on the left-hand side of (4.69). By employing the fact that the part of the boundary $S_t \cap \{x : \zeta(x) \neq 0\}$ can be described in the local

coordinates $\{y\}$ by the formula $y^i = s^i, i = 1, 2, y^3 = \bar{F}(s^1, s^2, t)$, we have $g^{\alpha\beta} = \delta^{\alpha\beta} + \varepsilon^{\alpha\beta}$, where $\varepsilon^{\alpha\beta} = \bar{F}_{s^\alpha}\bar{F}_{s^\beta}\left(1 + \bar{F}_{s^1}^2 + \bar{F}_{s^2}^2\right)^{-1}$. Assuming that $\operatorname{supp}\{\zeta\}$ is sufficiently small, we have that $|\bar{F}_s| \le \frac{1}{2}$. Then, performing summation over $s \in \{s^1, s^2\}$, we write the second term in the form

$$
\begin{aligned}
&\frac{\sigma}{2}\frac{d}{dt}\int_U \frac{1}{2}\tilde{\delta}^{\alpha\beta}\bar{n}\cdot\int_0^t \tilde{u}_{ss^\alpha}d\tau\,\bar{n}\cdot\int_0^t \tilde{u}_{ss^\beta}d\tau\sqrt{g}\,ds^1ds^2\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_U\left|\bar{n}\cdot\int_0^t \tilde{u}_{s^1s^2}d\tau\right|^2\sqrt{g}ds^1ds^2\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_U\sum_{i=1}^2\left(\frac{1}{2}\bar{n}\cdot\int_0^t \tilde{u}_{s^is^i}d\tau+2\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)^2\sqrt{g}ds^1ds^2\\
&\quad-4\sigma\frac{d}{dt}\int_U\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)^2\sqrt{g}ds^1ds^2,
\end{aligned}
\tag{4.70}
$$

where $\tilde{\delta}^{\alpha\beta} = \delta^{\alpha\beta} + 2\varepsilon^{\alpha\beta}$ and $\frac{3}{4}\xi^2 \le \tilde{\delta}^{\alpha\beta}\xi_\alpha\xi_\beta, \xi^2 = \xi_1^2 + \xi_2^2$.

We use (4.70) in (4.69) and then we go back to the variables ξ. Then from the resulting estimate and (4.36), after summing over all neighborhoods of the partition of unity, and going back to the variables x and from using (4.17), we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho|v|_{1,0}^2+\frac{1}{p\Psi(\rho)}|p_\sigma|_{1,0}^2\right)dt\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\left[\bar{n}\cdot\int_0^t v_{s^\alpha}d\tau\bar{n}\cdot\int_0^t v_{s^\beta}d\tau+\bar{n}\cdot v_{s^\alpha}\bar{n}\cdot v_{s^\beta}\right]ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\frac{1}{2}\tilde{\delta}^{\alpha\beta}\bar{n}\cdot\int_0^t v_{ss^\alpha}d\tau\bar{n}\cdot\int_0^t v_{ss^\beta}d\tau ds+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\left|\bar{n}\cdot\int_0^t v_{s^1s^2}d\tau\right|^2ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\sum_{i=1}^2\left(\frac{1}{2}\bar{n}\cdot\int_0^t v_{s^1s^2}d\tau+2\left(H(\cdot,0)+\frac{2}{R_0}\right)\right)^2ds\\
&\quad+\frac{\mu}{2}|v|_{2,1,\Omega_t}^2+|p_\sigma|_{1,0,\Omega_t}^2\\
&\le\delta_{17}\left(\left\|\int_0^t v d\tau\right\|_{2,S_t}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{0,S_t}^2+\|R(\cdot,t)-R(\cdot,0)\|_{2,S^1}^2\right)\\
&\quad+a_{32}(|v|_{1,0,\Omega_t}^2+\|p_\sigma\|_{0,\Omega_t}^2+|f|_{1,0,\Omega_t}^2)+a_{33}X_4Y_4\\
&\quad+4\sigma\frac{d}{dt}\int_{S_t}\left(H(\cdot,0)+\frac{2}{R_0}\right)^2ds.
\end{aligned}
\tag{4.71}
$$

By virtue of the interpolation inequality (1.12), we obtain

$$
\begin{aligned}
\left|\frac{d}{dt}\int_{S_t}\left(H(\cdot,0)+\frac{2}{R_0}\right)^2\right|ds&\le\delta_{18}\|v_{xx}\|_{0,\Omega_t}^2\\
&\quad+a_{34}(\|H(\cdot,0)+\frac{2}{R_0}\|_{0,S^1}^4+\|v\|_{0,\Omega_t}^2).
\end{aligned}
\tag{4.72}
$$

Expressing boundary conditions (4.1c) locally we have

$$
\sigma\hat{\Delta}_{\hat{S}_t}\int_0^t\tilde{u}dt'=-\sigma\left(\hat{\Delta}_{\hat{S}_t}\hat{\xi}+\frac{2}{R_0}\hat{n}\right)\hat{\zeta}-\hat{\mathbb{T}}_u(\tilde{u},\tilde{q}_\sigma)\hat{n}+l_1+l_2,
\tag{4.73}
$$

where

$$l_1^i = -\hat{B}^{ij}(\hat{u},\hat{\zeta})\hat{n}_j, \qquad l_2 = \sigma\left(2\hat{\nabla}\int_0^t \hat{u}d\tau\,\hat{\nabla}\hat{\zeta} + \int_0^t \hat{u}d\tau\,\hat{\nabla}^2\hat{\zeta}\right). \tag{4.74}$$

Multiply (4.73) by $\int_0^t \tilde{u}dt'$, then differentiate with respect to τ and multiply by $\int_0^t \tilde{u}_\tau dt'$. By integrating the result over $\hat{S}$ and summing over all neighborhoods of the partition of unity we obtain

$$\begin{aligned}\left\|\int_0^t v d\tau\right\|^2_{2,S_t} &\le \delta_{19}\left(\|v\|^2_{2,\Omega_t} + \|p_\sigma\|^2_{1,\Omega_t}\right)\\ &\quad + a_{35}\Bigg(\|v\|^2_{0,\Omega_t} + \|p_\sigma\|^2_{0,\Omega_t} + \left\|\int_0^t v d\tau\right\|^2_{0,\Omega_t}\\ &\qquad + \left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{0,S^1} + \|R(\cdot,t) - R(\cdot,0)\|^2_{2,S^1}\Bigg)\\ &\quad + a_{36}(\|v\|^2_{2,\Omega_t} + \|p_\sigma\|^2_{2,\Omega_t})\int_0^t \|v\|^2_{3,\Omega_\tau} d\tau.\end{aligned} \tag{4.75}$$

From (4.71), (4.72), (4.75), and sufficiently small δ_{17}, δ_{18}, δ_{19} we get (4.30). This concludes the proof.

Now we obtain an inequality for the third derivatives.

LEMMA 4.5. *For a sufficiently smooth solution of the problem* (4.1a–c) *the following inequality holds*:

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v_{xx}^2 + \frac{1}{p\Psi(\rho)}p_{\sigma xx}^2\right)dx\\ &\quad + \frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\frac{1}{2}\tilde{\delta}^{\alpha\beta}\bar{n}\cdot\int_0^t v_{s_1s_2s^\alpha}d\tau\bar{n}\cdot\int_0^t v_{s_1s_2s^\beta}d\tau ds\\ &\quad + \frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\left|\bar{n}\cdot\int_0^t v_{s^1s^2s}d\tau\right|^2 ds\\ &\quad + \frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\sum_{i=1}^2\left(\frac{1}{2}\bar{n}\cdot\int_0^t v_{ss^is^i}d\tau + 2\left(H(\cdot,0)+\frac{2}{R_0}\right),s\right)^2 ds\\ &\quad + \|v\|^2_{3,\Omega_t} + \|p_\sigma\|^2_{2,\Omega_t}\\ &\le \varepsilon_5\left(\|v_{xxt}\|^2_{0,\Omega_t} + \left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{1,S^1} + \|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\right)\\ &\quad + P_9\left(\|f\|^2_{1,\Omega_t} + |v|^2_{2,1,\Omega_t} + |p_\sigma|^2_{1,0,\Omega_t}\right)\\ &\quad + P_{10}\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^4_{1,S^1} + P_{11}X_5(1+X_5)Y_5,\end{aligned} \tag{4.76}$$

where the summation over the repeated indices (α,β) *and coordinates* $x, s_i = (s^1,s^2), i = 1,2, s = (s^1,s^2)$ *is assumed and*

$$X_5 = \|v\|^2_{3,\Omega_t} + |p_\sigma|^2_{2,1,\Omega_t} + \|v_t\|^2_{1,\Omega_t} + \int_0^t \|v\|^2_{3,\Omega_t} d\tau, \tag{4.77}$$

$$Y_5 = \|v\|^2_{4,\Omega_t} + \|p_\sigma\|^2_{3,\Omega_t} + |p_\sigma|^2_{2,1,\Omega_t} + \|v_t\|^2_{1,\Omega_t} + \int_0^t \|v\|^2_{4,\Omega_t} d\tau. \tag{4.78}$$

Proof. We use the introduced partition of unity. First we consider interior subdomains. We differentiate (4.27) twice with respect to ξ, multiply the result by $\tilde{u}_{\xi\xi}A$, and integrate over $\tilde{\Omega}$ to get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \eta \tilde{u}^2_{\xi\xi} A d\xi + \frac{\mu}{2}\int_{\tilde{\Omega}} (\nabla_{u^i}\tilde{u}^j_{\xi\xi} + \nabla_{u^j}\tilde{u}^i_{\xi\xi})^2 A d\xi \\
&\quad + (\nu-\mu)\|\nabla_u \cdot \tilde{u}_{\xi\xi}\|^2_{0,\tilde{\Omega}} - \int_{\tilde{\Omega}} \tilde{q}_{\sigma\xi\xi}\nabla_u \cdot \tilde{u}_{\xi\xi} A d\xi \\
&\le \delta_1(\|\partial^3_\xi u\|^2_{0,\tilde{\Omega}} + \|\partial^2_\xi q_\sigma\|^2_{0,\tilde{\Omega}}) + a_1(\|u\|^2_{2,\tilde{\Omega}} + \|q_\sigma\|^2_{1,\tilde{\Omega}} + \|\tilde{g}\|^2_{1,\tilde{\Omega}}) \\
&\quad + a_2 X_5(\tilde{\Omega})(1 + X_5(\tilde{\Omega}))Y_5(\tilde{\Omega}),
\end{aligned} \tag{4.79}$$

where

$$\begin{aligned}
X_5(\tilde{\Omega}) &= \|u\|^2_{3,\tilde{\Omega}} + |q_\sigma|^2_{2,1,\tilde{\Omega}} + \|u_t\|^2_{1,\tilde{\Omega}} + \int_0^t \|u\|^2_{3,\tilde{\Omega}} dt', \\
Y_5(\tilde{\Omega}) &= \|u\|^2_{4,\tilde{\Omega}} + \|q_\sigma\|^2_{3,\tilde{\Omega}} + |q_\sigma|^2_{2,1,\tilde{\Omega}} + \|u_t\|^2_{1,\tilde{\Omega}} + \int_0^t \|u\|^2_{4,\tilde{\Omega}} dt'.
\end{aligned} \tag{4.80}$$

From the continuity equation (4.27b) we obtain

$$-\int_{\tilde{\Omega}} \tilde{q}_{\sigma\xi\xi}\nabla_u \cdot \tilde{u}_{\xi\xi} A d\xi = \frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \frac{1}{q\Psi(\eta)}\tilde{q}^2_{\sigma\xi\xi} A d\xi + N_1, \tag{4.81}$$

where

$$|N_1| \le \delta_2 \|\tilde{q}_{\sigma\xi\xi}\|^2_{0,\tilde{\Omega}} + c\|u\|^2_{2,\tilde{\Omega}} + a_3 X_5(\tilde{\Omega})(1 + X_5(\tilde{\Omega}))Y_5(\tilde{\Omega}).$$

Using the form of k_1 (see (4.27a)) from (4.34) we obtain

$$\begin{aligned}
\|\tilde{u}\|^2_{3,\tilde{\Omega}} + \|\tilde{q}_\sigma\|^2_{2,\tilde{\Omega}} &\le a_4(\|\tilde{g}\|^2_{1,\tilde{\Omega}} + \|u\|^2_{2,\tilde{\Omega}} + \|q_\sigma\|^2_{1,\tilde{\Omega}} + \|\tilde{u}_t\|^2_{1,\tilde{\Omega}}) \\
&\quad + a_5\Bigg(\|q_\sigma\|^4_{2,\tilde{\Omega}} + \|q_\sigma\|^2_{2,\tilde{\Omega}}\|\tilde{u}_t\|^2_{1,\tilde{\Omega}} \\
&\qquad + \left\|\int_0^t u dt'\right\|^2_{3,\tilde{\Omega}} \left(1 + \left\|\int_0^t u dt'\right\|^2_{3,\tilde{\Omega}}\right)\|u\|^2_{3,\tilde{\Omega}}\Bigg) \\
&\quad + c\|\nabla_u \cdot \tilde{u}\|^2_{2,\tilde{\Omega}}.
\end{aligned} \tag{4.82}$$

Employing Lemma 5.1 from [35] in the case $G = \tilde{\Omega}$, $v = \tilde{u}_{\xi\xi}$, and (4.79), (4.81), and (4.82) we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \left(\eta\tilde{u}^2_{\xi\xi} + \frac{1}{q\Psi(\eta)}\tilde{q}^2_{\xi\xi}\right) A d\xi + \frac{\mu}{2}\|\tilde{u}\|^2_{3,\tilde{\Omega}} + \|\tilde{q}_\sigma\|^2_{2,\tilde{\Omega}} \\
&\le \delta_3(\|\partial^3_\xi u\|^2_{0,\tilde{\Omega}} + \|\partial^2_\xi q_\sigma\|^2_{0,\tilde{\Omega}}) + a_6(\|\tilde{g}\|^2_{1,\tilde{\Omega}} + |u|^2_{2,1,\tilde{\Omega}} + \|\tilde{q}_\sigma\|^2_{1,\tilde{\Omega}}) \\
&\quad + a_7 X_5(\tilde{\Omega})(1 + X_5(\tilde{\Omega}))Y_5(\tilde{\Omega}).
\end{aligned} \tag{4.83}$$

Now we consider a subdomain near the boundary. Differentiating (4.28a) twice with respect to τ, multiplying the result by $\tilde{u}_{\tau\tau}J$, and integrating over $\hat{\Omega}$ yields

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}_{\tau\tau}+\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma\tau\tau}\right)Jdz+\frac{\mu}{2}\|\tilde{u}_{\tau\tau}\|^2_{1,\hat{\Omega}}\\
&\qquad-\int_{\hat{S}}(\hat{n}\hat{T}(\tilde{u},\tilde{q}_\sigma))_{,\tau\tau}\cdot\tilde{u}_{\tau\tau}Jdz'\\
&\leq\delta_4(\|\hat{q}_{\sigma zz}\|^2_{0,\hat{\Omega}}+\|\hat{u}_{zzz}\|^2_{0,\hat{\Omega}})\\
&\qquad+a_8(\|\hat{u}\|^2_{2,\hat{\Omega}}+\|\hat{q}_\sigma\|^2_{1,\hat{\Omega}}+\|\tilde{g}\|^2_{1,\hat{\Omega}})+a_9X_5(\hat{\Omega})(1+X_5(\hat{\Omega})Y_5(\hat{\Omega}),
\end{aligned}
\tag{4.84}
$$

where the $X_5(\hat{\Omega}), Y_5(\hat{\Omega})$ are defined by (4.80) with $\hat{\Omega}$ instead of $\tilde{\Omega}$, and u, q_σ are replaced by $\hat{u}, \hat{q}_\sigma$. We have used Lemma 5.1 from [35] in the case $G=\hat{\Omega}, v=\tilde{u}_{\tau\tau}$, and

$$
-\int_{\hat{\Omega}}\tilde{q}_{\sigma\tau\tau}\hat{\nabla}\cdot\tilde{u}_{\tau\tau}Jdz=\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma\tau\tau}Jdz+N_2,
$$

where

$$
|N_2|\leq\delta_5\|\tilde{q}_{\sigma\tau\tau}\|^2_{0,\hat{\Omega}}+a_{10}\|\hat{u}\|^2_{2,\hat{\Omega}}+a_{11}X_5(\hat{\Omega})(1+X_5(\hat{\Omega}))Y_5(\hat{\Omega}).
$$

Considering the boundary term in (4.84) we obtain the expression

$$
\begin{aligned}
&\int_{\hat{S}}(\hat{n}\cdot\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,\tau\tau}\cdot\tilde{u}_{\tau\tau}Jdz'\\
&=-\int_{\hat{S}}\left(\hat{\Delta}_{\hat{S}_t}\hat{\xi}\cdot\hat{n}\hat{n}\hat{\zeta}+\frac{2}{R_0}\hat{n}\hat{\zeta}\right)_{,\tau\tau}\tilde{u}_{\tau\tau}Jdz'\\
&\quad-\sigma\int_{\hat{S}}\left(\hat{\Delta}_{\hat{S}_t}\int_0^t\tilde{u}d\tau\cdot\hat{n}\hat{n}\right)_{,\tau\tau}\tilde{u}_{\tau\tau}Jdz'+\int_{\hat{S}}(k_5+k_6)_{,\tau\tau}\tilde{u}_{\tau\tau}Jdz'.
\end{aligned}
\tag{4.85}
$$

Similarly, as in the case of (4.38), the first term on the right-hand side is equal to

$$
\begin{aligned}
&-\sigma\int_U\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\bar{n}\zeta\right)_{,s_1s_2}\tilde{u}_{s_1s_2}\sqrt{g}ds^1ds^2+N_3\\
&=\sigma\int_U\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\bar{n}\zeta\right)_{,s_1}\tilde{u}_{s_1s_2s_2}\sqrt{g}ds^1ds^2+N_4,
\end{aligned}
$$

where the summation over the repeated indices s_1, s_2 is assumed, and where

$$
\begin{aligned}
|N_4|\leq\delta_6&\left(\left\|\int_0^t\hat{u}d\tau\right\|^2_{3,\hat{S}}+\|\hat{u}_{zzz}\|^2_{0,\hat{\Omega}}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{1,\hat{S}}\right)\\
&+a_{12}(\|\xi\|_4)\|\hat{u}\|^2_{2,\hat{\Omega}}+a_{13}\|\hat{u}\|^2_{3,\hat{\Omega}}\left\|\int_0^t\hat{u}d\tau\right\|^2_{3,\hat{S}}.
\end{aligned}
$$

The second term on the right-hand side of (4.85) takes the form

$$
\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}g^{\alpha\beta}\bar{n}\cdot\int_0^t\tilde{u}_{s_1s_2s^\alpha}d\tau\bar{n}\cdot\int_0^t\tilde{u}_{s_1s_2s^\beta}d\tau\,ds+N_5,
$$

and

$$|N_5| \le \delta_7 \left(\left\| \int_0^t \tilde{u} d\tau \right\|_{3,\hat{S}}^2 + \|\tilde{u}_{zzz}\|_{0,\hat{\Omega}}^2 \right) + a_{14}\|\hat{u}\|_{2,\hat{\Omega}}^2 + a_{15}\|\hat{u}\|_{3,\hat{\Omega}}^2 \left\| \int_0^t \tilde{u} d\tau \right\|_{3,\hat{S}}^2.$$

Finally, the last term in (4.85) is estimated by

$$\delta_8 \left(\left\| \int_0^t \tilde{u} d\tau \right\|_{3,\hat{S}}^2 + \|\hat{u}_{zzz}\|_{0,\hat{\Omega}}^2 \right) + a_{16}\|\hat{u}\|_{0,\hat{\Omega}}^2.$$

By summarizing, we have

$$\begin{aligned}
&\int_{\hat{S}} (\hat{n}\mathbb{T}(\tilde{u}, \tilde{q}_\sigma))_{,\tau\tau} \tilde{u}_{\tau\tau} J dz' \\
&\quad \le -\frac{\sigma d}{2dt} \int_{\hat{S}} g^{\alpha\beta} \bar{n} \cdot \int_0^t \tilde{u}_{s_1 s_2 s^\alpha} d\tau \bar{n} \cdot \int_0^t \tilde{u}_{s_1 s_2 s^\beta} d\tau J dz' \\
&\quad\quad -\sigma \int_{\hat{S}} \left(\left(H(\cdot, 0) + \frac{2}{R_0} \right) \bar{n}\zeta \right)_{,s_1} \tilde{u}_{s_1 s_2 s_2} J dz' \\
&\quad\quad +\delta_9 \left(\left\| \int_0^t \tilde{u} d\tau \right\|_{3,\hat{S}}^2 + \|\hat{u}_{zzz}\|_{0,\hat{\Omega}}^2 + \left\| H(\cdot, 0) + \frac{2}{R_0} \right\|_{1,\hat{S}}^2 \right) \\
&\quad\quad +a_{17}\|\hat{u}\|_{2,\hat{\Omega}}^2 + a_{18}\|\hat{u}\|_{3,\hat{\Omega}}^2 \left\| \int_0^t \tilde{u} d\tau \right\|_{3,\hat{S}}^2.
\end{aligned} \tag{4.86}$$

By summarizing, we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt} \int_{\hat{\Omega}} \left(\hat{\eta}\tilde{u}_{\tau\tau}^2 + \frac{1}{\hat{q}\Psi(\hat{\eta})} \tilde{q}_{\sigma\tau\tau}^2 \right) J dz \\
&\quad +\frac{\sigma}{2}\frac{d}{dt} \int_{\hat{S}} g^{\alpha\beta} \bar{n} \cdot \int_0^t \tilde{u}_{s_1 s_2 s^\alpha} d\tau \bar{n} \cdot \int_0^t \tilde{u}_{s_1 s_2 s^\beta} d\tau ds \\
&\quad +\sigma \frac{d}{dt} \int_{\hat{S}} \left(\left(H(\cdot, 0) + \frac{2}{R_0} \right) \zeta \right)_{,s_1} \tilde{u}_{s_1 s_2 s_2} \cdot \bar{n} ds + \frac{\mu}{2} \|\tilde{u}_{\tau\tau}\|_{1,\hat{\Omega}}^2 \\
&\le \delta_{10} \left(\|\hat{u}_{zzz}\|_{0,\hat{\Omega}}^2 + \|\hat{q}_{\sigma zz}\|_{0,\hat{\Omega}}^2 + \left\| \int_0^t \tilde{u} d\tau \right\|_{3,\hat{S}}^2 \right. \\
&\quad \left. + \left\| H(\cdot, 0) + \frac{2}{R_0} \right\|_{1,\hat{S}}^2 + \|R(\cdot, t) - R(\cdot, 0)\|_{3,S^1}^2 \right) \\
&\quad +a_{19}(\|\hat{u}\|_{2,\hat{\Omega}}^2 + \|\hat{q}_\sigma\|_{1,\hat{\Omega}}^2 + \|\tilde{g}\|_{1,\hat{\Omega}}^2) + a_{20} X_5(\hat{\Omega})(1 + X_5(\hat{\Omega})) Y_5(\hat{\Omega}).
\end{aligned} \tag{4.87}$$

Differentiating the third component of (4.42) with respect to τ, multiplying the

result by $\tilde{q}_{\sigma n\tau}J$, and integrating over $\hat{\Omega}$ yields

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma n\tau}Jdz+\|\tilde{q}_{\sigma n\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le c\|\tilde{u}_{z\tau\tau}\|^2_{0,\hat{\Omega}}\\
&\qquad+(\delta_{11}+cd)(\|\hat{u}_{zzz}\|^2_{0,\hat{\Omega}}+\|\hat{q}_{\sigma zz}\|^2_{0,\hat{\Omega}})\|\hat{F}\|^2_{4+1/2,\hat{S}}\\
&\qquad+a_{21}(|\hat{u}|^2_{2,1,\hat{\Omega}}+\|\hat{q}_{\sigma}\|^2_{1,\hat{\Omega}}+\|\tilde{g}\|^2_{1,\hat{\Omega}})\\
&\qquad+a_{22}X_5(\hat{\Omega})(1+X_5(\hat{\Omega}))Y_5(\hat{\Omega}).
\end{aligned}
\tag{4.88}
$$

Differentiating the third component of (4.44) with respect to τ, multiplying the result by $\tilde{u}^3_{nn\tau}J$ and integrating over $\hat{\Omega}$ gives

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}|\tilde{u}^3_{n\tau}|^2Jdz+\frac{\mu+\nu}{2}\|\tilde{u}^3_{nn\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le c(\|\tilde{u}_{z\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma n\tau}\|^2_{0,\hat{\Omega}})\\
&\qquad+(\delta_{12}+cd)(\|\tilde{u}_{zzz}\|^2_{0,\hat{\Omega}}+\|\hat{q}_{\sigma zz}\|^2_{0,\hat{\Omega}})\|\hat{F}\|^2_{4+1/2,\hat{S}}\\
&\qquad+a_{23}\|\tilde{u}_t\|^2_{1,\hat{\Omega}}+a_{24}(\|\hat{u}\|^2_{2,\hat{\Omega}}+\|\hat{q}_{\sigma}\|^2_{1,\hat{\Omega}}+\|\tilde{g}\|^2_{1,\hat{\Omega}})\\
&\qquad+a_{25}X_5(\hat{\Omega})(1+X_5(\hat{\Omega}))Y_5(\hat{\Omega}).
\end{aligned}
\tag{4.89}
$$

By differentiating (4.46) twice with respect to τ, multiplying by $\tilde{u}'_{\tau\tau}J$, integrating over $\hat{\Omega}$, and using the boundary condition (4.47), we get

$$
\begin{aligned}
&\|\tilde{u}'_{z\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma\tau\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le(\delta_{13}+cd)(\|\hat{u}_{zzz}\|^2_{0,\hat{\Omega}}+\|\hat{q}_{\sigma zz}\|^2_{0,\hat{\Omega}})\|\hat{F}\|^2_{4+1/2,\hat{S}}\\
&\qquad+a_{26}(\|\operatorname{div}\tilde{u}_\tau\|^2_{1,\hat{\Omega}}+\|\tilde{u}_t\|^2_{1,\hat{\Omega}}+\|\hat{u}\|^2_{2,\hat{\Omega}}+\|\hat{q}_{\sigma}\|^2_{1,\hat{\Omega}}+\|\tilde{g}\|^2_{1,\hat{\Omega}})\\
&\qquad+a_{27}X_5(\hat{\Omega})Y_5(\hat{\Omega}).
\end{aligned}
\tag{4.90}
$$

Moreover, from (4.46) we obtain

$$
\begin{aligned}
\|\tilde{u}'_{nn\tau}\|^2_{1,\hat{\Omega}}\le{}&c(\|\tilde{u}'_{\tau\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma\tau\tau}\|^2_{0,\hat{\Omega}}+\|\operatorname{div}\tilde{u}_{\tau\tau}\|^2_{0,\hat{\Omega}})\\
&+(\delta_{14}+cd)(\|\hat{u}_{zzz}\|^2_{0,\hat{\Omega}}+\|\hat{q}_{\sigma zz}\|^2_{0,\hat{\Omega}})\|\hat{F}\|^2_{4+1/2,\hat{S}}\\
&+a_{28}(\|\tilde{u}_t\|^2_{1,\hat{\Omega}}+\|\tilde{u}\|^2_{2,\hat{\Omega}}+\|\hat{q}_{\sigma}\|^2_{1,\hat{\Omega}}+\|\hat{g}\|^2_{1,\hat{\Omega}})\\
&+a_{29}X_5(\hat{\Omega})(1+X_5(\hat{\Omega}))Y_5(\hat{\Omega}).
\end{aligned}
\tag{4.91}
$$

By summarizing, inequalities (4.87)–(4.91) we get the following:

(4.92)
$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left[\hat\eta(\tilde u_{\tau\tau}^2+|\tilde u_{n\tau}^3|^2)+\frac{1}{\hat q\hat\Psi(\hat\eta)}\tilde q_{\sigma z\tau}^2\right]Jdz\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{\tilde S}g^{\alpha\beta}\bar n\cdot\int_0^t\tilde u_{s_1s_2s^\alpha}d\tau\bar n\cdot\int_0^t\tilde u_{s_1s_2s^\beta}d\tau ds\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{\tilde S}\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)_{,s_1}\tilde u_{s_1s_2s_2}\cdot\bar n ds\\
&\qquad+\|\tilde u_\tau\|_{2,\hat\Omega}^2+\|\tilde q_{\sigma\tau}\|_{1,\hat\Omega}^2\\
&\quad\le(\delta_{15}+cd)(\|\hat u_{zzz}\|_{0,\hat\Omega}^2+\|\hat q_{\sigma zz}\|_{0,\hat\Omega}^2)\|\hat F\|_{4+1/2,\hat S}^2\\
&\qquad+\delta_{16}\left(\left\|\int_0^t\tilde u d\tau\right\|_{3,\hat S}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{1,S^1}^2+\|R(\cdot,t)-R(\cdot,0)\|_{3,S^1}^2\right)\\
&\qquad+a_{30}(|\hat u|_{2,1,\hat\Omega}^2+\|\hat q_\sigma\|_{1,\hat\Omega}^2+\|\tilde g\|_{1,\hat\Omega}^2)+a_{31}X_5(\hat\Omega)(1+X_5(\hat\Omega))Y_5(\hat\Omega).
\end{aligned}$$

By differentiating the third component of (4.42) with respect to n, multiplying the result by $\tilde q_{\sigma nn}J$, and then integrating over $\hat\Omega$ implies

(4.93)
$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\frac{\mu+\nu}{\hat q\hat\Psi(\hat\eta)}\tilde q_{\sigma nn}^2Jdz+\|\tilde q_{\sigma nn}\|_{0,\hat\Omega}^2\\
&\le c(\|\tilde u_\tau\|_{2,\hat\Omega}^2+(\delta_{16}+cd)(\|\tilde u_{zzz}\|_{0,\hat\Omega}^2+\|\tilde q_{\sigma zz}\|_{0,\hat\Omega}^2))\|\hat F\|_{4+1/2,\hat S}^{82}+\delta_{17}\left\|\int_0^t\tilde u d\tau\right\|_{3,\hat S}^2\\
&\quad+a_{32}(\|\tilde u_t\|_{1,\hat\Omega}^2+\|\hat u\|_{2,\hat\Omega}^2+\|\hat q_\sigma\|_{1,\hat\Omega}^2+\|\tilde g\|_{1,\hat\Omega}^2)\\
&\quad+a_{33}X_5(\hat\Omega)(1+X_5(\hat\Omega))Y_5(\hat\Omega).
\end{aligned}$$

We write (4.28a) in the form

$$\begin{aligned}
(\mu+\nu)\nabla_{z^i}\operatorname{div}\tilde u=&-\mu(\Delta\tilde u^i-\nabla_{z^i}\operatorname{div}\tilde u)+\hat\eta\tilde u_t^i-\hat\eta\tilde g^i-k_3^i\\
&-[\mu\nabla^2\tilde u^i+\nu\nabla_{z^i}\operatorname{div}\tilde u-\mu\hat\nabla^2\tilde u^i-\nu\hat\nabla_i\hat{\operatorname{div}}\tilde u]-\hat\nabla_i\tilde q_\sigma.
\end{aligned}\tag{4.94}$$

Differentiating the third component of (4.94) with respect to n gives

$$\begin{aligned}
\|(\operatorname{div}\tilde u)_{nn}\|_{0,\hat\Omega}^2\le&\,c(\|\tilde u_\tau\|_{2,\hat\Omega}^2+\|\tilde q_{\sigma nn}\|_{0,\tilde\Omega}^2)\\
&+(\delta_{18}+cd)\|\tilde u\|_{3,\tilde\Omega}^2\|\hat F\|_{4+1/2,\hat S}^2\\
&+a_{34}(|\hat u|_{2,1,\hat\Omega}^2+\|\hat q_\sigma\|_{1,\hat\Omega}^2+\|\tilde g\|_{1,\hat\Omega}^2)\\
&+a_{35}X_5(\hat\Omega)(1+X_5(\hat\Omega))Y_5(\hat\Omega).
\end{aligned}\tag{4.95}$$

Finally, differentiating (4.46) with respect to n yields

$$\begin{aligned}
\|\tilde u_{nnn}\|_{0,\hat\Omega}^2\le&\,c(\|\tilde u_{\tau\tau}\|_{1,\hat\Omega}^2+\|\tilde q_{\sigma n}\|_{1,\hat\Omega}^2+\|(\operatorname{div}\tilde u)_n\|_{1,\hat\Omega}^2)\\
&+(\delta_{19}+cd)\|\tilde u\|_{3,\hat\Omega}^2\|\hat F\|_{4+1/2,\hat S}^2\\
&+a_{36}(|\hat u|_{2,1,\hat\Omega}^2+\|\hat q_\sigma\|_{1,\hat\Omega}^2+\|\tilde g\|_{1,\hat\Omega}^2)\\
&+a_{37}X_5(\hat\Omega)(1+X_5(\hat\Omega))Y_5(\hat\Omega).
\end{aligned}\tag{4.96}$$

From (4.92), (4.93), (4.95), and (4.96) we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left[\hat{\eta}(\tilde{u}_{\tau\tau}^2+|\tilde{u}_{n\tau}^3|^2)+\frac{\mu+\nu}{\hat{q}\psi(\hat{\eta})}\tilde{q}_{\sigma zz}^2\right]Jdz\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{\tilde{S}}g^{\alpha\beta}\bar{n}\cdot\int_0^t\tilde{u}_{s_1s_2s^\alpha}d\tau\bar{n}\cdot\int_0^t\tilde{u}_{s_1s_2s^\beta}d\tau ds\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{\tilde{S}}\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)_{,s_1}\tilde{u}_{s_1s_2s_2}\cdot\bar{n}ds+\|\tilde{u}\|_{3,\hat{\Omega}}^2+\|\tilde{q}_\sigma\|_{2,\hat{\Omega}}^2\\
&\qquad\le c(\delta_{20}+cd)(\|\hat{u}_{zzz}\|_{0,\hat{\Omega}}^2+\|\hat{q}_{\sigma zz}\|_{0,\hat{\Omega}}^2)\\
&\qquad+\delta_{21}\left(\left\|\int_0^t\tilde{u}d\tau\right\|_{3,\hat{S}}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{1,S^1}^2+\|R(\cdot,t)-R(\cdot,0)\|_{3,S^1}^2\right)\\
&\qquad+a_{38}(|\hat{u}|_{2,1,\hat{\Omega}}^2+\|\hat{q}_\sigma\|_{1,\hat{\Omega}}^2+\|\tilde{g}\|_{1,\hat{\Omega}}^2)+a_{39}X_5(\hat{\Omega})(1+X_5(\hat{\Omega}))Y_5(\hat{\Omega}).
\end{aligned}
\tag{4.97}
$$

To obtain the full second derivative of u under the derivative with respect to time, we examine the expression

$$
\begin{aligned}
\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}\tilde{u}_{zz}^2Jdz&=\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}_{zz}\cdot\tilde{u}_{zzt}J+\frac{1}{2}\hat{\eta}_t\tilde{u}_{zz}^2J+\frac{1}{2}\hat{\eta}\tilde{u}_{zz}^2J_t\right)dz\\
&\le\delta_{22}(\|\tilde{u}_{zzt}\|_{0,\hat{\Omega}}^2+\|\tilde{u}_{zz}\|_{1,\hat{\Omega}}^2)+a_{40}\|\hat{u}\|_{2,\hat{\Omega}}^2\|\tilde{u}\|_{2,\hat{\Omega}}^2,
\end{aligned}
\tag{4.98}
$$

where we have used the relations

$$
\hat{\eta}_t+\hat{\eta}\hat{\nabla}\cdot\hat{u}=0\quad\text{and}\quad J_t=J\hat{\nabla}\cdot\hat{u}.
\tag{4.99}
$$

Employing (4.98) in (4.97) and using the fact that δ_{22} is sufficiently small, we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}_{zz}^2+\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}_{\sigma zz}^2\right)Jdz\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{\tilde{S}}g^{\alpha\beta}\bar{n}\cdot\int_0^t\tilde{u}_{s_1s_2s^\alpha}d\tau\bar{n}\cdot\int_0^t\tilde{u}_{s_1s_2s^\beta}d\tau\,ds\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{\tilde{S}}\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)_{,s_1}\tilde{u}_{s_1s_2s_2}\cdot\bar{n}ds\\
&\qquad+\|\tilde{u}\|_{3,\hat{\Omega}}^2+\|\tilde{q}_\sigma\|_{2,\hat{\Omega}}^2\\
&\qquad\le c(\delta_{23}+cd)(\|\hat{u}_{zzt}\|_{0,\hat{\Omega}}^2+\|\hat{u}_{zzz}\|_{0,\hat{\Omega}}^2+\|\hat{q}_{\sigma zz}\|_{0,\hat{\Omega}}^2)\\
&\qquad+\delta_{24}\left(\left\|\int_0^t\tilde{u}d\tau\right\|_{3,\hat{S}}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{1,S^1}^2+\|R(\cdot,t)-R(\cdot,0)\|_{3,S^1}^2\right)\\
&\qquad+a_{41}(|\hat{u}|_{2,1,\hat{\Omega}}^2+\|\hat{q}_\sigma\|_{1,\hat{\Omega}}^2+\|\tilde{g}\|_{1,\hat{\Omega}}^2)\\
&\qquad+a_{42}X_5(\hat{\Omega})(1+X_5(\hat{\Omega}))Y_5(\hat{\Omega}).
\end{aligned}
\tag{4.100}
$$

Now we examine the second and the third terms in the left-hand side of (4.100). Applying the same considerations as they were used in the case of inequalities (4.69)

and (4.70), we obtain that both terms are equal to

$$
\begin{aligned}
&\frac{\sigma}{2}\frac{d}{dt}\int_U \frac{1}{2}\tilde{\delta}^{\alpha\beta}\bar{n}\cdot\int_0^t \tilde{u}_{s_1s_2s^\alpha}d\tau\,\bar{n}\cdot\int_0^t \tilde{u}_{s_1s_2s^\beta}d\tau\sqrt{g}ds^1ds^2\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_U\left|\bar{n}\cdot\int_0^t \tilde{u}_{ss_1s_2}d\tau\right|^2\sqrt{g}ds^1ds^2\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_U\sum_{i=1}^2\left(\frac{1}{2}\bar{n}\cdot\int_0^t \tilde{u}_{ss^is^i}d\tau+2\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)_{,s}\right)^2\sqrt{g}ds^1ds^2\\
&\quad-4\sigma\frac{d}{dt}\int_U\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)_{,s}^2\sqrt{g}ds^1ds^2,
\end{aligned}
\tag{4.101}
$$

where $\tilde{\delta}^{\alpha\beta}$ is defined in (4.70).

Now, going back to the variables ξ in the inequality (4.101), summing over all neighborhoods of the partition of unity (for the interior neighborhoods we use the inequality (4.83)), and then going back to the variables x, we obtain for sufficiently small δ_1 and d the inequality

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v_{xx}^2+\frac{1}{\rho\Psi(\rho)}p_{\sigma xx}^2\right)dx\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\frac{1}{2}\tilde{\delta}^{\alpha\beta}\bar{n}\cdot\int_0^t v_{s_1s_2s^\alpha}d\tau\,\bar{n}\cdot\int_0^t v_{s_1s_2s^\beta}d\tau\,ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\left|\bar{n}\cdot\int_0^t v_{ss^1s^2}d\tau\right|^2 ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\sum_{i=1}^2\left(\frac{1}{2}\bar{n}\cdot\int_0^t v_{ss^is^i}d\tau+2\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,s}\right)^2 ds\\
&\quad+\frac{\mu}{2}\|v\|_{3,\Omega_t}^2+\|p_\sigma\|_{2,\Omega_t}^2\le\delta_{24}\left(\left\|\int_0^t v d\tau\right\|_{3,S_t}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{1,S^1}^2\right)\\
&\quad+a_{42}(|v|_{2,1,\Omega_t}^2+\|\hat{p}_\sigma\|_{1,\Omega_t}^2+\|f\|_{1,\Omega_t}^2)\\
&\quad+a_{43}X_5(1+X_5)Y_5+4\sigma\frac{d}{dt}\int_{S_t}\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,s}\right)^2 ds.
\end{aligned}
\tag{4.102}
$$

In virtue of the Young and Hölder inequalities we get

$$
\begin{aligned}
&\left|\frac{d}{dt}\int_{S_t}\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,s}\right)^2 ds\right|\\
&\qquad\le\delta_{25}\|v_{xxx}\|_{0,\Omega_t}^2+a_{44}\left(\|v\|_{0,\Omega_t}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{1,S^1}^4\right).
\end{aligned}
\tag{4.103}
$$

Finally, using (4.73) and repeating the considerations following (4.75) gives

$$
\begin{aligned}
\left\| \int_0^t v d\tau \right\|^2_{3,S_t} &\le \delta_{26}(\|v\|^2_{3,\Omega_t} + \|p_\sigma\|^2_{2,\Omega_t}) \\
&\quad + a_{45}\left(\left\| \int_0^t v d\tau \right\|^2_{0,\Omega_t} + \|v\|^2_{0,\Omega_t} + \|p_\sigma\|^2_{0,\Omega_t} \right. \\
&\quad \left. + \left\| H(\cdot,0) + \frac{2}{R_0} \right\|^2_{1,S^1} + \|R(\cdot,t) - R(\cdot,0)\|^2_{3,S^1} \right) \\
&\quad + a_{46}(\|v\|^2_{3,\Omega_t} + \|p_\sigma\|^2_{2,\Omega_t}) \int_0^t \|v\|^2_{4,\Omega_\tau} d\tau.
\end{aligned} \tag{4.104}
$$

Therefore, from (4.102)–(4.104) for sufficiently small δ_{24}, δ_{25}, δ_{26}, we obtain (4.76). This concludes the proof.

To estimate the first term in the right-hand side of (4.77) we need the following result.

LEMMA 4.6. *For sufficiently smooth solutions of problem* (4.1a–c) *we have*

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t} \left(\rho v_{xt}^2 + \frac{1}{\rho \Psi(\rho)} p_{\sigma xt}^2 \right) dx + \frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta} \bar{n} \cdot v_{ss^\alpha} \bar{n} \cdot v_{ss^\beta} ds \\
&\quad + \|v_t\|^2_{2,\Omega_t} + \|p_{\sigma t}\|^2_{1,\Omega_t} \\
&\le \varepsilon_6 \left(\|v_{xtt}\|^2_{0,\Omega_t} + \|v_{xxx}\|^2_{0,\Omega_t} + \left\| H(\cdot,0) + \frac{2}{R_0} \right\|^2_{0,s^1} \right) \\
&\quad + P_{12}(|v|^2_{2,0,\Omega_t} + |p_\sigma|^2_{1,0,\Omega_t} + |f|^2_{1,0,\Omega_t}) \\
&\quad + P_{13} X_6 (1 + X_6) Y_6,
\end{aligned} \tag{4.105}
$$

where

$$
\begin{aligned}
X_6 &= |v|^2_{3,1,\Omega_t} + |p_\sigma|^2_{2,0,\Omega_t} + \int_0^t \|v\|^2_{3,\Omega_\tau} d\tau, \\
Y_6 &= |v|^2_{4,2,\Omega_t} + |p_\sigma|^2_{3,1,\Omega_\tau} + \int_0^t \|v\|^2_{4,\Omega_\tau} d\tau.
\end{aligned} \tag{4.106}
$$

Proof. We use the partition of unity. First we obtain the inequality in an interior subdomain. Differentiating (4.27a) with respect to t and ξ, multiplying the result by $\tilde{u}_{t\xi}$, and integrating over $\tilde{\Omega}$ yields

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \eta \tilde{u}^2_{t\xi} A d\xi + \frac{\mu}{2}\int_{\tilde{\Omega}} (\nabla_{u^i} \tilde{u}^j_{t\xi} + \nabla_{u^j} \tilde{u}^i_{t\xi})^2 A d\xi \\
&\quad + (\nu - \mu)\|\nabla_u \cdot \tilde{u}_{t\xi}\|^2_{0,\tilde{\Omega}} - \int_{\tilde{\Omega}} \tilde{q}_{\sigma t\xi} \nabla_u \cdot \tilde{u}_{t\xi} A d\xi \\
&\le \delta_1 \|\tilde{u}_{t\xi}\|^2_{1,\tilde{\Omega}} + a_1(\|u_t\|^2_{1,\tilde{\Omega}} + \|q_{\sigma t}\|^2_{0,\tilde{\Omega}} + |\tilde{g}|^2_{1,0,\tilde{\Omega}}) \\
&\quad + a_2 X_6(\tilde{\Omega}) Y_6(\tilde{\Omega}),
\end{aligned} \tag{4.107}
$$

where

$$X_6(\tilde{\Omega}) = |u|^2_{3,1,\tilde{\Omega}} + |q_\sigma|^2_{2,0,\tilde{\Omega}} + \left\| \int_0^t u dt' \right\|^2_{3,\tilde{\Omega}},$$

$$Y_6(\tilde{\Omega}) = |u|^2_{4,2,\tilde{\Omega}} + |q_\sigma|^2_{3,1,\tilde{\Omega}} + \left\| \int_0^t u dt' \right\|^2_{4,\tilde{\Omega}}.$$

By the continuity equation (4.27b) we have

$$-\int_{\tilde{\Omega}} \tilde{q}_{\sigma t\xi} \nabla_u \cdot \tilde{u}_{t\xi} A d\xi = \frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \frac{1}{q\Psi(\eta)} \tilde{q}^2_{\sigma t\xi} A d\xi + N_1, \tag{4.108}$$

where

$$|N_1| \le \delta_1 \|\tilde{q}_{\sigma t}\|^2_{1,\tilde{\Omega}} + a_3 |u|^2_{2,1,\tilde{\Omega}} + a_4 X_6(\tilde{\Omega})(1 + X_6(\tilde{\Omega}))Y_6(\tilde{\Omega}).$$

From (4.34) we obtain

$$\begin{aligned} \|\tilde{u}_t\|^2_{2,\tilde{\Omega}} + \|\tilde{q}_{\sigma t}\|^2_{1,\tilde{\Omega}} &\le c\|(\nabla_u \cdot \tilde{u})_t\|^2_{1,\tilde{\Omega}} + a_5(\|\tilde{u}_{tt}\|^2_{0,\tilde{\Omega}} \\ &\quad + |u|^2_{2,1,\tilde{\Omega}} + |q_{\sigma t}|^2_{1,0,\tilde{\Omega}} + |\tilde{g}|^2_{1,0,\tilde{\Omega}}) + a_6 X_6(\tilde{\Omega}) Y_6(\tilde{\Omega}). \end{aligned} \tag{4.109}$$

Employing Lemma 5.1 from [35] in the case $G = \tilde{\Omega}$, $v = \tilde{u}_{\xi\xi}$, and (4.107)–(4.109) we obtain for sufficiently small δ_1 and δ_2,

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \left(\eta \tilde{u}^2_{t\xi} + \frac{1}{q\Psi(\eta)} \tilde{q}^2_{\sigma t\xi} \right) A d\xi + \|\tilde{u}_t\|^2_{2,\tilde{\Omega}} + \|\tilde{q}_{\sigma t}\|^2_{1,\tilde{\Omega}} \\ &\quad \le a_7(\|\tilde{u}_{tt}\|^2_{0,\tilde{\Omega}} + |u|^2_{2,1,\tilde{\Omega}} + |q_\sigma|^2_{1,0,\tilde{\Omega}} + |\tilde{g}|^2_{1,0,\tilde{\Omega}}) \\ &\quad + a_8 X_6(\tilde{\Omega})(1 + X_6(\tilde{\Omega}))Y_6(\tilde{\Omega}). \end{aligned} \tag{4.110}$$

Now we obtain an estimate in a subdomain near the boundary. Differentiating (4.28a) with respect to t and τ, multiplying the result by $\tilde{u}_{t\tau} J$, and integrating over $\hat{\Omega}$ yields

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}} \left(\hat{\eta} \tilde{u}^2_{t\tau} + \frac{1}{\hat{q}\Psi(\hat{\eta})} \tilde{q}^2_{\sigma t\tau} \right) J dz + \frac{\mu}{2}\|\tilde{u}_{t\tau}\|^2_{1,\hat{\Omega}} \\ &\quad - \int_{\hat{S}} (\hat{n}\hat{\mathbb{T}}(\tilde{u}, \tilde{q}_\sigma))_{,t\tau} \cdot \tilde{u}_{t\tau} J dz' \le \delta_3(\|\tilde{u}_{zzt}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_\sigma\|^2_{1,\hat{\Omega}}) \\ &\quad + a_9(|\hat{u}|^2_{2,0,\hat{\Omega}} + |\hat{q}_\sigma|^2_{1,0,\hat{\Omega}} + |\tilde{g}|^2_{1,0,\hat{\Omega}}) \\ &\quad + a_{10} X_6(\hat{\Omega})(1 + X_6(\hat{\Omega}))Y_6(\hat{\Omega}), \end{aligned} \tag{4.111}$$

where $X_6(\hat{\Omega})$, $Y_6(\hat{\Omega})$ are equal to $X_6(\tilde{\Omega})$, $Y_6(\tilde{\Omega})$ with $\hat{u}$, $\hat{q}_\sigma$, $\hat{\Omega}$ instead of u, q_σ, $\hat{\Omega}$, respectively. Moreover, to obtain (4.111) we have used Lemma 5.1 from [35] in the case $G = \hat{\Omega}$, $v = \tilde{u}_{t\tau}$, and

$$-\int_{\hat{\Omega}} \tilde{q}_{\sigma t\tau} \hat{\nabla} \cdot \tilde{u}_{t\tau} J dz = \frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}} \frac{1}{\hat{q}\Psi(\hat{\eta})} \tilde{q}^2_{\sigma t\tau} J dz + N_2,$$

and

$$|N_2| \le \delta_4 \|\tilde{q}_{\sigma t}\|^2_{1,\hat{\Omega}} + a_{11}|\hat{u}|^2_{2,1,\hat{\Omega}} + a_{12}X_6(\hat{\Omega})(1+X_6(\hat{\Omega}))Y_6(\hat{\Omega}).$$

Let us consider the boundary term in (4.111). Using the boundary condition (4.28c) we obtain

$$\begin{aligned} &-\int_{\hat{S}} (\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,t\tau} \cdot \tilde{u}_{t\tau} J dz' \\ &= -\sigma \int_{\hat{S}} \left(\hat{\Delta}_{\hat{S}_t} \hat{\xi} \cdot \hat{n}\hat{n}\hat{\zeta} + \frac{2}{R_0}\hat{n}\hat{\zeta} \right)_{,t\tau} \tilde{u}_{t\tau} J dz' \\ &\quad -\sigma \int_{\hat{S}} \left(\hat{\Delta}_{\hat{S}_t} \int_0^t \tilde{u} d\tau \cdot \hat{n}\hat{n} \right)_{,t\tau} \tilde{u}_{t\tau} J dz' + \int_{\hat{S}} (k_5 + k_6)_{,t\tau} \tilde{u}_{t\tau} J dz'. \end{aligned} \tag{4.112}$$

The first term on the right-hand side of (4.112) is estimated by

$$\begin{aligned} &\delta_5 \left(\|\tilde{u}_t\|^2_{2,\hat{\Omega}} + \|\hat{u}\|^2_{3,\hat{\Omega}} + \left\| H(\cdot,0) + \frac{2}{R_0} \right\|^2_{1,S^1} \right) + a_{13}(\|\hat{u}\|^4_{3,\hat{\Omega}} + \|\hat{u}\|^2_{3,\hat{\Omega}} \|\hat{u}_t\|^2_{2,\hat{\Omega}}) \\ &\quad + a_{14}(\|\hat{u}\|^2_{0,\hat{\Omega}} + \|\hat{u}_t\|^2_{0,\hat{\Omega}}). \end{aligned}$$

The second term is equal to

$$\frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta} \bar{n} \cdot \tilde{u}_{ss^\alpha} \bar{n} \cdot \tilde{u}_{ss^\beta} ds + N_3,$$

where

$$\begin{aligned} |N_3| \le{}& \delta_6 \left(\|\tilde{u}_t\|^2_{2,\hat{\Omega}} + \|\tilde{u}\|^2_{3,\hat{\Omega}} + \left\| \int_0^t \tilde{u} d\tau \right\|^2_{2,\hat{S}} \right) \\ &+ a_{15} \left(\|\hat{u}\|^2_{3,\hat{\Omega}} \left\| \int_0^t \hat{u} d\tau \right\|^2_{4,\hat{\Omega}} + \|\hat{u}\|^2_{3,\hat{S}} \left\| \int_0^t \hat{u} d\tau \right\|^2_{3,\hat{\Omega}} \right) + a_{16}\|\hat{u}\|^2_{0,\hat{\Omega}}. \end{aligned}$$

Finally, the last term is bounded by

$$\begin{aligned} &\delta_7 \left(\|\hat{u}_t\|^2_{2,\hat{\Omega}} + \|\hat{u}\|^2_{3,\hat{\Omega}} \right) + a_{17}(\|\hat{u}_t\|^2_{0,\hat{\Omega}} + \|\hat{u}\|^2_{0,\hat{\Omega}}) \\ &\quad + a_{18}(\|\hat{u}\|^4_{2,\hat{\Omega}} + \|\hat{u}\|^2_{3,\hat{\Omega}} \left\| \int_0^t \hat{u} d\tau \right\|^2_{3,\hat{\Omega}}). \end{aligned}$$

By summarizing, we have

$$\begin{aligned} &\int_{\hat{S}} (\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,t\tau} \tilde{u}_{t\tau} J dz' \\ &\le -\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}} g^{\alpha\beta} \bar{n} \cdot \tilde{u}_{ss^\alpha} \bar{n} \cdot \tilde{u}_{ss^\beta} J dz' \\ &\quad + \delta_8 \left(\|\hat{u}_t\|^2_{2,\hat{\Omega}} + \|\hat{u}\|^2_{3,\hat{\Omega}} + \left\| \int_0^t \tilde{u} d\tau \right\|^2_{2,\hat{S}} + \left\| H(\cdot,0) + \frac{2}{R_0} \right\|^2_{1,S^1} \right) \end{aligned} \tag{4.113}$$

$$+a_{19}\left(\|\hat{u}\|^4_{3,\hat{\Omega}}+\|\hat{u}\|^2_{3,\hat{\Omega}}\|\tilde{u}_t\|^2_{2,\hat{\Omega}}+\|\hat{u}\|^2_{3,\hat{\Omega}}\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,\hat{\Omega}}+\|\hat{u}\|^2_{3,\hat{S}}\left\|\int_0^t\hat{u}d\tau\right\|^2_{3,\hat{\Omega}}\right)$$
$$+a_{20}(\|\hat{u}\|^2_{0,\hat{\Omega}}+\|\hat{u}_t\|^2_{0,\hat{\Omega}}).$$

By exploiting (4.113) in (4.111), it follows that

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}^2_{i\tau}+\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma t\tau}\right)Jdz+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}g^{\alpha\beta}\bar{n}\cdot\tilde{u}_{ss^\alpha}\bar{n}\cdot\tilde{u}_{ss^\beta}Jdz'\\&\quad+\frac{\mu}{4}\|\tilde{u}_{t\tau}\|^2_{1,\hat{\Omega}}\\&\le\delta_9\left(\|\hat{u}_{tzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma tz}\|^2_{0,\hat{\Omega}}+\|\hat{u}_{zzz}\|^2_{0,\hat{\Omega}}+\left\|\int_0^t\tilde{u}d\tau\right\|^2_{2,\hat{S}}\right)\\&\quad+a_{21}(|\hat{u}|^2_{2,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{1,0,\hat{\Omega}}+|\tilde{g}|^2_{1,0,\hat{\Omega}})+a_{22}X_6(\hat{\Omega})(1+X_6(\hat{\Omega}))Y_6(\hat{\Omega}).\end{aligned}\tag{4.114}$$

By differentiating the third component of (4.42) with respect to t, multiplying the result by $\tilde{q}_{\sigma nt}J$, and integrating over $\hat{\Omega}$ yields

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+v}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma nt}Jdz+\|\tilde{q}_{\sigma nt}\|^2_{0,\hat{\Omega}}\\&\le(\delta_{10}+cd)\|\hat{F}\|^2_{4+1/2,\hat{S}}(\|\tilde{u}_{zzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zt}\|^2_{0,\hat{\Omega}})+c\|\tilde{u}_{t\tau}\|^2_{1,\hat{\Omega}}\\&\quad+a_{23}(|\hat{u}|^2_{2,0,\hat{\Omega}}+\|\hat{q}_{\sigma t}\|^2_{0,\hat{\Omega}}+|\tilde{g}|^2_{1,0,\hat{\Omega}})\\&\quad+a_{24}X_6(\hat{\Omega})(1+X_6(\hat{\Omega}))Y_6(\hat{\Omega}).\end{aligned}\tag{4.115}$$

By differentiating the third component of (4.44) with respect to t, multiplying the result by $\tilde{u}^3_{nnt}J$, and integrating over $\hat{\Omega}$ implies

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}|\tilde{u}^3_{nt}|^2Jdz+\frac{\mu+\nu}{2}\|\tilde{u}^3_{nnt}\|^2_{0,\hat{\Omega}}\\&\le(\delta_{11}+cd)\|\hat{F}\|^2_{4+1/2,\hat{S}}(\|\tilde{u}_{zzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zt}\|^2_{0,\hat{\Omega}})+\delta_{12}\|\tilde{u}_{ztt}\|^2_{0,\hat{\Omega}}\\&\quad+c(\|\tilde{u}_{z\tau t}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma nt}\|^2_{0,\hat{\Omega}})+a_{25}(|\hat{u}|^2_{2,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{1,0,\hat{\Omega}}+|\tilde{g}|^2_{1,0,\hat{\Omega}})\\&\quad+a_{26}X_6(\hat{\Omega})(1+X_6(\hat{\Omega}))Y_6(\hat{\Omega}).\end{aligned}\tag{4.116}$$

By differentiating (4.46) with respect to t and τ, multiplying by $\tilde{u}'_{t\tau}J$, integrating over $\hat{\Omega}$, and using (4.47) we get the following:

$$\begin{aligned}&\|\tilde{u}'_{zt\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma t\tau}\|^2_{0,\hat{\Omega}}\\&\le(\delta_{13}+cd)\|\hat{F}\|^2_{4+1/2,\hat{S}}(\|\tilde{u}_{zzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zt}\|^2_{0,\hat{\Omega}})\\&\quad+c\|(\operatorname{div}\tilde{u}')_{,\tau t}\|^2_{0,\hat{\Omega}}+a_{27}(|\hat{u}|^2_{2,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{1,0,\hat{\Omega}}+|\tilde{g}|^2_{1,0,\hat{\Omega}})\\&\quad+a_{28}X_6(\hat{\Omega})(1+X_6(\hat{\Omega}))Y_6(\hat{\Omega}).\end{aligned}\tag{4.117}$$

Moreover, from (4.46) we obtain

(4.118)
$$\begin{aligned}\|\tilde{u}'_{nnt}\|^2_{0,\hat{\Omega}} \le{}& (\delta_{14}+cd)\|\hat{F}\|^2_{4+1/2,\hat{S}}(\|\tilde{u}_{zzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zt}\|^2_{0,\hat{\Omega}})\\ &+c(\|\tilde{u}_{zt\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma t\tau}\|^2_{0,\hat{\Omega}})+a_{29}(|\hat{u}|^2_{2,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{1,0,\hat{\Omega}}+|\tilde{g}|^2_{1,0,\hat{\Omega}})\\ &+a_{30}X_6(\hat{\Omega})(1+X_6(\hat{\Omega}))Y_6(\hat{\Omega}).\end{aligned}$$

Finally, we have

$$\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}\tilde{u}^2_{zt}Jdz \le \delta_{15}\|\tilde{u}_{ztt}\|^2_{0,\hat{\Omega}}+c(\|\tilde{u}_t\|^2_{1,\hat{\Omega}}+\|\tilde{u}_{zt}\|^2_{0,\hat{\Omega}}\|\hat{u}\|^2_{3,\hat{\Omega}}). \tag{4.119}$$

From (4.113), (4.115)–(4.119) we obtain, for sufficiently small δ's,

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}^2_{zt}+\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma zt}\right)Jdz+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}g^{\alpha\beta}\bar{n}\cdot\tilde{u}_{ss^\alpha}\bar{n}\cdot\tilde{u}_{ss^\beta}Jdz'\\ &\quad+\|\tilde{u}_{zzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zt}\|^2_{0,\hat{\Omega}}\\ &\le \delta_{16}\|\tilde{u}_{ztt}\|^2_{0,\hat{\Omega}}+(\delta_{17}+cd)(\|\hat{u}_{zzt}\|^2_{0,\hat{\Omega}}+\|\hat{q}_{\sigma zt}\|^2_{0,\hat{\Omega}})\\ &\quad+a_{18}\left(\|\hat{u}_{zzz}\|^2_{0,\hat{\Omega}}+\left\|\int_0^t\tilde{u}d\tau\right\|^2_{2,\hat{S}}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{1,\hat{S}}\right)\\ &\quad+a_{31}(|\hat{u}|^2_{2,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{1,0,\hat{\Omega}}+|\tilde{g}|^2_{1,0,\hat{\Omega}})\\ &\quad+a_{32}X_6(\hat{\Omega})(1+X_6(\hat{\Omega}))Y_6(\hat{\Omega}).\end{aligned} \tag{4.120}$$

By going back to the variables ξ in (4.120), summing over all neighborhoods of the partition of unity (where we use (4.111) for the interior subdomains), then going back to the variables x and using estimate (4.104), we obtain (4.105) for sufficiently small δ's and d. This concludes the proof.

To estimate the first term in the right-hand side of (4.105) we need the following result.

LEMMA 4.7. *For a sufficiently smooth solution of the problem* (4.1), *the following inequality holds*:

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v^2_{tt}+\frac{1}{p\Psi(\rho)}p^2_{\sigma tt}\right)dx+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\bar{n}\cdot\tilde{v}_{s^\alpha t}\bar{n}\cdot\tilde{v}_{s^\beta t}ds\\ &\quad+\|v_{tt}\|^2_{1,\Omega_t}+\|p_{\sigma tt}\|^2_{0,\Omega_t}\\ &\le c\|v_t\|^2_{1,\Omega_t}+\varepsilon_7\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{0,S^1}\\ &\quad+P_{14}(\|f_{tt}\|^2_{0,\Omega_t}+|f|^2_{1,0,\Omega_t})+P_{15}X_7(1+X_7)Y_7,\end{aligned} \tag{4.121}$$

where

$$X_7=|v|^2_{3,1,\Omega_t}+|p_\sigma|^2_{2,0,\Omega_t},\qquad Y_7=|v|^2_{4,2,\Omega_t}+|p_\sigma|^2_{3,1,\Omega_t}+\int_0^t\|u\|^2_{2,\Omega_t}d\tau. \tag{4.122}$$

Proof. Differentiating (4.1a) twice with respect to t, multiplying by v_{tt}, and integrating over Ω_t yields

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v_{tt}^2+\frac{1}{p\Psi(\rho)}p_{\sigma tt}^2\right)dx+\frac{\mu}{2}\|v_{tt}\|_{1,\Omega_t}^2\\ &\quad-\int_{S_t}(n_iT^{ij}(v,p_\sigma))_{,tt}\cdot v_{tt}^i ds\le\delta_1(\|v_{tt}\|_{1,\Omega_t}^2+\|p_{\sigma tt}\|_{0,\Omega_t}^2)\\ &\quad+a_1(|f|_{1,0,\Omega_t}^2+\|f_{tt}\|_{0,\Omega_t}^2)+a_2X_7(1+X_7)Y_7,\end{aligned}\tag{4.123}$$

where we have used

$$\int_{\Omega_t}p_{\sigma tt}\operatorname{div}v_{tt}\,dx=\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\frac{1}{p\Psi(\rho)}p_{\sigma tt}^2dx+N_1,$$

Lemma 5.4 from [35], and

$$|N_1|\le\delta_2\|p_{\sigma tt}\|_{0,\Omega_t}^2+a_3X_7(1+X_7)Y_7.$$

Employing boundary condition (4.1c) we obtain

$$\begin{aligned}\int_{S_t}(\bar n\cdot T(v,p_\sigma))_{,tt}\cdot v_{tt}ds&=\sigma\int_{S_t}\left(g^{\alpha\beta}x_{s^\alpha s^\beta}\cdot\bar n\bar n+\frac{2}{R_0}\bar n\right)_{,tt}\cdot v_{tt}ds\\ &=-\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\bar n\cdot v_{s^\alpha t}\bar n\cdot v_{s^\beta t}ds+N_2,\end{aligned}\tag{4.124}$$

where

$$\begin{aligned}|N_2|\le\delta_3&\left(\|v_{xtt}\|_{0,\Omega_t}^2+\|v_{xxt}\|_{0,\Omega_t}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{0,S^1}^2\right)+a_4\|v_t\|_{0,\Omega_t}^2\\ &+a_5(\|v\|_{2,\Omega_t}^2+\|v_t\|_{2,\Omega_t}^2)\left(\|v\|_{3,\Omega_t}^2+\left\|\int_0^t v\,d\tau\right\|_{2,\Omega_t}^2\right).\end{aligned}$$

Moreover, by the continuity equation (4.2) we have

$$\|p_{\sigma tt}\|_{0,\Omega_t}^2\le c\|v_t\|_{1,\Omega_t}^2+a_6X_7(1+X_7)Y_7.\tag{4.125}$$

Hence, (4.123) and (4.125) imply (4.121). This concludes the proof.

Summarizing, from Lemmas 4.5–4.7 we obtain the following.

LEMMA 4.8.

(4.126)

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho|D_{x,t}^2u|^2+\frac{1}{p\Psi(\rho)}|D_{x,t}^2p_\sigma|^2\right)dx\\ &\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\frac{1}{2}\tilde\delta^{\alpha\beta}\bar n\cdot\int_0^t v_{s_1s_2s^\alpha}d\tau\bar n\cdot\int_0^t v_{s_1s_2s^\beta}d\tau ds+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\left|\bar n\cdot\int_0^t v_{ss^1s^2}d\tau\right|^2ds\\ &\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\sum_{i=1}^2\left(\frac{1}{2}\bar n\cdot\int_0^t v_{ss^1s^2}d\tau+2\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,s}\right)^2ds\\ &\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\bar n\cdot v_{ss^\alpha}\bar n\cdot v_{ss^\beta}ds+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\bar n\cdot v_{s^\alpha t}\bar n\cdot v_{s^\beta t}ds\end{aligned}$$

$$
\begin{aligned}
&+|v|^2_{3,1,\Omega_t}+|p_\sigma|^2_{2,0,\Omega_t}\\
&\le P_{16}\left(|v|^2_{2,0,\Omega_t}+|p_\sigma|^2_{1,0,\Omega_t}+\left\|\int_0^t v\,d\tau\right\|^2_{0,\Omega_t}\right)\\
&\quad+\varepsilon_8\left(\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{1,S^1}+\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\right)\\
&\quad+P_{17}\left(|f|^2_{1,0,\Omega_t}+\|f_{tt}\|^2_{0,\Omega_t}\right)+P_{18}\left(X_8(1+X_8)Y_8+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^4_{1,S^1}\right),
\end{aligned}
$$

where the summation over repeated indices $(\alpha,\beta=1,2)$ *and coordinates* $(x,s_i=(s^1,s^2),i=1,2,s=(s^1,s^2))$ *is assumed and*

$$
\begin{aligned}
X_8&=|v|^2_{3,1,\Omega_t}+|p_\sigma|^2_{2,0,\Omega_t}+\int_0^t\|v\|^2_{3,\Omega_\tau}d\tau,\\
Y_8&=|v|^2_{4,2,\Omega_t}+|p_\sigma|^2_{3,1,\Omega_t}+\int_0^t\|v\|^2_{4,\Omega_t}d\tau.
\end{aligned}
\tag{4.127}
$$

Finally, we obtain inequalities for the fourth derivatives.

LEMMA 4.9. *For a sufficiently smooth solution of the problem* (4.1a–c) *the following estimate holds:*

(4.128)

$$
\begin{aligned}
&\frac12\frac{d}{dt}\int_{\Omega_t^*}\left(\rho v^2_{xxx}+\frac{1}{p\Psi(\rho)}p^2_{\sigma xxx}\right)dx\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\frac12\tilde\delta^{\alpha\beta}\bar n\cdot\int_0^t v_{s_1s_2s_3s^\alpha}d\tau\,\bar n\cdot\int_0^t v_{s_1s_2s_3s^\beta}d\tau\,ds\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\left|\bar n\cdot\int_0^t v_{s_1s_2s^1s^2}d\tau\right|^2ds\\
&\qquad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\sum_{i=1}^2\left(\frac12\bar n\cdot\int_0^t v_{s_1s_2s^is^i}d\tau+2\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,s_1s_2}\right)^2ds\\
&\qquad+\|v\|^2_{4,\Omega_t}+\|p_\sigma\|^2_{3,\Omega_t}\\
&\le\varepsilon_8\left(\|v_{xxxt}\|^2_{0,\Omega_t}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}+\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1}\right)\\
&\qquad+P_{19}\left(|v|^2_{3,2,\Omega_t}+\|p_\sigma\|^2_{2,\Omega_t}+\left\|\int_0^t v\,d\tau\right\|^2_{0,\Omega_t}\right.\\
&\qquad\qquad\left.+\|R(\cdot,t)-R(\cdot,0)\|^2_{0,S^1}+\|f\|^2_{2,\Omega_t}\right)\\
&\qquad+P_{20}\left(X_9(1+X_9^2)Y_9+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^4_{2,S^1}\right.\\
&\qquad\qquad\left.+\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\left\|\int_0^t v\,d\tau\right\|^2_{4,S_t}\right.
\end{aligned}
$$

$$+\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\left\|\int_0^t v d\tau\right\|^2_{3,S_t}\Bigg),$$

where the summation over repeated indices $(\alpha,\beta=1,2)$ *and coordinates* $(x,s_i=(s^1,s^2),i=1,2,3)$ *is assumed and*

$$\begin{aligned} X_9&=|v|^2_{3,2,\Omega_t}+|p_\sigma|^2_{3,2,\Omega_t}+\int_0^t\|v\|^2_{3,\Omega_t}d\tau,\\ Y_9&=|v|^2_{4,3,\Omega_t}+|p_\sigma|^2_{3,2,\Omega_t}+\int_0^t\|v\|^2_{4,\Omega_t}d\tau. \end{aligned} \tag{4.129}$$

Proof. We use the partition of unity. First, we consider interior subdomains. We differentiate (4.27a) three times with respect to ξ, multiply by $\tilde u_{\xi\xi\xi}A$, and integrate over $\tilde\Omega$ to get

$$\begin{aligned} &\frac12\frac{d}{dt}\int_{\tilde\Omega}\left(\eta\tilde u^2_{\xi\xi\xi}+\frac{1}{q\Psi(\eta)}\tilde q^2_{\sigma\xi\xi\xi}\right)Ad\xi\\ &\qquad+\frac{\mu}{2}\int_{\tilde\Omega}(\nabla_{u^i}\tilde u^j_{\xi\xi\xi}+\nabla_{u^j}\tilde u^i_{\xi\xi\xi})^2Ad\xi+(\nu-\mu)\|\nabla_u\cdot\tilde u_{\xi\xi\xi}\|^2_{0,\tilde\Omega}\\ &\le\delta_1(\|\tilde u_{\xi\xi\xi}\|^2_{1,\tilde\Omega}+\|\tilde q_{\sigma\xi\xi\xi}\|^2_{0,\tilde\Omega})+a_1(\|u\|^2_{3,\tilde\Omega}+\|q_\sigma\|^2_{2,\tilde\Omega}+\|\tilde g\|^2_{2,\tilde\Omega})\\ &\qquad+a_2X_9(\tilde\Omega)(1+X_9^2(\tilde\Omega))Y_9(\tilde\Omega), \end{aligned} \tag{4.130}$$

where

$$\begin{aligned} X_9(\tilde\Omega)&=|u|^2_{3,2,\tilde\Omega}+|q_\sigma|^2_{3,2,\tilde\Omega}+\int_0^t\|u\|^2_{3,\tilde\Omega}dt',\\ Y_9(\tilde\Omega)&=|u|^2_{4,3,\tilde\Omega}+|q_\sigma|^2_{3,2,\tilde\Omega}+\int_0^t\|u\|^2_{4,\tilde\Omega}dt'. \end{aligned}$$

We have also used

$$-\int_{\tilde\Omega}\tilde q_{\sigma\xi\xi\xi}\nabla_u\cdot\tilde u_{\xi\xi\xi}Ad\xi=\frac12\frac{d}{dt}\int_{\tilde\Omega}\frac{1}{q\Psi(\eta)}\tilde q^2_{\sigma\xi\xi\xi}Ad\xi+N_1,$$

where

$$|N_1|\le\delta_2\|\tilde q_{\sigma\xi\xi\xi}\|^2_{0,\tilde\Omega}+a_3\|u\|^2_{3,\tilde\Omega}+a_4X_9(\tilde\Omega)(1+X_9^2(\tilde\Omega))Y_9(\tilde\Omega).$$

Moreover, the following relation has been employed, too:

$$\begin{aligned} &\left|\int_{\tilde\Omega}[(\nabla_u\nabla_u\tilde u)_{\xi\xi\xi}-\nabla_u\nabla_u\tilde u_{\xi\xi\xi}]\cdot\tilde u_{\xi\xi\xi}Ad\xi+\int_{\tilde\Omega}[(\nabla_u\tilde q_\sigma)_{\xi\xi\xi}-\nabla_u\tilde q_{\sigma\xi\xi\xi}]\cdot\tilde u_{\xi\xi\xi}Ad\xi\right|\\ &\quad\le\delta_3\|\tilde u_{\xi\xi\xi}\|^2_{1,\tilde\Omega}+a_5X_9(\tilde\Omega)(1+X_9^2(\tilde\Omega))Y_9(\tilde\Omega). \end{aligned}$$

From the problem (4.34) we obtain

$$\begin{aligned} \|\tilde u\|^2_{4,\tilde\Omega}+\|\tilde q_\sigma\|^2_{3,\tilde\Omega}&\le c\|\nabla_u\cdot\tilde u\|^2_{3,\tilde\Omega}+a_6(|u|^2_{3,2,\tilde\Omega}+\|\tilde q_\sigma\|^2_{2,\tilde\Omega}+\|\tilde g\|^2_{2,\tilde\Omega})\\ &\quad+a_7X_9(\tilde\Omega)(1+X_9^2(\tilde\Omega))Y_9(\tilde\Omega). \end{aligned} \tag{4.131}$$

From (4.130) and (4.131) for sufficiently small δ_1 we have

$$\begin{aligned}(4.132)\quad &\frac{1}{2}\frac{d}{dt}\int_{\tilde\Omega}\left(\eta\tilde u_{\xi\xi\xi}^2+\frac{1}{q\Psi(\eta)}\tilde q_{\sigma\xi\xi\xi}^2\right)Ad\xi+\|\tilde u\|_{4,\tilde\Omega}^2+\|\tilde q_\sigma\|_{3,\tilde\Omega}^2\\ &\le a_8(|u|_{3,2,\tilde\Omega}^2+\|q_\sigma\|_{2,\tilde\Omega}^2+\|\tilde g\|_{2,\tilde\Omega}^2)+a_9X_9(\tilde\Omega)(1+X_9(\tilde\Omega))Y_9(\tilde\Omega),\end{aligned}$$

where Lemma 5.1 from [35] in the case $G=\tilde\Omega$ and $v=\tilde u_{\xi\xi\xi}$ has been used.

Now we consider a neighborhood of the boundary. Differentiating (4.28a) three times with respect to τ, multiplying by $\tilde u_{\tau\tau\tau}J$, and integrating over $\hat\Omega$ yields

$$\begin{aligned}(4.133)\quad &\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left[\hat\eta\tilde u_{\tau\tau\tau}^2+\frac{1}{\hat q\Psi(\hat\eta)}\tilde q_{\sigma\tau\tau\tau}^2\right]Jdz+\frac{\mu}{2}\|\tilde u_{\tau\tau\tau}\|_{1,\hat\Omega}^2\\ &\quad+(\nu-\mu)\|\hat\nabla\cdot\tilde u_{\tau\tau\tau}\|_{0,\hat\Omega}^2-\int_{\hat S}(\hat n\hat{\mathbb{T}}(\tilde u,\tilde q_\sigma))_{,\tau\tau\tau}\cdot\tilde u_{\tau\tau\tau}Jdz'\\ &\le\delta_4(\|\tilde u_{z\tau\tau\tau}\|_{0,\hat\Omega}^2+\|\tilde q_{\sigma z\tau\tau}\|_{0,\hat\Omega}^2)+a_{10}(|\hat u|_{3,2,\hat\Omega}^2+\|\hat q_\sigma\|_{2,\hat\Omega}^2+\|\tilde g\|_{2,\hat\Omega}^2)\\ &\quad+a_{11}X_9(\hat\Omega)(1+X_9^2(\hat\Omega))Y_9(\hat\Omega),\end{aligned}$$

where $X_9(\hat\Omega),Y_9(\hat\Omega)$ has the form $X_9(\tilde\Omega),Y_9(\tilde\Omega)$ with $\hat u,\hat q_\sigma,\hat\Omega$ instead of $u,q_\sigma,\tilde\Omega$, respectively. Moreover, we have used

$$-\int_{\hat\Omega}\tilde q_{\sigma\tau\tau\tau}\hat\nabla\cdot\tilde u_{\tau\tau\tau}Jdz=\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\frac{1}{\hat q\Psi(\hat\eta)}\tilde q_{\sigma\tau\tau\tau}^2Jdz+N_2,$$

where

$$|N_2|\le\delta_5\|\tilde q_{\sigma\tau\tau\tau}\|_{0,\hat\Omega}^2+a_{12}\|\hat u\|_{3,\hat\Omega}^2+a_{13}X_9(\hat\Omega)(1+X_9^2(\hat\Omega))Y_9(\hat\Omega).$$

We have also employed the following considerations:

$$\begin{aligned}&\left|\int_{\hat\Omega}[(\hat\nabla\hat\nabla\tilde u)_{,\tau\tau\tau}-\hat\nabla\hat\nabla\tilde u_{\tau\tau\tau}]\cdot\tilde u_{\tau\tau\tau}Jdz+\int_{\hat\Omega}[(\hat\nabla\tilde q_\sigma)_{,\tau\tau\tau}-\hat\nabla\tilde q_{\sigma\tau\tau\tau}]\cdot\tilde u_{\tau\tau\tau}Jdz\right|\\ &\le\delta_6(\|\tilde u_{zz\tau\tau}\|_{0,\hat\Omega}^2+\|\tilde q_{\sigma z\tau\tau}\|_{0,\hat\Omega}^2)+a_{12}(\|\hat u\|_{3,\hat\Omega}^2+\|\hat q_\sigma\|_{2,\hat\Omega}^2)\\ &\quad+a_{13}X_9(\hat\Omega)(1+X_9^2(\hat\Omega))Y_9(\hat\Omega),\end{aligned}$$

and

$$\begin{aligned}&\left|\int_{\hat S}[(\hat n\hat{\mathbb{T}}(\tilde u,\tilde q_\sigma))_{,\tau\tau\tau}-\hat n\hat{\mathbb{T}}(\tilde u_{\tau\tau\tau},\tilde q_{\sigma\tau\tau\tau})]\cdot\tilde u_{\tau\tau\tau}Jdz'\right|\\ &\quad\le\delta_7(\|\tilde u_{z\tau\tau\tau}\|_{0,\hat\Omega}^2+\|\tilde q_{\sigma z\tau\tau}\|_{0,\hat\Omega}^2)+a_{14}(\|\hat u\|_{3,\hat\Omega}^2+\|\hat q_\sigma\|_{2,\hat\Omega}^2)\\ &\qquad+a_{15}X_9(\hat\Omega)(1+X_9^2(\hat\Omega))Y_9(\hat\Omega),\end{aligned}$$

where δ_6 and δ_7 have been assumed sufficiently small.

Finally, Lemma 5.1 from [35] in the case $G=\hat\Omega$ and $v=\tilde u_{\tau\tau\tau}$ has been used.

Using the Lagrangian coordinates and boundary condition (4.28c) we rewrite the boundary term in (4.133) as follows:

$$
\begin{aligned}
&-\int_{\hat{S}} (\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,\tau\tau\tau}\cdot\tilde{u}_{\tau\tau\tau}Jdz' \\
&= -\sigma\int_{\hat{S}}\left(\hat{\Delta}_{\hat{S}_t}\hat{\xi}\cdot\hat{n}\hat{n}\hat{\zeta}+\frac{2}{R_0}\hat{n}\hat{\zeta}\right)_{,\tau\tau\tau}\tilde{u}_{\tau\tau\tau}Jdz' \\
&\quad -\sigma\int_{\hat{S}}\left(\hat{\Delta}_{\hat{S}_t}\int_0^t\tilde{u}d\tau\cdot\hat{n}\hat{n}\right)_{,\tau\tau\tau}\cdot\tilde{u}_{\tau\tau\tau}Jdz'+\int_{\hat{S}}(k_5+k_6)_{,\tau\tau\tau}\tilde{u}_{\tau\tau\tau}Jdz'.
\end{aligned}
\tag{4.134}
$$

Similarly, as in the cases of (4.38) and (4.85), the first term in the right-hand side of (4.134) is estimated by

$$
\begin{aligned}
&\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,\tau\tau}\hat{\zeta}\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau_i\tau_i}ds+\delta_8\left(\|\tilde{u}_{\tau\tau\tau}\|^2_{1,\hat{S}}+\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,\hat{S}}\right.\\
&\quad\left.+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}+\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1}\right)\\
&\quad+a_{16}\left(\|\hat{u}\|^2_{3,\hat{\Omega}}+\left\|\int_0^t\hat{u}d\tau\right\|^2_{0,\hat{\Omega}}+\|R(\cdot,t)-R(\cdot,0)\|^2_{0,\hat{S}}\right)\\
&\quad+a_{17}\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,\hat{S}}\\
&\quad+a_{18}\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\left\|\int_0^t\hat{u}d\tau\right\|^2_{3,\hat{S}}.
\end{aligned}
$$

We estimate the second term by

$$
\begin{aligned}
&\delta_9\left(\left\|\int_0^t\tilde{u}d\tau\right\|^2_{4,\hat{S}}+\|\tilde{u}_{\tau\tau\tau}\|^2_{1,\hat{\Omega}}\right)+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}g^{\alpha\beta}\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau s^\alpha}dt'\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau s^\beta}dt'ds\\
&\quad+a_{19}\|\tilde{u}\|^2_{3,\hat{\Omega}}+a_{20}\|\hat{u}\|^2_{3,\hat{\Omega}}\left\|\int_0^t\tilde{u}d\tau\right\|^2_{4,\hat{S}}.
\end{aligned}
$$

Finally, the last term is bounded by

$$
c\|\hat{u}\|^2_{3,\hat{S}}+\delta_{10}\left(\|\hat{u}_{\tau\tau\tau}\|^2_{1,\hat{\Omega}}+\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,\hat{S}}\right).
$$

Summarizing, we have

$$
\begin{aligned}
&\int_{\hat{S}}(\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,\tau\tau\tau}\cdot\tilde{u}_{\tau\tau\tau}Jdz'\le-\frac{\sigma d}{2dt}\int_{\hat{S}}g^{\alpha\beta}\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau s^\alpha}dt'\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau s^\beta}dt'ds\\
&\quad-\frac{\sigma d}{2dt}\int_{\hat{S}}\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,\tau\tau}\hat{\zeta}\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau_i\tau_i}dt'ds\\
&\quad+\delta_{11}\left(\|\tilde{u}_{\tau\tau\tau}\|^2_{1,\hat{\Omega}}+\left\|\int_0^t\tilde{u}d\tau\right\|^2_{4,\hat{S}}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,\hat{S}}+\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1}\right)
\end{aligned}
\tag{4.135}
$$

$$+a_{21}\left(\|\hat{u}\|^2_{3,\hat{\Omega}}+\left\|\int_0^t\hat{u}d\tau\right\|^2_{0,\hat{\Omega}}+\|R(\cdot,t)-R(\cdot,0)\|^2_{0,S^1}\right)$$

$$+a_{22}\left(\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,S^1}\right.$$

$$\left.+\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\left\|\int_0^t\hat{u}d\tau\right\|^2_{3,\hat{S}}\right)$$

$$+a_{23}\|\hat{u}\|^2_{3,\hat{\Omega}}\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,\hat{S}}.$$

From (4.133) and (4.135) we obtain

(4.136)

$$\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}^2_{\tau\tau\tau}+\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma\tau\tau\tau}\right)Jdz$$

$$+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}g^{\alpha\beta}\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau s^\alpha}dt'\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau s^\beta}dt'ds$$

$$+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,\tau\tau}\hat{\zeta}\hat{n}\cdot\int_0^t\tilde{u}_{\tau\tau\tau_i\tau_i}dt'ds$$

$$+\frac{\mu}{4}\|\tilde{u}_{\tau\tau\tau}\|^2_{1,\hat{\Omega}}+(\nu-\mu)\|\hat{\nabla}\cdot\tilde{u}_{\tau\tau\tau}\|^2_{0,\hat{\Omega}}$$

$$\le\delta_{12}\left(\|\hat{u}_{z\tau\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma z\tau\tau}\|^2_{0,\hat{\Omega}}+\left\|\int_0^t\tilde{u}d\tau\right\|^2_{4,\hat{S}}\right.$$

$$\left.+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,\hat{S}}+\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1}\right)$$

$$+a_{24}\left(|\hat{u}|^2_{3,2,\hat{\Omega}}+\left\|\int_0^t\tilde{u}d\tau\right\|^2_{0,\hat{\Omega}}+\|\hat{q}_\sigma\|^2_{2,\hat{\Omega}}+\|R(\cdot,t)-R(\cdot,0)\|^2_{0,S^1}+\|\tilde{g}\|^2_{2,\hat{\Omega}}\right)$$

$$+a_{25}\left(\|\hat{u}\|^2_{3,2,\hat{\Omega}}\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,\hat{S}}+\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\left\|\int_0^t\hat{u}d\tau\right\|^2_{4,\hat{S}}\right.$$

$$\left.+\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\left\|\int_0^t\hat{u}d\tau\right\|^2_{3,\hat{S}}\right)$$

$$+a_{26}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).$$

Differentiating the third component of (4.42) twice with respect to τ, multiplying the result by $\tilde{q}_{\sigma n\tau\tau}J$, and integrating over $\hat{\Omega}$ gives

(4.137)

$$\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma n\tau\tau}Jdz+\frac{1}{2}\|\tilde{q}_{\sigma n\tau\tau}\|^2_{0,\hat{\Omega}}$$

$$\le(\delta_{13}+cd)(\|\hat{u}_{zz\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma n\tau\tau}\|^2_{0,\hat{\Omega}})$$

$$+c\|\tilde{u}_{\tau\tau\tau}\|^2_{1,\hat{\Omega}}+a_{27}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|q_\sigma\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})$$

$$+a_{28}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).$$

Differentiating the third component of (4.44) twice with respect to τ, multiplying the result by $\tilde{u}^3_{nn\tau\tau}J$, and integrating over $\hat{\Omega}$ implies

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}|\tilde{u}^3_{n\tau\tau}|^2 Jdz+\frac{\mu+\nu}{2}\|\tilde{u}^3_{nn\tau\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le(\delta_{14}+cd)(\|\tilde{u}_{zz\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma z\tau\tau}\|^2_{0,\hat{\Omega}})+\delta_{15}\|\tilde{u}_{n\tau\tau t}\|^2_{0,\hat{\Omega}}\\
&\qquad+c(\|\tilde{u}_{z\tau\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma n\tau\tau}\|^2_{0,\hat{\Omega}})+a_{29}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\tilde{q}_{\sigma}\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\qquad+a_{30}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}\tag{4.138}$$

Differentiating (4.46) three times with respect to τ, multiplying by $\tilde{u}'_{\tau\tau\tau}J$, integrating over $\hat{\Omega}$, and using the boundary condition (4.47) we obtain

$$\begin{aligned}
&\|\tilde{u}'_{z\tau\tau\tau}\|^2_{0,\hat{\Omega}}+\tilde{q}_{\sigma\tau\tau\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le(\delta_{16}+cd)(\|\hat{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\hat{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})\\
&\qquad+c\|\mathrm{div}\tilde{u}_{\tau\tau}\|^2_{1,\hat{\Omega}}+a_{31}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_{\sigma}\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\qquad+a_{32}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}\tag{4.139}$$

Moreover, from (4.46) we find

$$\begin{aligned}
\|\tilde{u}'_{nn\tau\tau}\|^2_{0,\hat{\Omega}}&\le(\delta_{17}+cd)\,(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})\\
&\quad+c(\|\tilde{u}'_{\tau\tau\tau\tau}\|^2_{0,\hat{\Omega}}+\|(\mathrm{div}\,\tilde{u})_{\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma\tau\tau\tau}\|^2_{0,\hat{\Omega}})\\
&\quad+a_{33}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_{\sigma}\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})+a_{34}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}\tag{4.140}$$

To summarize, from (4.137)–(4.140) we obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}|\tilde{u}^3_{n\tau\tau}|^2+\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma n\tau\tau}\right)Jdz+\|\tilde{q}_{\sigma z\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{u}_{zz\tau\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le(\delta_{18}+cd)(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})+\delta_{19}\|\tilde{u}_{n\tau\tau t}\|^2_{0,\hat{\Omega}}\\
&\qquad+c\|\tilde{u}_{z\tau\tau\tau}\|^2_{0,\hat{\Omega}}+a_{35}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_{\sigma}\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\qquad+a_{36}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}\tag{4.141}$$

Differentiating the third component of (4.42) with respect to n and τ, multiplying by $\tilde{q}_{\sigma nn\tau}J$, and integrating over $\hat{\Omega}$ yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma nn\tau}Jdz+\|\tilde{q}_{\sigma nn\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le(\delta_{20}+cd)(\|\tilde{u}_{zzz\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zz\tau}\|^2_{0,\hat{\Omega}})\\
&\qquad+c\|\tilde{u}_{zz\tau\tau}\|^2_{0,\hat{\Omega}}+a_{37}(|\hat{u}|^2_{3,2,\hat{\Omega}}+|\hat{q}_{\sigma}|^2_{2,1,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\qquad+a_{38}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}\tag{4.142}$$

Differentiating the third component of (4.94) with respect to n and τ gives

$$
\begin{aligned}
\|(\operatorname{div}\tilde{u})_{nn\tau}\|^2_{0,\hat{\Omega}} &\le (\delta_{21}+cd)(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})\\
&\quad + c(\|\tilde{u}_{zz\tau\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma nn\tau}\|^2_{0,\hat{\Omega}})+a_{39}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_\sigma\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\quad + a_{40}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}
\tag{4.143}
$$

Next we differentiate (4.46) with respect to n and τ. Hence we get

$$
\begin{aligned}
\|\tilde{u}_{nn\tau\tau}\|^2_{0,\hat{\Omega}} &\le (\delta_{22}+cd)(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})\\
&\quad + c(\|\tilde{u}_{zz\tau\tau}\|^2_{0,\hat{\Omega}}+\|(\operatorname{div}\tilde{u})_{zn\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zn\tau}\|^2_{0,\hat{\Omega}})\\
&\quad + a_{41}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_\sigma\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})+a_{42}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}
\tag{4.144}
$$

From (4.141)–(4.144) we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}|\tilde{u}^3_{n\tau\tau}|^2+\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma zz\tau}\right)Jdz+\|\tilde{u}_{zzz\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zz\tau}\|^2_{0,\hat{\Omega}}\\
&\quad\le(\delta_{23}+cd)(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})+\delta_{24}\|\tilde{u}_{n\tau\tau t}\|^2_{0,\hat{\Omega}}\\
&\quad+c\|\tilde{u}_{z\tau\tau\tau}\|^2_{0,\hat{\Omega}}+a_{43}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_\sigma\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\quad+a_{44}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}
\tag{4.145}
$$

Differentiating the third component of (4.42) twice with respect to n, multiplying the result by $\tilde{q}_{\sigma nnn}J$, and integrating over $\hat{\Omega}$ yields

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma nnn}Jdz+\|\tilde{q}_{\sigma nnn}\|^2_{0,\hat{\Omega}})\\
&\quad\le(\delta_{25}+cd)(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})\\
&\quad+c\|\tilde{u}_{zzz\tau}\|^2_{0,\hat{\Omega}}+a_{45}(|\hat{u}|^2_{3,2,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\quad+a_{46}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}
\tag{4.146}
$$

Differentiating the third component of (4.94) twice with respect to n implies

$$
\begin{aligned}
\|(\operatorname{div}\tilde{u})_{nnn}\|^2_{0,\hat{\Omega}} &\le (\delta_{26}+cd)\,(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})\\
&\quad + c(\|\tilde{u}_{zzz\tau}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma nnn}\|^2_{0,\hat{\Omega}})+a_{47}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_\sigma\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})\\
&\quad + a_{48}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}
\tag{4.147}
$$

We differentiate (4.46) twice with respect to n. Hence after integrating over $\hat{\Omega}$ we obtain

$$
\begin{aligned}
\|\tilde{u}_{nnnn}\|^2_{0,\hat{\Omega}} &\le (\delta_{27}+cd)\,(\|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzz}\|^2_{0,\hat{\Omega}})\\
&\quad + c(\|\tilde{u}_{zzz\tau}\|^2_{0,\hat{\Omega}}+\|(\operatorname{div}\tilde{u})_{znn}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma znn}\|^2_{0,\hat{\Omega}})\\
&\quad + a_{49}(|\hat{u}|^2_{3,2,\hat{\Omega}}+\|\hat{q}_\sigma\|^2_{2,\hat{\Omega}}+\|\tilde{g}\|^2_{2,\hat{\Omega}})+a_{50}X_9(\hat{\Omega})(1+X_9^2(\hat{\Omega}))Y_9(\hat{\Omega}).
\end{aligned}
\tag{4.148}
$$

Finally, using that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\hat\eta\tilde u_{zzz}^2 J dz &\le \delta_{28}(\|\tilde u_{zzzt}\|_{0,\hat\Omega}^2+\|\tilde u_{zzzz}\|_{0,\hat\Omega}^2)\\ &\quad + c\|\tilde u\|_{3,\hat\Omega}^2 + c(|\hat q_\sigma|_{2,1,\hat\Omega}^2+\|\hat u\|_{3,\hat\Omega}^2)\|\tilde u\|_{3,\hat\Omega}^2,\end{aligned}\tag{4.149}$$

from (4.145)–(4.149), we obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left(\hat\eta\tilde u_{zzz}^2+\frac{\mu+\nu}{\hat q\Psi(\hat\eta)}\tilde q_{\sigma zzz}^2\right)Jdz+\|\tilde u_{zzzz}\|_{0,\hat\Omega}^2+\|\tilde q_{\sigma zzz}\|_{0,\hat\Omega}^2\\ &\quad\le \delta_{29}\|\tilde u_{zzzt}\|_{0,\hat\Omega}^2+(\delta_{30}+cd)(\|\hat u_{zzzz}\|_{0,\hat\Omega}^2+\|\hat q_{\sigma zzz}\|_{0,\hat\Omega}^2)\\ &\quad+c\|\tilde u_{z\tau\tau\tau}\|_{0,\hat\Omega}^2+a_{51}(|\hat u|_{3,2,\hat\Omega}^2+|\hat q_\sigma|_{2,1,\hat\Omega}^2+\|\tilde g\|_{2,\hat\Omega}^2)\\ &\quad+a_{52}X_9(\hat\Omega)(1+X_9^2(\hat\Omega))Y_9(\hat\Omega).\end{aligned}\tag{4.150}$$

From (4.136) and (4.150) it follows that

(4.151)

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left(\hat\eta\tilde u_{zzz}^2+\frac{\mu+\nu}{\hat q\Psi(\hat\eta)}\tilde q_{\sigma zzz}^2\right)Jdz\\ &\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat S}g^{\alpha\beta}\hat n\cdot\int_0^t\tilde u_{\tau\tau\tau s^\alpha}dt'\hat n\cdot\int_0^t\tilde u_{\tau\tau\tau s^\beta}dt'ds\\ &\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat S}\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,\tau\tau}\hat\zeta\hat n\cdot\int_0^t\tilde u_{\tau\tau\tau_i\tau_i}dt'ds\\ &\quad+\|\tilde u_{zzzz}\|_{0,\hat\Omega}^2+\tilde q_{\sigma zzz}\|_{0,\hat\Omega}^2\\ &\le \delta_{31}\left(\|\tilde u_{zzz\tau}\|_{0,\hat\Omega}^2+\left\|\int_0^t\tilde u d\tau\right\|_{4,\hat S}^2+\|H(\cdot,0)+\frac{2}{R_0}\|_{2,S^1}^2+\|R(\cdot,t)-R(\cdot,0)\|_{4,S^1}^2\right)\\ &\quad+(\delta_{32}+cd)(\|\hat u_{zzzz}\|_{0,\hat\Omega}^2+\|\hat q_{\sigma zzz}\|_{0,\hat\Omega}^2)\\ &\quad+a_{53}\left(|\hat u|_{3,2,\hat\Omega}^2+\left\|\int_0^t\hat u d\tau\right\|_{0,\hat\Omega}^2+|\hat q_\sigma|_{2,1,\hat\Omega}^2\right.\\ &\qquad\left.+\|R(\cdot,t)-R(\cdot,0)\|_{0,S^1}^2+\|\tilde g\|_{2,\hat\Omega}^2\right)\\ &\quad+a_{54}\left(\|\hat u\|_{3,\hat\Omega}^2\left\|\int_0^t\hat u d\tau\right\|_{4,\hat S}^2+\|R(\cdot,t)-R(\cdot,0)\|_{3,S^1}^2\left\|\int_0^t\hat u d\tau\right\|_{4,\hat S}^2\right.\\ &\qquad\left.+\|R(\cdot,t)-R(\cdot,0)\|_{4+1/2,S^1}^2\left\|\int_0^t\hat u d\tau\right\|_{3,\hat S}^2\right)\\ &\quad+a_{55}X_9(\hat\Omega)(1+X_9^2(\hat\Omega))Y_9(\hat\Omega).\end{aligned}$$

Now we examine the second and the third terms in the left-hand side of (4.151). Applying the same considerations as in the case of (4.69), (4.70), and (4.101) we find

that both terms are equal to

$$
\begin{aligned}
&\frac{\sigma}{2}\frac{d}{dt}\int_{\hat S}\frac{1}{2}\tilde\delta^{\alpha\beta}\hat n\cdot\int_0^t \tilde u_{s_1s_2s_3s^\alpha}d\tau\hat n\cdot\int_0^t\tilde u_{s_1s_2s_3s^\beta}dtJdz'\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat S}\left|\hat n\cdot\int_0^t\tilde u_{s_1s_2s^is^i}d\tau\right|^2Jdz'\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat S}\sum_{i=1}^2\left(\frac{1}{2}\hat n\cdot\int_0^t\tilde u_{s_1s_2s^is^i}d\tau+2\left(\left(H(\cdot,0)+\frac{2}{R_0}\right)\zeta\right)_{,s_1s_2}\right)^2Jdz'\\
&\quad-4\sigma\frac{d}{dt}\int_{\hat S}\left(H(\cdot,0)+\frac{2}{R_0}\right)^2_{,s_1s_2}Jdz'.
\end{aligned}
\tag{4.152}
$$

Going back to the variable ξ in (4.151) and (4.152), summing the result and (4.132) over all neighborhoods of the partition of unity, using that δ_{32} and d are sufficiently small, and, finally, going back to the variables x we obtain

(4.153)

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v_{xxx}^2+\frac{1}{p\Psi(\rho)}p_{\sigma xxx}^2\right)dx\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\frac{1}{2}\tilde\delta^{\alpha\beta}\bar n\cdot\int_0^t v_{s_1s_2s_3s^\alpha}d\tau\,\bar n\cdot\int_0^t v_{s_1s_2s_3s^\beta}d\tau\,ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\left|\bar n\cdot\int_0^t v_{s_1s_2s^1s^2}d\tau\right|^2ds\\
&\quad+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}\sum_{i=1}^2\left(\frac{1}{2}\bar n\cdot\int_0^t v_{s_1s_2s^is^i}d\tau+2\left(H(\cdot,0)+\frac{2}{R_0}\right)_{,s_1s_2}\right)^2ds\\
&\le\delta_{33}\left(\|v_{xxxt}\|^2_{0,\Omega_t}+\left\|\int_0^t vd\tau\right\|^2_{4,S_t}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}+\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1}\right)\\
&\quad+a_{55}\left(|v|^2_{3,2,\Omega_t}+\|p_\sigma\|^2_{2,\Omega_t}+\left\|\int_0^t vd\tau\right\|^2_{0,\Omega_t}+\|R(\cdot,t)-R(\cdot,0)\|^2_{0,S^1}+\|f\|^2_{2,\Omega_t}\right)\\
&\quad+a_{56}\left(X_9(+X_9^2)Y_9+\|v\|^2_{3,\Omega_t}\left\|\int_0^t vd\tau\right\|^2_{4,S_t}+\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\left\|\int_0^t vd\tau\right\|^2_{4,S_t}\right.\\
&\qquad\left.+\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\left\|\int_0^t vd\tau\right\|^2_{3,S_t}\right)\\
&\quad+a_{57}\left|\frac{d}{dt}\int_{S_t}\left(H(\cdot,0)+\frac{2}{R_0}\right)^2_{,ss}ds\right|.
\end{aligned}
$$

We have the estimates

$$
\begin{aligned}
\left\|\int_0^t vd\tau\right\|^2_{4,S_t}&\le\delta_{34}(\|v_{xxxx}\|^2_{0,\Omega_t}+\|p_{\sigma xxx}\|^2_{0,\Omega_t})\\
&\quad+a_{58}\left(\|R(\cdot,t)-R(\cdot,0)\|^2_{4,s^1}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}+\left\|\int_0^t vd\tau\right\|^2_{0,\Omega_t}\right.
\end{aligned}
$$

(4.154)
$$+\|v\|^2_{0,\Omega_t}+\|p_\sigma\|^2_{0,\Omega_t}\Big)$$
$$+a_{59}(\|v\|^2_{3,S_t}+\|p_\sigma\|^2_{2,S_t})\left\|\int_0^t v\,d\tau\right\|^2_{4,\Omega_t}\left(1+\left\|\int_0^t v\,d\tau\right\|^2_{3,\Omega_t}\right),$$

and

$$\left|\frac{d}{dt}\int_{S_t}\left(H(\cdot,0)+\frac{2}{R_0}\right)^2_{,ss}ds\right|\le\delta_{35}\|v\|^2_{3,\Omega_t}+c\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^4_{2,S}. \tag{4.155}$$

Using (4.154) and (4.155) in (4.153) implies (4.128) for sufficiently small δ_{34} and δ_{35}. This concludes the proof.

To estimate the first term in the right-hand side of (4.128) we need the following result.

LEMMA 4.10. *For a sufficiently smooth solution of the problem* (4.1a–c) *the following inequality holds*:

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v^2_{xxt}+\frac{\mu+\nu}{p\Psi(\rho)}p^2_{\sigma xxt}\right)dx+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}v_{s_1s_2s^\alpha}\cdot\bar n v_{s_1s_2s^\beta}\cdot\bar n ds\\&\quad+\|v_{xxxt}\|^2_{0,\Omega_t}+\|p_{\sigma xxt}\|^2_{0,\Omega_t}\\&\le\varepsilon_{10}\left(\|v_{xxxx}\|^2_{0,\Omega_t}+\|v_{xxtt}\|^2_{0,\Omega_t}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{0,S^1}\right)\\&\quad+P_{21}(|v|^2_{3,1,\Omega_t}+|p_\sigma|^2_{2,1,\Omega_t}+|f|^2_{2,1,\Omega_t})+P_{22}X_{10}(1+X^2_{10})Y_{10},\end{aligned} \tag{4.156}$$

where $\varepsilon_{10}\in(0,1)$, *the summation over repeated indices* $(\alpha,\beta=1,2)$ *and coordinates* $(x,s_i=(s^1,s^2),i=1,2)$ *is assumed, and*

$$\begin{aligned}X_{10}&=|v|^2_{3,2,\Omega_t}+|p_\sigma|^2_{3,1,\Omega_t}+\int_0^t\|v\|^2_{3,\Omega_t}d\tau,\\Y_{10}&=|v|^2_{4,3,\Omega_t}+|p_\sigma|^2_{3,1,\Omega_t}+\int_0^t\|v\|^2_{4,\Omega_t}d\tau.\end{aligned} \tag{4.157}$$

Proof. We also use the partition of unity. First we consider the interior subdomains. Differentiating (4.27a) twice with respect to ξ and once with respect to time, multiplying by $\tilde u_{t\xi\xi}A$, and integrating over $\tilde\Omega$ yields

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\tilde\Omega}\left(\eta\tilde u^2_{i\xi\xi}+\frac{1}{q\Psi(\eta)}\tilde q^2_{\sigma t\xi\xi}\right)Ad\xi\\&\quad+\frac{\mu}{2}\int_{\tilde\Omega}(\nabla_{u^i}\tilde u^j_{t\xi\xi}+\nabla_{u^j}\tilde u^i_{t\xi\xi})^2Ad\xi+(\nu-\mu)\|\nabla_u\cdot\tilde u_{t\xi\xi}\|^2_{0,\tilde\Omega}\\&\le\delta_1(\|\tilde u_{t\xi\xi}\|^2_{1,\tilde\Omega}+\|\tilde q_{\sigma t\xi\xi}\|^2_{0,\tilde\Omega})+a_1(|\tilde u|^2_{3,2,\tilde\Omega}+|\tilde q_\sigma|^2_{2,1,\tilde\Omega}+|\tilde g|^2_{2,1,\tilde\Omega})\\&\quad+a_2X_{10}(\tilde\Omega)(1+X^2_{10}(\tilde\Omega))Y_{10}(\tilde\Omega),\end{aligned} \tag{4.158}$$

where

$$X_{10}(\tilde{\Omega}) = |u|^2_{3,2,\tilde{\Omega}} + |q_\sigma|^2_{3,1,\tilde{\Omega}} + \int_0^t \|u\|^2_{3,\tilde{\Omega}} dt,$$

$$Y_{10}(\tilde{\Omega}) = |u|^2_{4,3,\tilde{\Omega}} + |q_\sigma|^2_{3,1,\tilde{\Omega}} + \int_0^t \|u\|^2_{4,\tilde{\Omega}} dt.$$

Moreover, the following considerations have been used:

$$\begin{aligned} -\int_{\tilde{\Omega}} \tilde{q}_{\sigma t\xi\xi} \nabla_u \cdot \tilde{u}_{t\xi\xi} A d\xi &= \frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \frac{1}{q\Psi(\eta)} \tilde{q}^2_{\sigma t\xi\xi} A d\xi + N_1, \\ &\Big| \int_{\tilde{\Omega}} [(\nabla_u \nabla_u \tilde{u})_{,t\xi\xi} - \nabla_u \nabla_u \tilde{u}_{,t\xi\xi}] \cdot \tilde{u}_{t\xi\xi} A d\xi \\ &\quad + \int_{\tilde{\Omega}} [(\nabla_u \tilde{q}_\sigma)_{,t\xi\xi} - \nabla_u \tilde{q}_{\sigma t\xi\xi}] \cdot \tilde{u}_{t\xi\xi} A d\xi \Big| \\ &\le \delta_2 \|\tilde{u}_{t\xi\xi}\|^2_{1,\tilde{\Omega}} + a_3 X_{10}(\tilde{\Omega})(1 + X^2_{10}(\tilde{\Omega}))Y_{10}(\tilde{\Omega}), \end{aligned}$$

where

$$N_1 \le \delta_3 \|\tilde{q}_{\sigma t\xi\xi}\|^2_{0,\tilde{\Omega}} + a_4 |\tilde{u}|^2_{3,2,\tilde{\Omega}} + a_5 X_{10}(\tilde{\Omega})(1 + X^2_{10}(\tilde{\Omega}))Y_{10}(\tilde{\Omega}).$$

From (4.34) we obtain

(4.159)
$$\begin{aligned} \|\tilde{u}_t\|^2_{3,\tilde{\Omega}} + \|\tilde{q}_{\sigma t}\|^2_{2,\tilde{\Omega}} &\le c\|\nabla_u \cdot \tilde{u}_t\|^2_{2,\tilde{\Omega}} \\ &\quad + a_6(|u|^2_{3,1,\tilde{\Omega}} + |q_\sigma|^2_{2,1,\tilde{\Omega}} + |\tilde{g}|^2_{2,1,\tilde{\Omega}}) \\ &\quad + a_7 X_{10}(\tilde{\Omega})(1 + X^2_{10}(\tilde{\Omega}))Y_{10}(\tilde{\Omega}). \end{aligned}$$

Now, by applying Lemma 5.1 from [35] for $G = \tilde{\Omega}$ and $v = \tilde{u}_{t\xi\xi}$ from (4.158) and (4.159) for sufficiently small δ_1, we have

(4.160)
$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}} \left(\eta \tilde{u}^2_{t\xi\xi} + \frac{1}{q\Psi(\eta)} \tilde{q}^2_{\sigma t\xi\xi}\right) A d\xi + \|\tilde{u}_t\|^2_{3,\tilde{\Omega}} + \|\tilde{q}_{\sigma t}\|^2_{2,\tilde{\Omega}} \\ &\quad \le a_8(|u|^2_{3,1,\tilde{\Omega}} + |q_\sigma|^2_{2,1,\tilde{\Omega}} + |\tilde{g}|^2_{2,1,\tilde{\Omega}}) + a_9 X_{10}(\tilde{\Omega})(1 + X^2_{10}(\tilde{\Omega}))Y_{10}(\tilde{\Omega}). \end{aligned}$$

Now we consider boundary subdomains. Differentiating (4.28a) with respect to t and twice with respect to τ, multiplying by $\tilde{u}_{t\tau\tau} J$, integrating over $\tilde{\Omega}$, and applying Lemma 5.1 from [35] for $G = \hat{\Omega}$ and $v = \tilde{u}_{t\tau\tau}$ gives

(4.161)
$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}} \left[\hat{\eta}\tilde{u}^2_{t\tau\tau} + \frac{1}{\hat{q}\Psi(\hat{\eta})} \tilde{q}^2_{\sigma t\tau\tau}\right] J dz + \|\tilde{u}_{t\tau\tau}\|^2_{1,\hat{\Omega}} \\ &\qquad + (\nu - \mu)\|(\operatorname{div} \tilde{u})_{,t\tau\tau}\|^2_{0,\hat{\Omega}} - \int_{\hat{S}} (\hat{n}\hat{T}(\tilde{u}, \tilde{q}_\sigma))_{,t\tau\tau} \cdot \tilde{u}_{t\tau\tau} J dz' \\ &\quad \le \delta_4(\|\tilde{u}_{zz\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z\tau t}\|^2_{0,\hat{\Omega}}) \\ &\qquad + a_{10}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}} + |\tilde{g}|^2_{2,1,\hat{\Omega}}) \\ &\qquad + a_{11} X_{10}(\hat{\Omega})(1 + X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}), \end{aligned}$$

where we have used the following relations:

$$-\int_{\hat{\Omega}} \tilde{q}_{\sigma\tau\tau t}\hat{\nabla}\cdot\tilde{u}_{\tau\tau t}Jdz = -\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma\tau\tau t}Jdz + N_2,$$

$$\int_{\hat{\Omega}}[(\hat{\nabla}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,\tau\tau t} - \hat{\nabla}\hat{\mathbb{T}}(\tilde{u}_{\tau\tau t},\tilde{q}_{\sigma\tau\tau t})]\tilde{u}_{\tau\tau t}Jdz$$

$$\le \delta_5(\|\tilde{u}_{zz\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z\tau t}\|^2_{0,\hat{\Omega}}) + a_{12}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}})$$

$$+a_{13}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega})$$

and

$$N_2 \le \delta_6\|\tilde{q}_{\sigma\tau\tau t}\|^2_{0,\hat{\Omega}} + a_{14}|\hat{u}|^2_{3,2,\hat{\Omega}} + a_{15}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).$$

Moreover, $X_{10}(\hat{\Omega})$, $Y_{10}(\hat{\Omega})$ are obtained from $X_{10}(\tilde{\Omega})$, $Y_{10}(\tilde{\Omega})$, replacing u, q_σ, $\tilde{\Omega}$, by $\hat{u}$, $\hat{q}_\sigma$, $\hat{\Omega}$, respectively.

In view of the boundary condition (4.28c) the boundary term in (4.161) can be estimated in the following way:

$$-\int_{\hat{S}}(\hat{n}\hat{\mathbb{T}}(\tilde{u},\tilde{q}_\sigma))_{,t\tau\tau}\cdot\tilde{u}_{t\tau\tau}Jdz' = -\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}} g^{\alpha\beta}\tilde{u}_{\tau\tau s^\alpha}\cdot\bar{n}\tilde{u}_{\tau\tau s^\beta}\cdot\bar{n}Jdz' + N_3, \tag{4.162}$$

where

$$|N_3| \le \delta_7\left(\|\hat{u}_{t\tau\tau z}\|^2_{0,\hat{\Omega}} + \|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}} + \left\|\int_0^t \tilde{u}d\tau\right\|^2_{4,\hat{\Omega}} + \left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{0,S^1}\right)$$

$$+ a_{16}|\hat{u}|^2_{3,1,\hat{\Omega}} + a_{17}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).$$

From (4.161) and (4.162) we obtain

(4.163)

$$\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left[\hat{\eta}\tilde{u}^2_{t\tau\tau} + \frac{1}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma t\tau\tau}\right]Jdz + \frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}} g^{\alpha\beta}\tilde{u}_{\tau\tau s^\alpha}\cdot\bar{n}\tilde{u}_{\tau\tau s^\beta}\cdot\bar{n}Jdz'$$

$$+\|\tilde{u}_{t\tau\tau}\|^2_{1,\hat{\Omega}} + (\nu-\mu)\|(\mathrm{div}\tilde{u})_{,t\tau\tau}\|^2_{0,\hat{\Omega}}$$

$$\le \delta_8\left(\|\tilde{u}_{t\tau\tau z}\|^2_{0,\hat{\Omega}} + \|\tilde{u}_{zzzz}\|^2_{0,\hat{\Omega}} + \|\hat{q}_{\sigma t\tau\tau}|^2_{0,\hat{\Omega}} + \left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{0,S^1}\right)$$

$$+a_{18}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}} + |\tilde{g}|^2_{2,1,\hat{\Omega}}) + a_{19}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).$$

Differentiating the third component of (4.42) with respect to τ and t, multiplying the result by $\tilde{q}_{\sigma n\tau t}J$, and integrating over $\hat{\Omega}$ implies

$$\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma n\tau t}Jdz + \|\tilde{q}_{\sigma n\tau t}\|^2_{0,\hat{\Omega}}$$

$$\le (\delta_9 + cd)(\|\tilde{u}_{zz\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z\tau t}\|^2_{0,\hat{\Omega}})$$

$$+c\|\tilde{u}_{z\tau\tau t}\|^2_{0,\hat{\Omega}} + a_{20}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\tilde{q}_\sigma|^2_{2,1,\hat{\Omega}} + |\tilde{g}|^2_{2,1,\hat{\Omega}})$$

$$+a_{21}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}). \tag{4.164}$$

Differentiating the third component of (4.44) with respect to τ and t, multiplying the result by $\tilde{u}^3_{nn\tau t}J$, and integrating over $\hat{\Omega}$ gives

(4.165)
$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}|\tilde{u}^3_{n\tau t}|^2 J dz + \frac{\mu+\nu}{2}\|\tilde{u}^3_{nn\tau t}\|^2_{0,\hat{\Omega}} \\
&\quad \le \delta_{10}\|\tilde{u}_{zztt}\|^2_{0,\hat{\Omega}} \\
&\qquad +(\delta_{10}+cd)(\|\tilde{u}_{zzzt}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z\tau t}\|^2_{0,\hat{\Omega}}) + c(\|\tilde{u}_{z\tau\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma n\tau t}\|^2_{0,\hat{\Omega}}) \\
&\qquad +a_{22}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}} + |\tilde{g}|^2_{2,1,\hat{\Omega}}) + a_{23}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}$$

From the problem (4.46)–(4.48) we have

(4.166)
$$\begin{aligned}
\|\tilde{u}'_{n\tau\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma\tau\tau t}\|^2_{0,\hat{\Omega}} &\le (\delta_{12}+cd)\left(\|\tilde{u}_{zzzt}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}}\right) \\
&\quad +c(\|\tilde{u}^3_{nn\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{u}_{z\tau\tau t}\|^2_{0,\hat{\Omega}}) \\
&\quad +a_{24}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}} + |\tilde{g}|^2_{2,1,\hat{\Omega}}) \\
&\quad +a_{25}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}),
\end{aligned}$$

where prim denotes that only components u^1, u^2 are taken into consideration. Moreover, from (4.46) we get

(4.167)
$$\begin{aligned}
\|\tilde{u}'_{nn\tau t}\|^2_{0,\hat{\Omega}} &\le c\,(\,\|(\mathrm{div}\tilde{u})_{,\tau\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma\tau\tau t}\|^2_{0,\hat{\Omega}}) \\
&\quad +(\delta_{13}+cd)(\|\tilde{u}_{zzzt}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}}) \\
&\quad +a_{26}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}} + |\tilde{g}|^2_{2,1,\hat{\Omega}}) \\
&\quad +a_{27}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}$$

Summarizing, from (4.164)–(4.167) we obtain

(4.168)
$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}|\tilde{u}^3_{n\tau t}|^2 + \frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma n\tau t}\right)Jdz + \|\tilde{u}_{nn\tau t}\|^2_{0,\hat{\Omega}} + \|\tilde{q}_{\sigma z\tau t}\|^2_{0,\hat{\Omega}} \\
&\quad \le \delta_{14}\|\tilde{u}_{tt}\|^2_{2,\hat{\Omega}} + c\|\tilde{u}_{\tau\tau t}\|^2_{1,\hat{\Omega}} + (\delta_{15}+cd)(\|\hat{u}_{zzzt}\|^2_{0,\hat{\Omega}} + \|\hat{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}}) \\
&\qquad +a_{28}(|\hat{u}|^2_{3,1,\hat{\Omega}} + |\hat{q}_\sigma|^2_{2,1,\hat{\Omega}} + |\tilde{g}|^2_{2,1,\hat{\Omega}}) \\
&\qquad +a_{29}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}$$

Differentiating the third component of (4.42) with respect to t and n, multiplying the

result by $\tilde{q}_{\sigma nnt}J$, and integrating over $\hat{\Omega}$ yields

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma nnt}Jdz\\
&\quad+\frac{1}{2}\|\tilde{q}_{\sigma nnt}\|^2_{0,\hat{\Omega}}\le c\|\tilde{u}_{zz\tau t}\|^2_{0,\hat{\Omega}}\\
&\quad+(\delta_{16}+cd)(\|\tilde{u}_{zzzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}})\\
&\quad+a_{30}(|\hat{u}|^2_{3,1,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}+|\tilde{g}|^2_{2,1,\hat{\Omega}})+a_{31}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}
\tag{4.169}
$$

Differentiating the third component of (4.44) with respect to n and t, multiplying the result by $\tilde{u}^3_{nnnt}J$, and integrating over $\hat{\Omega}$ implies

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}|\tilde{u}^3_{nnt}|^2Jdz+\frac{\mu+\nu}{2}\|\tilde{u}^3_{nnnt}\|^2_{0,\hat{\Omega}}\\
&\quad\le c(\|\tilde{u}_{zz\tau t}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma nnt}\|^2_{0,\hat{\Omega}})\\
&\qquad+\delta_{17}\|\tilde{u}_{tt}\|^2_{2,\hat{\Omega}}+(\delta_{18}+cd)(\|\tilde{u}_{zzzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}})\\
&\qquad+a_{32}(|\hat{u}|^2_{3,1,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}+|\tilde{g}|^2_{2,1,\hat{\Omega}})\\
&\qquad+a_{33}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}
\tag{4.170}
$$

Finally, from (4.46) we get

$$
\begin{aligned}
\|\tilde{u}'_{nnnt}\|^2_{0,\hat{\Omega}}&\le c\left((\mathrm{div}\tilde{u})_{,n\tau t}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma n\tau t}\|^2_{0,\hat{\Omega}}\right)\\
&\quad+c(\delta_{19}+cd)(\|\tilde{u}_{zzzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}})\\
&\quad+a_{34}(|\hat{u}|^2_{3,1,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}+|\tilde{g}|^2_{2,1,\hat{\Omega}})+a_{35}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}
\tag{4.171}
$$

Hence, from (4.168)–(4.171) it follows that

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}|\tilde{u}^3_{znt}|^2+\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma znt}\right)Jdz+\|\tilde{u}_{znnt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}}\\
&\quad\le\delta_{20}\|\tilde{u}_{zztt}\|^2_{0,\hat{\Omega}}+c\|\tilde{u}_{z\tau\tau t}\|^2_{0,\hat{\Omega}}+(\delta_{21}+cd)(\|\tilde{u}_{zzzt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma zzt}\|^2_{0,\hat{\Omega}})\\
&\qquad+a_{36}(|\hat{u}|^2_{3,1,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,1,\hat{\Omega}}+|\tilde{g}|^2_{2,1,\hat{\Omega}})+a_{37}X_{10}(\hat{\Omega})(1+X^2_{10}(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}
\tag{4.172}
$$

To obtain a full derivative $\tilde{u}_{zzt}$ under the integral over $\hat{\Omega}$ and under derivative with respect to time, we need the following:

$$
\begin{aligned}
\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}\tilde{u}^2_{zzt}Jdz&\le\delta_{22}\|\tilde{u}_{zztt}\|^2_{0,\hat{\Omega}}+c\|\tilde{u}_{zzt}\|^2_{0,\hat{\Omega}}\\
&\quad+c(\|\hat{u}\|^2_{3,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,1,\hat{\Omega}})|\hat{u}|^2_{3,2,\hat{\Omega}}.
\end{aligned}
\tag{4.173}
$$

From (4.163), (4.172), and (4.173) we obtain for sufficiently small δ and d,

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\left(\hat{\eta}\tilde{u}_{zzt}^2+\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}_{\sigma zzt}^2\right)Jdz+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat{S}}g^{\alpha\beta}\tilde{u}_{\tau\tau^\alpha}\cdot\bar{n}\tilde{u}_{\tau\tau^\beta}\cdot\bar{n}Jdz'\\
&\qquad+\|\tilde{u}_{zzzt}\|_{0,\hat{\Omega}}^2+\|\tilde{q}_{\sigma zzt}\|_{0,\hat{\Omega}}^2\\
&\qquad\le\delta_{23}\Bigg(\|\hat{u}_{zzzz}\|_{0,\hat{\Omega}}^2+\|\hat{u}_{zztt}\|_{0,\hat{\Omega}}^2+\|\hat{u}_{zzzt}\|_{0,\hat{\Omega}}^2\\
&\qquad+\|\hat{q}_{\sigma zzt}\|_{0,\hat{\Omega}}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{0,S^1}^2\Bigg)\\
&\qquad+a_{38}(|\hat{u}|_{3,1,\hat{\Omega}}^2+|\hat{q}_\sigma|_{2,1,\hat{\Omega}}^2+|\tilde{g}|_{2,1,\hat{\Omega}}^2)\\
&\qquad+a_{39}X_{10}(\hat{\Omega})(1+X_{10}^2(\hat{\Omega}))Y_{10}(\hat{\Omega}).
\end{aligned}
\tag{4.174}
$$

Going back to the variables ξ in (4.174) and using (4.160), we sum the inequalities over all neighborhoods of the partition of unity. Then, going back to the variables x in the followed estimate and using smallness of δ_{23}, we obtain (4.156). This concludes the proof.

To estimate the first term in the right-hand side of (4.156) we need the following result.

LEMMA 4.11. *For a sufficiently smooth solution of the problem* (4.1a–c) *the following inequality holds*:

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v_{xtt}^2+\frac{\mu+\nu}{p\Psi(\rho)}q_{\sigma xtt}^2\right)dx+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\bar{n}\cdot v_{tss^\alpha}\bar{n}\cdot v_{tss^\beta}ds\\
&\qquad+\|v_{tt}\|_{2,\Omega_t}^2+\|p_{\sigma tt}\|_{1,\Omega_t}^2\\
&\qquad\le\varepsilon_{11}\left(\|v_{xxxt}\|_{0,\Omega_t}^2+\|v_{xttt}\|_{0,\Omega_t}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{0,S^1}^2\right)\\
&\qquad+P_{23}(|v|_{3,0,\Omega_t}^2+|p_\sigma|_{2,0,\Omega_t}^2+|f|_{2,0,\Omega_t}^2)+P_{24}X_{11}(1+X_{11}^2)Y_{11},
\end{aligned}
\tag{4.175}
$$

where

$$
\begin{aligned}
X_{11}&=|v|_{3,0,\Omega_t}^2+|p_\sigma|_{3,0,\Omega_t}^2+\int_0^t\|v\|_{3,\Omega_\tau}^2d\tau,\\
Y_{11}&=|v|_{4,1,\Omega_t}^2+|p_\sigma|_{3,0,\Omega_t}^2+\int_0^t\|v\|_{4,\Omega_t}^2d\tau.
\end{aligned}
\tag{4.176}
$$

Proof. We also use the partition of unity. First we consider interior subdomains. Differentiating (4.27a) twice with respect to t and once with respect to ξ; multiplying

the result by $\tilde{u}_{tt\xi}A$ and integrating over $\hat{\Omega}$ yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}}\left(\eta\tilde{u}_{tt\xi}^2+\frac{1}{q\Psi(\eta)}\tilde{q}_{\sigma tt\xi}^2\right)Ad\xi\\
&\quad+\frac{\mu}{2}\int_{\tilde{\Omega}}(\nabla_{u^i}\tilde{u}_{tt\xi}^j+\nabla_{u^j}\tilde{u}_{tt\xi}^i)^2Ad\xi+(\nu-\mu)\|\nabla_u\cdot\tilde{u}_{tt\xi}\|_{0,\tilde{\Omega}}^2\\
&\quad\le\delta_1(\|\tilde{u}_{tt\xi\xi}\|_{0,\tilde{\Omega}}^2+\|\tilde{q}_{\sigma tt\xi}\|_{0,\tilde{\Omega}}^2)+a_1(|u|_{3,1,\tilde{\Omega}}^2+|q_\sigma|_{2,0,\tilde{\Omega}}^2+|\tilde{g}|_{2,0,\tilde{\Omega}}^2)\\
&\qquad+a_2X_{11}(\tilde{\Omega})(1+X_{11}^2(\tilde{\Omega}))Y_{11}(\tilde{\Omega}),
\end{aligned}\tag{4.177}$$

where

$$X_{11}(\tilde{\Omega})=|u|_{3,0,\tilde{\Omega}}^2+|q_\sigma|_{3,0,\tilde{\Omega}}^2+\int_{\tilde{\Omega}}\|u\|_{3,\tilde{\Omega}}^2dt,$$

$$Y_{11}(\tilde{\Omega})=|u|_{4,1,\tilde{\Omega}}^2+|q_\sigma|_{3,0,\tilde{\Omega}}^2+\int_{\tilde{\Omega}}\|u\|_{4,\tilde{\Omega}}^2dt.$$

We have also used the fact that

$$-\int_{\tilde{\Omega}}\tilde{q}_{\sigma tt\xi}\nabla_u\cdot\tilde{u}_{tt\xi}Ad\xi=\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}}\frac{1}{q\Psi(\eta)}\tilde{q}_{\sigma tt\xi}^2Ad\xi+N_1,$$

$$\left|\int_{\tilde{\Omega}}[(\nabla_u\nabla_u\tilde{u})_{,tt\xi}-\nabla_u\nabla_u\tilde{u}_{tt\xi}]\cdot\tilde{u}_{tt\xi}Ad\xi+\int_{\tilde{\Omega}}[(\nabla_u\tilde{q}_\sigma)_{,tt\xi}-\nabla_u\tilde{q}_{\sigma tt\xi}]\cdot\tilde{u}_{tt\xi}Ad\xi\right|$$

$$\le\delta_2\|\tilde{u}_{tt\xi\xi\xi}\|_{0,\tilde{\Omega}}^2+a_3X_{11}(\tilde{\Omega})(1+X_{11}^2(\tilde{\Omega}))Y_{11}(\tilde{\Omega}),$$

where

$$|N_1|\le\delta_3\|\tilde{q}_{\sigma tt\xi}\|_{0,\tilde{\Omega}}^2+a_4|\hat{u}|_{3,1,\hat{\Omega}}^2+a_5X_{11}(\tilde{\Omega})(1+X_{11}^2(\tilde{\Omega}))Y_{11}(\tilde{\Omega}).$$

From (4.34) we obtain

$$\begin{aligned}
\|\tilde{u}_{tt}\|_{2,\tilde{\Omega}}^2+\|\tilde{q}_{\sigma tt}\|_{1,\tilde{\Omega}}^2\le{}&c\|u_{zztt}\|_{0,\tilde{\Omega}}^2\\
&+a_6(|u|_{3,0,\tilde{\Omega}}^2+|q_\sigma|_{2,0,\tilde{\Omega}}^2+|\tilde{g}|_{2,0,\tilde{\Omega}}^2)\\
&+a_7X_{11}(\tilde{\Omega})(1+X_{11}^2(\tilde{\Omega}))Y_{11}(\tilde{\Omega}).
\end{aligned}\tag{4.178}$$

Now from (4.177) and (4.178) for sufficiently small δ_1 and Lemma 5.1 from [35] for $G=\tilde{\Omega}$ and $v=\tilde{u}_{tt\xi}$, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\tilde{\Omega}}\left(\eta\tilde{u}_{tt\xi}^2+\frac{1}{q\Psi(\eta)}\tilde{q}_{\sigma tt\xi}^2\right)Ad\xi+\|\tilde{u}_{tt}\|_{2,\tilde{\Omega}}^2+\|\tilde{q}_{\sigma tt}\|_{1,\tilde{\Omega}}^2\\
&\quad\le a_8(|u|_{3,0,\tilde{\Omega}}^2+|q_\sigma|_{2,0,\tilde{\Omega}}^2+|\tilde{g}|_{2,0,\tilde{\Omega}}^2)+a_9X_{11}(\tilde{\Omega})(1+X_{11}^2(\tilde{\Omega}))Y_{11}(\tilde{\Omega}).
\end{aligned}\tag{4.179}$$

Consider the boundary subdomains. Differentiating (4.28a) twice with respect to t and once with respect to τ, multiplying the result by $\tilde{u}_{tt\tau}J$, and integrating over $\hat{\Omega}$

gives

$$\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left[\hat\eta\tilde u_{tt\tau}^2+\frac{1}{\hat q\Psi(\hat\eta)}\tilde q_{\sigma tt\tau}^2\right]Jdz+\|\tilde u_{tt\tau}\|_{1,\hat\Omega}^2$$

$$+(\nu-\mu)\|(\operatorname{div}\tilde u)_{,tt\tau}\|_{0,\hat\Omega}^2-\int_{\hat S}(\hat n\hat{\mathbb{T}}(\tilde u,\tilde q_\sigma))_{,tt\tau}\cdot\tilde u_{tt\tau}Jdz'$$

$$\le\delta_4(\|\tilde q_{\sigma tt\tau}\|_{0,\hat\Omega}^2+\|\tilde u_{zztt}\|_{0,\hat\Omega}^2) \tag{4.180}$$

$$+a_{10}(|\hat u|_{3,0,\hat\Omega}^2+|q_\sigma|_{2,0,\hat\Omega}^2+|\tilde g|_{2,0,\hat\Omega}^2)$$

$$+a_{11}X_{11}(\hat\Omega)(1+X_{11}^2(\hat\Omega))Y_{11}(\hat\Omega),$$

where we have used Lemma 5.1 from [35] in the case $G=\hat\Omega$, $v=\tilde u_{tt\tau}$, and

$$\left|\int_{\hat\Omega}[(\hat\nabla\hat{\mathbb{T}}(\tilde u,\tilde q_\sigma))_{,tt\tau}-\hat\nabla\hat{\mathbb{T}}(\tilde u_{tt\tau},\tilde q_{\sigma tt\tau})]\tilde u_{tt\tau}Jdz\right|$$

$$\le\delta_5(\|\tilde u_{zztt}\|_{0,\hat\Omega}^2+\|\tilde q_{\sigma ztt}\|_{0,\hat\Omega}^2)+a_{12}|\tilde u|_{3,1,\hat\Omega}^2+a_{13}X_{11}(\hat\Omega)(1+X_{11}^2(\hat\Omega))Y_{11}(\hat\Omega),$$

$$\int_{\hat\Omega}\hat\nabla\cdot\tilde u_{tt\tau}\tilde q_{\sigma tt\tau}Jdz=-\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\frac{1}{\hat q\Psi(\hat\eta)}\tilde q_{\sigma tt\tau}^2Jdz+N_2,$$

and

$$|N_2|\le\delta_6\|\tilde q_{\sigma tt\tau}\|_{0,\hat\Omega}^2+a_{14}|\hat u|_{3,1,\hat\Omega}^2+a_{15}X_{11}(\hat\Omega)(1+X_{11}^2(\hat\Omega))Y_{11}(\hat\Omega).$$

Moreover, $X_{11}(\hat\Omega)$, $Y_{11}(\hat\Omega)$ are obtained from $X_{11}(\tilde\Omega)$, $Y_{11}(\tilde\Omega)$, replacing u, q_σ, $\tilde\Omega$, by $\hat u$, $\hat q_\sigma$, $\hat\Omega$, respectively.

Employing the boundary condition (4.28c) the boundary term in (4.180) can be expressed in the following way:

$$\int_{\hat S}(\hat n\hat{\mathbb{T}}(\tilde u,\tilde q_\sigma))_{,tt\tau}\cdot\tilde u_{tt\tau}Jdz'\le-\frac{\sigma}{2}\frac{d}{dt}\int_{\hat S}g^{\alpha\beta}\bar n\cdot\tilde u_{tss^\alpha}\bar n\cdot\tilde u_{tss^\beta}Jdz'+N_3, \tag{4.181}$$

where

$$|N_3|\le\delta_7\left(\|\hat u_{ztt\tau}\|_{0,\hat\Omega}^2+\|\hat u_{zzzt}\|_{0,\hat\Omega}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{2,S^1}^2\right)+a_{16}|\hat u|_{3,1,\hat\Omega}^2$$

$$+a_{17}X_{11}(\hat\Omega)(1+X_{11}(\hat\Omega))Y_{11}(\hat\Omega).$$

From (4.180) and (4.181) we obtain for sufficiently small δ_7,

(4.182)

$$\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left(\hat\eta\tilde u_{tt\tau}^2+\frac{1}{\hat q\Psi(\hat\eta)}\tilde q_{\sigma tt\tau}^2\right)Jdz'+\frac{\sigma}{2}\frac{d}{dt}\int_{\hat S}g^{\alpha\beta}\bar n\cdot\tilde u_{tss^\alpha}\bar n\cdot\tilde u_{tss^\beta}Jdz'$$

$$+\|\tilde u_{tt\tau}\|_{1,\hat\Omega}^2+(\nu-\mu)\|(\operatorname{div}\tilde u)_{,tt\tau}\|_{0,\hat\Omega}^2$$

$$\le\delta_8\left(\|\hat u_{ztt\tau}\|_{0,\hat\Omega}^2+\|\hat u_{zzzt}\|_{0,\hat\Omega}^2+\|\tilde q_{\sigma tt\tau}\|_{0,\hat\Omega}^2+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|_{2,S^1}^2\right)$$

$$+a_{18}(|\hat u|_{3,0,\hat\Omega}^2+|\hat q_\sigma|_{2,0,\hat\Omega}^2+|\tilde g|_{2,0,\hat\Omega}^2)+a_{19}X_{11}(\hat\Omega)(1+X_{11}^2(\hat\Omega))Y_{11}(\hat\Omega)).$$

Differentiating the third component of (4.42) twice with respect to t, multiplying the result by $\tilde{q}_{\sigma ntt}J$, and integrating over $\hat{\Omega}$ implies

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\frac{\mu+\nu}{\hat{q}\Psi(\hat{\eta})}\tilde{q}^2_{\sigma ntt}Jdz+\frac{1}{2}\|\tilde{q}_{\sigma ntt}\|^2_{0,\hat{\Omega}}\\
&\quad\le c\|\tilde{u}_{z\tau tt}\|^2_{0,\hat{\Omega}}\\
&\quad\quad+(\delta_9+cd)(\|\tilde{u}_{zztt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma ztt}\|^2_{0,\hat{\Omega}})\\
&\quad\quad+a_{20}(|\hat{u}|^2_{3,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,0,\hat{\Omega}}+|\tilde{g}|^2_{2,0,\hat{\Omega}})\\
&\quad\quad+a_{21}X_{11}(\hat{\Omega})(1+X^2_{11}(\hat{\Omega}))Y_{11}(\hat{\Omega}).
\end{aligned}\tag{4.183}$$

Differentiating the third component of (4.44) twice with respect to t, multiplying the result by $\tilde{u}^3_{nntt}J$, and integrating over $\hat{\Omega}$, we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat{\Omega}}\hat{\eta}|\tilde{u}^3_{nnt}|^2Jdz+\frac{\mu+\nu}{2}\|\tilde{u}^3_{nnnt}\|^2_{0,\hat{\Omega}}\\
&\quad\le c(\|\tilde{u}_{z\tau tt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma ntt}\|^2_{0,\hat{\Omega}})+\delta_{10}\|\tilde{u}_{ztttt}\|^2_{0,\hat{\Omega}}\\
&\quad\quad+(\delta_{11}+cd)(\|\tilde{u}_{zztt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma ztt}\|^2_{0,\hat{\Omega}})\\
&\quad\quad+a_{22}(|\hat{u}|^2_{3,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,0,\hat{\Omega}}+|\tilde{g}|^2_{2,0,\hat{\Omega}})+a_{23}X_{11}(\hat{\Omega})(1+X^2_{11}(\hat{\Omega}))Y_{11}(\hat{\Omega}).
\end{aligned}\tag{4.184}$$

From the problem (4.46)–(4.48) we have

$$\begin{aligned}
&\|\tilde{u}'_{z\tau tt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma\tau tt}\|^2_{0,\hat{\Omega}}\\
&\quad\le(\delta_{12}+cd)\,(\|\tilde{u}_{zztt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma ztt}\|^2_{0,\hat{\Omega}})\\
&\quad\quad+c\|\mathrm{div}\;\tilde{u}_{tt}\|^2_{1,\hat{\Omega}}+a_{24}(|\hat{u}|^2_{3,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,0,\hat{\Omega}}+|\tilde{g}|^2_{2,0,\hat{\Omega}})\\
&\quad\quad+a_{25}X_{11}(\hat{\Omega})(1+X^2_{11}(\hat{\Omega}))Y_{11}(\hat{\Omega}).
\end{aligned}\tag{4.185}$$

Moreover, from (4.46) it follows

$$\begin{aligned}
\|\tilde{u}'_{nntt}\|^2_{0,\hat{\Omega}}&\le c\,(\|(\mathrm{div}\;\tilde{u})_{,\tau tt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma\tau tt}\|^2_{0,\hat{\Omega}})\\
&\quad+(\delta_{13}+cd)(\|\tilde{u}_{zztt}\|^2_{0,\hat{\Omega}}+\|\tilde{q}_{\sigma ztt}\|^2_{0,\hat{\Omega}})\\
&\quad+a_{26}(|\hat{u}|^2_{3,0,\hat{\Omega}}+|\hat{q}_\sigma|^2_{2,0,\hat{\Omega}}+|\tilde{g}|^2_{2,0,\hat{\Omega}})+a_{27}X_{11}(\hat{\Omega})(1+X^2_{11}(\hat{\Omega}))Y_{11}(\hat{\Omega}).
\end{aligned}\tag{4.186}$$

Summarizing from (4.183)–(4.186) we have

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left(\hat\eta|\tilde u^3_{ntt}|^2+\frac{\mu+\nu}{\hat q\Psi(\hat\eta)}\tilde q^2_{\sigma ntt}\right)Jdz+\|\tilde u_{zztt}\|^2_{0,\hat\Omega}+\|\tilde q_{\sigma ztt}\|^2_{0,\hat\Omega}\\
&\quad\le\delta_{14}\|\tilde u_{ttt}\|^2_{1,\hat\Omega}+(\delta_{15}+cd)(\|\tilde u_{zztt}\|^2_{0,\hat\Omega}+\|\tilde q_{\sigma ztt}\|^2_{0,\hat\Omega})\\
&\qquad+c\|\tilde u_{z\tau tt}\|^2_{0,\hat\Omega}+a_{28}(|\hat u|^2_{3,0,\hat\Omega}+|\hat q_\sigma|^2_{2,0,\hat\Omega}+|\tilde g|^2_{2,0,\hat\Omega})\\
&\qquad+a_{29}X_{11}(\hat\Omega)(1+X^2_{11}(\hat\Omega))Y_{11}(\hat\Omega).
\end{aligned}
\tag{4.187}
$$

Finally, using the inequality

$$
\begin{aligned}
\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\hat\eta\tilde u^2_{ztt}Jdz&\le\delta_{16}\|\tilde u_{ttt}\|^2_{1,\hat\Omega}+c\|\tilde u_{tt}\|^2_{1,\hat\Omega}\\
&\quad+c(\|\hat u\|^2_{3,\hat\Omega}+|\hat q_\sigma|^2_{3,2,\hat\Omega})\|u_{ztt}\|^2_{0,\hat\Omega}.
\end{aligned}
\tag{4.188}
$$

From (4.182) and (4.187) we obtain

(4.189)

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\hat\Omega}\left(\hat\eta\tilde u^2_{ztt}+\frac{\mu+\nu}{\hat q\Psi(\hat\eta)}\tilde q^2_{\sigma ztt}\right)Jdz+\|\tilde u_{tt}\|^2_{2,\hat\Omega}+\|\tilde q_{\sigma tt}\|^2_{1,\hat\Omega}\\
&\quad\le\delta_{17}\left(\|\tilde u_{zzzt}\|^2_{0,\hat\Omega}+\|\tilde u_{ztttt}\|^2_{0,\hat\Omega}+\|\hat u_{zztt}\|^2_{0,\hat\Omega}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}\right)\\
&\qquad+a_{30}(|\hat u|^2_{3,0,\hat\Omega}+|\hat q_\sigma|^2_{2,0,\hat\Omega}+|\tilde g|^2_{2,0,\hat\Omega})+a_{31}X_{11}(\hat\Omega)(1+X^2_{11}(\hat\Omega))Y_{11}(\hat\Omega).
\end{aligned}
$$

Going back to the variables ξ in (4.189), then summing the result and (4.180) over all neighborhoods of the partition of unity, we finally obtain (4.175) after going back to the variables x and assuming that δ_{17} is sufficiently small. This concludes the proof.

Finally, to obtain an estimate for the second term in the right-hand side of the inequality (4.175) we have to show the result.

LEMMA 4.12. *For a sufficiently smooth solution of the problem* (4.1a–c) *we have*

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v^2_{ttt}+\frac{1}{p\Psi(\rho)}p^2_{\sigma ttt}\right)dx+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\bar n\cdot v_{tts^\alpha}\bar n\cdot v_{tts^\beta}ds\\
&\qquad+\|v_{ttt}\|^2_{1,\Omega_t}+\|p_{\sigma ttt}\|^2_{0,\Omega_t}\\
&\quad\le c\|v_{xtt}\|^2_{0,\Omega_t}+P_{25}(|v|^2_{3,0,\Omega_t}+\|f_{ttt}\|^2_{0,\Omega_t}+|f|^2_{2,0,\Omega_t})\\
&\qquad+P_{26}X_{12}(1+X^3_{12})Y_{12},
\end{aligned}
\tag{4.190}
$$

where $X_{12}=|v|^2_{3,0,\Omega_t}+|p_\sigma|^2_{3,0,\Omega_t}$, $Y_{12}=|v|^2_{4,1,\Omega_t}+|p_\sigma|^2_{3,0,\Omega_t}$.

Proof. Differentiating (4.1a) three times with respect to t, multiplying the result by v_{ttt}, integrating over Ω_t and using Lemma 5.5 from [35], we obtain

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_t}\left(\rho v^2_{ttt}+\frac{1}{p\Psi(\rho)}p^2_{\sigma ttt}\right)dx+\|v_{ttt}\|^2_{1,\Omega_t}\\
&\qquad-\int_{S_t}(n_iT^{ij}(v,p_\sigma))_{,ttt}\cdot v^j_{ttt}ds\le\delta_1\|v_{ttt}\|^2_{0,\Omega_t}\\
&\qquad+c(\|f_{ttt}\|^2_{0,\Omega_t}+|f|^2_{2,0,\Omega_t})+cX_{12}(1+X^3_{12})Y_{12},
\end{aligned}
\tag{4.191}
$$

where by the boundary condition (4.1c) the boundary term has the form

$$-\int_{S_t}(n_iT^{ij}(v,p_\sigma))_{,ttt}\,v^i_{ttt}ds=\frac{\sigma}{2}\frac{d}{dt}\int_{S_t}g^{\alpha\beta}\bar{n}\cdot v_{tts^\alpha}\bar{n}\cdot v_{tts^\beta}ds+N_1, \tag{4.192}$$

where

$$|N_1|\le\delta_2(\|v_{tttx}\|^2_{0,\Omega_t}+\|v_{ttxx}\|^2_{0,\Omega_t})+a_1|v|^2_{3,0,\Omega_t}+a_2X_{12}(1+X_{12}^2)Y_{12}.$$

By (4.2) we have

$$\|p_{\sigma ttt}\|^2_{0,\Omega_t}\le c\|v_{xtt}\|^2_{0,\Omega_t}+cX_{12}(1+X_{12}^2)Y_{12}. \tag{4.193}$$

Therefore, from (4.191)–(4.193) we obtain (4.190). This concludes the proof.

From the above lemmas for sufficiently small ε's we have the following result.

THEOREM 4.13. *Let*

$$\begin{aligned}\varphi(t)\equiv&\frac{1}{2}\int_{\Omega_t}\left(\rho|v|^2_{3,0}+\frac{1}{p\Psi(\rho)}|p_\sigma|^2_{3,0}\right)dx\\&+\frac{\sigma}{2}\int_{S_t}\tilde{\delta}^{\alpha\beta}\sum_{|k|\le2}\bar{n}\cdot\int_0^t\partial_s^kv_{s_1s^\alpha}dt'\bar{n}\cdot\int_0^t\partial_s^kv_{s_1s^\beta}dt'ds\\&+\frac{\sigma}{2}\int_{S_t}\sum_{|k|\le2}\left|\bar{n}\cdot\int_0^t\partial_s^kv_{s^1s^2}dt'\right|^2ds\\&+\frac{\sigma}{2}\int_{S_t}\sum_{|k|\le2}\sum_{i=1}^2\left(\frac{1}{2}\bar{n}\cdot\int_0^t\partial_s^kv_{s^is^i}dt'+2\partial_s^k\left(H(\cdot,0)+\frac{2}{R_0}\right)\right)^2ds\\&+\frac{\sigma}{2}\int_{S_t}g^{\alpha\beta}\left(\bar{n}\cdot\int_0^tv_{s^\alpha}dt'\bar{n}\cdot\int_0^tv_{s^\beta}dt'\right.\\&\qquad\left.+\sum_{|k|\le2}D^k_{t,s}v_{s^\alpha}\cdot\bar{n}D^k_{t,s}v_{s^\beta}\cdot\bar{n}\right)ds,\end{aligned} \tag{4.194}$$

$$\Phi_0(t)\equiv|v|^2_{4,1,\Omega_t}+|p_\sigma|^2_{3,0,\Omega_t},$$

$$\psi(t)\equiv\|v\|^2_{0,\Omega_t}+\|p_\sigma\|^2_{0,\Omega_t}+\|R(\cdot,t)-R(\cdot,0)\|^2_{0,S^1}.$$

For sufficiently smooth solutions of the problem (4.1a–c) *the following estimate holds*:

(4.195)

$$\begin{aligned}&\frac{d}{dt}\varphi+\Phi_0(t)\\&\quad\le c_1P(X)X(1+X^3)Y+c_2F+c_3\psi(t)\\&\qquad+c_4\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^4_{2,S^1}+\delta_1c_5\left(\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}+\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1}\right)\\&\qquad+c_6\left(\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\left\|\int_0^tvd\tau\right\|^2_{3,S_t}\right.\\&\qquad\qquad\left.+\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\left\|\int_0^tvd\tau\right\|^2_{4,S_t}\right),\end{aligned}$$

where $c_1, i = 1\text{–}6$ *depend on* ρ_*, ρ^*, T, a, b, *constants from theorems of imbedding and Korn inequalities* (*see* §5 *in* [35]); δ_1 *then is a small parameter, and*

$$
\begin{aligned}
X &= |v|^2_{3,0,\Omega_t} + |p_\sigma|^2_{3,0,\Omega_t} + \int_0^t \|v\|^2_{3,\Omega_{t'}} dt', \\
Y &= |v|^2_{4,1,\Omega_t} + |p_\sigma|^2_{3,0,\Omega_t} + \int_0^t \|v\|^2_{4,\Omega_{t'}} dt', \\
F &= |f|^2_{2,0,\Omega_t} + \|f_{ttt}\|^2_{0,\Omega_t}.
\end{aligned} \tag{4.196}
$$

Repeating the proof of Theorem 4.13 in the proper way we obtain the following.

THEOREM 4.14. *For sufficiently smooth solutions of the problem* (4.1a–c) *we have*

$$
\begin{aligned}
&\frac{d}{dt}\varphi(t) + \Phi_0(t) \\
&\quad \le c_7 P\left(\varphi_0 + \int_0^t \|v\|^2_{4,\Omega_t} d\tau\right)\left(\varphi_0 + \int_0^t \|v\|^2_{4,\Omega_t} d\tau\right) \\
&\quad \cdot \left[1 + \left(\varphi_0 + \int_0^t \|v\|^2_{4,\Omega_\tau} d\tau\right)^3\right] \Phi_0 + c_8 F + c_9 \psi(t) \\
&\qquad + c_{10} \left\| H(\cdot,0) + \frac{2}{R_0} \right\|^4_{2,S^1} \\
&\qquad + \delta_2 c_{11} \left(\left\| H(\cdot,0) + \frac{2}{R_0} \right\|^2_{2,S^1} + \|R(\cdot,t) - R(\cdot,0)\|^2_{4,S^1} \right) \\
&\qquad + c_{12} \left(\|R(\cdot,t) - R(\cdot,0)\|^2_{4+1/2,S^1} \left\| \int_0^t v d\tau \right\|^2_{3,S_t} \right. \\
&\qquad\qquad \left. + \|R(\cdot,t) - R(\cdot,0)\|^2_{3,S^1} \left\| \int_0^t v d\tau \right\|^2_{4,S_t} \right),
\end{aligned} \tag{4.197}
$$

where δ_2 *is a small parameter,* c_7–c_{12} *have the same properties as in Theorem* 4.13 *and*

$$
\varphi_0(t) = |v|^2_{3,2,\Omega_t} + |p_\sigma|^2_{3,2,\Omega_t}. \tag{4.198}
$$

5. Global existence. Let us introduce the spaces

$$
\mathcal{M}(t) = \left\{ (v, p_\sigma) : \varphi(t) + \int_0^t \Phi_0(\tau) d\tau < \infty \right\},
$$

and

$$
\mathcal{N}(t) = \{(v, p_\sigma) : \varphi(t) < \infty\},
$$

where $\varphi(t)$ and $\Phi_0(t)$ are defined in (4.194).

In this section, to prove the global existence of the solutions, we assume that the external force vanishes, so that

$$
f = 0. \tag{5.1}
$$

First we prove the following.

LEMMA 5.1. *Let the initial data* $v_0, p_{\sigma 0}, S$ *of* (1.1a–e) *be such that* $(v(0), p_\sigma(0)) \in \mathcal{N}(0)$ *and* $S \in H^{4+1/2}$. *Let*

$$\int_\Omega \rho_0 v_0 \cdot \eta dx = 0, \qquad \int_\Omega \rho_0 x dx = 0,$$

where $\eta = a + b \times x$, *and* a, b *are constant vectors.*

Let the initial data $v_0, p_{\sigma 0}, S$ *and the parameters of* (1.1a–e) $(p_0, \sigma, d, A, \kappa, M)$ *be such that*

$$\begin{aligned} &\varphi(0) \le \varepsilon_1, \qquad \chi(0) \equiv \left\| H(\cdot, 0) + \frac{2}{R_0} \right\|^2_{2+1/2, S^1} \le \varepsilon_2, \\ &\vartheta(t) \equiv \sup_{\tau \le t} \|R(\cdot, \tau) - R_0\|^2_{0, S^1} \le c_0 \varepsilon_0, \qquad t \le T, \end{aligned} \tag{5.2}$$

where $\varepsilon_0, \varepsilon_1, \varepsilon_2$ *are sufficiently small* (ε_0 *appears in Remark* 2.7—*see* (2.65)).

Then there exists a local solution of problem (1.1a–e) *such that* $v, p_\sigma \in \mathcal{M}(t), t \le T$, *where* T *is the time of local existence* (*see Theorem* 3.1) *and*

$$\varphi(t) + \int_0^t \Phi_0(\tau) d\tau \le c_1(\varphi(0) + \chi(0) + \vartheta(t)) \equiv c_1 A \le c_1(\varepsilon_0 + \varepsilon_1 + \varepsilon_2). \tag{5.3}$$

Moreover, we have that $S_t \in H^{4+1/2}$.

Proof. Take

$$v_0, p_{\sigma 0} \in H^3(\Omega), \qquad S \in H^{4+1/2} \tag{5.4}$$

such that the assumptions of the lemma hold. Then, in view of Theorem 3.1 and Remark 3.2 there exists a local solution of (1.1) such that

$$u \in W_2^{4,2}(\Omega^T), \qquad q_\sigma \in W_2^{3,3/2}(\Omega^T) \cap C(0, T; \Gamma_0^{3,3/2}(\Omega)), \tag{5.5}$$

where T is the time of local existence. Next, in view of Theorem 2.5, $S_t \in H^{4+1/2}$. Moreover, (3.6) implies

$$\begin{aligned} &\|u\|_{4,\Omega^T} + \|q_\sigma\|_{3,\Omega^T} + |q_\sigma|_{3,0,\infty,\Omega^T} \\ &\qquad \le c \left(\|v_0\|_{3,\Omega} + \|p_{\sigma 0}\|_{3,\Omega} + \left\| H(\cdot, 0) + \frac{2}{R_0} \right\|_{2+1/2, S^1} \right) \\ &\qquad \le c(\varphi(0) + \chi(0)) \le c(\varepsilon_1 + \varepsilon_2), \end{aligned} \tag{5.6}$$

where the last inequality follows from (5.2) and v, p_σ in (5.6) are written in the Lagrangian coordinates, so $u(\xi, t) = v(x(\xi, t), t)$, $q_\sigma(\xi, t) = p_\sigma(x(\xi, t), t)$.

Integrating the equation of continuity we have

$$\eta(\xi, t) = \rho_0(\xi) \exp\left[- \int_0^t \operatorname{div}_u u \, d\tau \right];$$

so, using (5.6), we have that $\eta_{tt} \in L_\infty(0,T;L_2(\Omega)) \cap L_2(0,T;H^1(\Omega))$, $\eta_t \in L_\infty(0,T;H^2(\Omega)) \cap L_2(0,T;H^3(\Omega))$, and the following estimates are valid:

$$\begin{aligned} &\sup_t (\|\eta_{tt}\|^2_{0,\Omega} + \|\eta_t\|^2_{2,\Omega} + \|\eta\|^2_{3,\Omega}) \\ &\qquad + \|\eta_{tt}\|^2_{1,2,\Omega^T} + \|\eta_t\|^2_{3,2,\Omega^T} \\ &\quad \le \varphi_1(T, \varphi(0) + \chi(0))(\varphi(0) + \chi(0)) \\ &\quad \le c(\varepsilon_1 + \varepsilon_2). \end{aligned} \tag{5.7}$$

Writing (4.2) in the Lagrangian coordinates, we have

$$q_{\sigma t} + q\Psi(\eta)\mathrm{div}_u u = 0;$$

so integration with respect to time yields

$$q_\sigma = q_\sigma(0) - \int_0^t q\Psi(\eta)\mathrm{div}_u u d\tau. \tag{5.8}$$

Using the estimate (5.6) for the local solutions we also obtain that q_σ belongs to the same spaces as η above and the following estimate for solution (5.8) holds:

$$\begin{aligned} N_1 \equiv &\sup_t (\|q_{\sigma tt}\|^2_{0,\Omega} + \|q_{\sigma t}\|^2_{2,\Omega} + \|q_\sigma\|^2_{3,\Omega}) \\ &+ \|q_{\sigma tt}\|^2_{1,2,\Omega^T} + \|q_{\sigma t}\|^2_{3,2,\Omega^T} \\ &\le \varphi_2(T, \varphi(0) + \chi(0))(\varphi(0) + \chi(0)) \\ &\le c(\varepsilon_1 + \varepsilon_2). \end{aligned} \tag{5.9}$$

In the above considerations we have used the imbeddings

$$\begin{aligned} N_2 &\equiv \sup_t (\|u\|^2_{3,\Omega} + \|u_t\|^2_{1,\Omega}) \\ &\le c(\|u\|^2_{4,\Omega^T} + \|u(0)\|^2_{3,\Omega} + |u(0)|^2_{1,0,\Omega}) \\ &\le c(\varphi(0) + \chi(0)) \le c(\varepsilon_1 + \varepsilon_2), \end{aligned} \tag{5.10}$$

and

$$\begin{aligned} \int_0^t |u_x|_{\infty,\Omega} d\tau &\le cT^{1/2} \left(\int_0^t \|u\|^2_{3,\Omega} d\tau \right)^{1/2} \le cT^{1/2} \|u\|_{4,\Omega^T} \\ &\le cT^{1/2}(\varphi(0) + \chi(0)) \le c(\varepsilon_1 + \varepsilon_2). \end{aligned} \tag{5.11}$$

By repeating the proof of Lemma 4.10, we have

$$\begin{aligned} &\frac{d}{dt} \|v_{xxt}\|^2_{0,\Omega_t} + \frac{\sigma}{2} \frac{d}{dt} \int_{S_t} g^{\alpha\beta} \bar{n} \cdot v_{sss^\alpha} \bar{n} \cdot v_{sss^\beta} ds \\ &\quad + \|v_{xxt}\|^2_{1,\Omega_t} \le (\varepsilon_1' + cN) \|v_{xttt}\|^2_{0,\Omega_t} + c \left(N, \int_0^T M(\tau) d\tau \right) M, \end{aligned} \tag{5.12}$$

where $N = N_1 + N_2$ and M is such an expression that

$$\int_0^t M d\tau \le c(\varphi(0) + \chi(0))$$

holds in virtue of the estimates for the local solution.

Similarly, using Lemma 4.11, we have

$$\begin{aligned}
&\frac{d}{dt}\left(\|v_{xtt}\|^2_{0,\Omega_t}+\|p_{\sigma xtt}\|^2_{0,\Omega_t}\right)+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}\bar{n}\cdot v_{tss^\alpha}\bar{n}\cdot v_{tss^\beta}ds\\
&\qquad+\|v_{xtt}\|^2_{1,\Omega_t}+\|p_{\sigma xtt}\|^2_{0,\Omega_t}\\
&\quad\le(\varepsilon_2'+cN)(\|v_{xttt}\|^2_{0,\Omega_t}+\|v_{xxtt}\|^2_{0,\Omega_t}+\|v_{xxxt}\|^2_{0,\Omega_t})\\
&\qquad+c\|v_{ttt}\|^2_{0,\Omega_t}+c\left(N,\int_0^T M(\tau)d\tau\right)M.
\end{aligned}\tag{5.13}$$

Finally, Lemma 4.12 implies

$$\begin{aligned}
&\frac{d}{dt}\|v_{ttt}\|^2_{0,\Omega_t}+\frac{\sigma}{2}\frac{d}{dt}\int_{S_t} g^{\alpha\beta}\bar{n}\cdot v_{tts^\alpha}\bar{n}\cdot v_{tts^\beta}ds\\
&\qquad+\|v_{ttt}\|^2_{1,\Omega_t}\le\varepsilon_3'\|v_{xxtt}\|^2_{0,\Omega_t}+cM\|v_t\|^2_{2,\Omega_t}+c(N)M,
\end{aligned}\tag{5.14}$$

where in virtue of the equation of continuity (4.2) we have

$$\|p_{\sigma ttt}\|^2_{0,\Omega_t}\le c\|v_{xtt}\|^2_{0,\Omega_t}+c(N)M.$$

From (5.12)–(5.14) and the last inequality we obtain for sufficiently small $\varepsilon_1',\varepsilon_2',\varepsilon_3'$, N and $\int_0^T Md\tau$ that

$$\begin{aligned}
&\sup_t(\|v_{xxt}\|^2_{0,\Omega_t}+\|v_{xtt}\|^2_{0,\Omega_t}+\|v_{ttt}\|^2_{0,\Omega_t}+\|p_{\sigma xtt}\|^2_{0,\Omega_t}+\|p_{\sigma ttt}\|^2_{0,\Omega_t})\\
&\quad+\sup_t\int_{S_t} g^{\alpha\beta}\left(\bar{n}\cdot v_{sss^\alpha}\bar{n}\cdot v_{sss^\beta}+\bar{n}\cdot v_{tss^\alpha}\bar{n}\cdot v_{tss^\beta}+\bar{n}\cdot v_{tts^\alpha}\bar{n}\cdot v_{tts^\beta}\right)ds\\
&\quad+\int_0^t(\|v_{xxt}\|^2_{1,\Omega_t}+\|v_{xtt}\|^2_{1,\Omega_t}+\|v_{ttt}\|^2_{1,\Omega_t})dt\le c\left(N,\int_0^t Md\tau\right).
\end{aligned}$$

Using (4.104) and (4.154) in (4.195) we have

$$\begin{aligned}
&\frac{d}{dt}\varphi+\Phi_0\\
&\quad\le c_1P(X)X(1+X^3)Y+c_2\psi(t)+c_3\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^4_{2,S^1}\\
&\qquad+\varepsilon c_4\left(\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1}+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}\right)\\
&\qquad+c_5(\|v\|^2_{4,\Omega_t}+\|p_\sigma\|^2_{3,\Omega_t}+\|R(\cdot,t)-R_0\|^2_{0,S^1}\\
&\qquad+\|R(\cdot,0)-R_0\|^2_{4+1/2,S^1})\int_0^t\|v\|^2_{3,\Omega_t}d\tau\\
&\qquad+c_6[\varepsilon(\|v\|^2_{4,\Omega_t}+\|p_\sigma\|^2_{3,\Omega_t})+\|R(\cdot,t)-R_0\|^2_{0,S^1}+\|R(\cdot,0)-R_0\|^2_{3,S^1}]\\
&\qquad\cdot\Bigg[\varepsilon(\|v\|^2_{4,\Omega_t}+\|p_\sigma\|^2_{3,\Omega_t})+\|v\|^2_{0,\Omega_t}+\|p_\sigma\|^2_{0,\Omega_t}+\int_0^t\|v\|^2_{0,\Omega_t}d\tau\\
&\qquad\quad+\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}+\|R(\cdot,t)-R_0\|^2_{0,S^1}+\|R(\cdot,t)-R_0\|^2_{0,S^1}\\
&\qquad\quad+(\|v\|^2_{3,\Omega_t}+\|p_\sigma\|^2_{2,\Omega_t})\left(1+\int_0^t\|v\|^2_{3,\Omega_\tau}d\tau\right)\int_0^t\|v\|^2_{4,\Omega_\tau}d\tau\Bigg].
\end{aligned}\tag{5.15}$$

Integrating (5.15) with respect to t and using the fact that $\varepsilon, \varepsilon_0, \varepsilon_1, \varepsilon_2$ are sufficiently small, we obtain

$$\begin{aligned}\varphi(t) &+ \int_0^t \Phi_0(\tau)d\tau \\ &\le c\int_0^t \psi(\tau)d\tau + c\left(\left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{2,S^1} + \sup_{\tau\le t}\|R(\cdot,\tau) - R_0\|^2_{0,S^1}\right) + \varphi(0) \\ &\le c(\varepsilon_0 + \varepsilon_1 + \varepsilon_2).\end{aligned}$$

Hence, we have shown that $(v, p_\sigma) \in \mathcal{M}(t), t \le T$ and (5.3) holds. We have to emphasize that to prove the above result the standard technique of mollifiers or differences should be used. This concludes the proof.

LEMMA 5.2. *Assume that there exists a local solution to* (1.1a–e), *which belongs to* $\mathcal{M}(t), t \le T$. *Let the assumptions of Lemma* 2.2 *be satisfied. Then there exists* $\delta = \delta(\delta', \varepsilon) \in (0,1)$ *such that*

$$\|p_\sigma\|^2_{0,\Omega_t} \le c_2\delta, \tag{5.16}$$

where $\delta' \in (0,1), \delta = c_3\varepsilon_1\delta' + c(\delta')\varepsilon_0$, $c(\delta')$ *is a decreasing function of* δ', *and* ε_0 *is taken from Remark* 2.7.

Proof. Let $\bar{p}_{\Omega_t} = (1/|\Omega_t|)\int_{\Omega_t} p dx$ and $p_{\Omega_t} = p - \bar{p}_{\Omega_t}$. Then

$$\|p_\sigma\|_{0,\Omega_t} \le \|p_{\Omega_t}\|_{0,\Omega_t} + \|\bar{p}_{\Omega_t} - p_0 - q_0\|_{0,\Omega_t}. \tag{5.17}$$

We introduce a function ϑ as a solution of the problem

$$\begin{aligned}\operatorname{div}\vartheta &= p_{\Omega_t} \quad \text{in } \Omega_t, \\ \vartheta &= 0 \quad \text{on } S_t.\end{aligned} \tag{5.18}$$

In view of Lemma 2.2 in [6] there exists $\vartheta \in \overset{0}{W}{}^1_2(\Omega_t) = \{u \in W^1_2(\Omega_t) : u_{|S_t} = 0\}$ such that

$$\|\vartheta\|_{1,\Omega_t} \le c\|p_{\Omega_t}\|_{0,\Omega_t}. \tag{5.19}$$

Multiplying (1.1a) written in the form

$$\rho v_t + \rho v\cdot\nabla v + \nabla p_{\Omega_t} - \operatorname{div}\mathbb{D}(v) = 0,$$

by ϑ, integrating the result over Ω_t, and performing integration by parts, we obtain

$$\int_{\Omega_t} p_{\Omega_t}\operatorname{div}\vartheta dx = \int_{\Omega_t}\mathbb{D}(v)\operatorname{div}\vartheta dx + \int_{\Omega_t}\rho(v_t + v\cdot\nabla v)\vartheta dx.$$

Taking (5.19) into account and using the fact that our local solution is such that $|\rho|_{\infty,\Omega_t} + |v|_{\infty,\Omega_t} \le c$, we have

$$\|p_{\Omega_t}\|^2_{0,\Omega_t} \le c(\|v_x\|^2_{0,\Omega_t} + \|v_t\|^2_{0,\Omega_t}). \tag{5.20}$$

To estimate the second term in the right-hand side of (5.17) we use

$$\|\bar{p}_{\Omega_t} - p_0 - q_0\|^2_{0,\Omega_t} = |\Omega_t||\bar{p}_{\Omega_t} - p_0 - q_0|^2 \tag{5.21}$$

and the relation

$$\bar{p}_{\Omega_t} - p_0 - q_0 = \bar{n} \cdot \mathbb{D}(v) \cdot \bar{n} - \sigma \left(H(\cdot, t) + \frac{2}{R_0} \right) - (p - \bar{p}_{\Omega_t}) \quad \text{on } S_t, \tag{5.22}$$

which follows form the boundary condition (1.1d).

From (5.21) and (5.22) we have

(5.23)

$$\begin{aligned} \|\bar{p}_{\Omega_t} - p_0 - q_0\|^2_{0,\Omega_t} &\le c \frac{|\Omega_t|}{|S_t|} \left(\|p - \bar{p}_{\Omega_t}\|^2_{0,S_t} + \|\mathbb{D}(v)\|^2_{0,S_t} + \left\| H(\cdot, t) + \frac{2}{R_0} \right\|^2_{0,S^1} \right) \\ &\le \varepsilon_1' (\|p_x\|^2_{0,\Omega_t} + \|v_{xx}\|^2_{0,\Omega_t}) + c(\varepsilon_1') (\|p - \bar{p}_{\Omega_t}\|^2_{0,\Omega_t} + \|v\|^2_{0,\Omega_t}) \\ &\quad + c \left\| H(\cdot, t) + \frac{2}{R_0} \right\|^2_{0,S^1}, \end{aligned}$$

where $c(\varepsilon_1')$ is a decreasing function of ε_1'.

Inequalities (5.17), (5.20), and (5.23) imply

$$\begin{aligned} \|p_\sigma\|^2_{0,\Omega_t} &\le \varepsilon_1' (\|p_x\|^2_{0,\Omega_t} + \|v_{xx}\|^2_{0,\Omega_t}) \\ &\quad + c(\|v_x\|^2_{0,\Omega_t} + \|v_t\|^2_{0,\Omega_t} + \|v\|^2_{0,\Omega_t}) + c \left\| H(\cdot, t) + \frac{2}{R_0} \right\|^2_{0,S^1} \\ &\le \varepsilon_2' (\|v\|^2_{4,\Omega^T} + \sup_{t \le T} \|p_\sigma\|^2_{1,\Omega_t}) + c \left(\|v\|^2_{0,\Omega_t} + \left\| H(\cdot, t) + \frac{2}{R_0} \right\|^2_{0,S^1} \right). \end{aligned} \tag{5.24}$$

Finally,

$$\begin{aligned} \left\| H(\cdot, t) + \frac{2}{R_0} \right\|^2_{0,S^1} &\le c \|R(\cdot, t) - R_0\|^2_{2,S^1} \\ &\le \varepsilon_3' \left\| H(\cdot, t) + \frac{2}{R_0} \right\|^2_{1,S^1} + c \|R(\cdot, t) - R_0\|^2_{0,S^1}. \end{aligned} \tag{5.25}$$

From (5.24) and (5.25) it follows

$$\begin{aligned} \|p_\sigma\|^2_{0,\Omega_t} &\le \varepsilon_4' \left(\|v\|^2_{4,\Omega^T} + \sup_{t \le T} \|p_\sigma\|^2_{1,\Omega_t} + \left\| H(\cdot, t) + \frac{2}{R_0} \right\|^2_{1,S^1} \right) \\ &\quad + c(\|v\|^2_{0,\Omega_t} + \|R(\cdot, t) - R_0\|^2_{0,S^1}). \end{aligned} \tag{5.26}$$

From (5.6) and Remark 2.7 (see (2.65)), the inequality (5.26) follows (5.16). This concludes the proof.

Now we have the following.

LEMMA 5.3. *Let $(v, p_\sigma) \in \mathcal{M}(t), t \le T$ be a solution of the problem* (1.1a–e). *Then $u(\xi, t) = v(x(\xi, t), t) \in C(t_0 + \lambda, T; H^4(\Omega))$ and the estimate holds*

$$\sup_{t_0 + \lambda < t \le T} \|v\|^2_{4,\Omega_t} \le c(\varphi(0) + \chi(0)) \equiv cA, \tag{5.27}$$

where $t_0 > 0, \lambda > 0, t_0 + \lambda < T$.

Proof. Let $\zeta_\lambda(t) \in C^\infty$ be such that $\zeta_\lambda(t) = 1$ for $t \geq t_0 + \lambda, \zeta_\lambda(t) = 0$ for $t \leq t_0 + \lambda/2, 0 \leq \zeta_\lambda(t) \leq 1, |\dot{\zeta}_\lambda(t)| \leq c/\lambda$, where $\dot{\zeta}_\lambda = (d/dt)\zeta_\lambda$. Let $u_\lambda = u\zeta_\lambda, q_\lambda = q\zeta_\lambda$. Then they are solutions of the problem (see (3.1) and (5.1))

(5.28)

$$
\begin{aligned}
&\eta u_{\lambda t} - \mu\nabla_u^2 u_\lambda - v\nabla_u\nabla_u \cdot u_\lambda = \nabla_u q_\lambda + \eta u\dot{\zeta}_\lambda \quad \text{in } \Omega^T,\\
&\Pi_0 \Pi\mathbb{D}_u(u_\lambda)\bar{n} = 0 \quad \text{on } S^T,\\
&\bar{n}_0\mathbb{D}_u(u_\lambda)\bar{n} - \sigma\bar{n}_0\int_0^t \Delta_{S_\tau}(\tau)u_\lambda(\tau)d\tau\\
&\quad = \int_0^t [\dot{\zeta}_\lambda\bar{n}_0\mathbb{T}(u,q_\sigma)\bar{n} + \sigma\bar{n}_0\zeta_\lambda\dot{\Delta}_{S_\tau}(\tau)\left(\xi + \int_0^\tau u(\tau')d\tau'\right) + \zeta_\lambda\partial_\tau(q_0\bar{n}_0\cdot\bar{n})]d\tau,\\
&+ q_\lambda\bar{n}_0\cdot\bar{n} \equiv \int_0^t B(\tau)d\tau + q_\lambda\bar{n}_0\cdot\bar{n} \quad \text{on } S^T,\\
&u_\lambda|_{t=0} = 0 \quad \text{in } \Omega,
\end{aligned}
$$

where $q_\lambda = q_\sigma\zeta_\lambda$ is treated as a given function, $\Pi_0 g = g - \bar{n}_0(\bar{n}_0\cdot g)$, $\Pi g = g - \bar{n}(\bar{n}\cdot g)$. The second boundary condition (5.28) follows from the following integration by parts:

$$
\begin{aligned}
0 &= \int_0^t \zeta_\lambda(\tau)\partial_\tau\left[\bar{n}_0\mathbb{T}(u,q_\sigma)\bar{n} - \sigma\bar{n}_0\Delta_{S_\tau}(\tau)\left(\xi + \int_0^\tau u(\tau')d\tau'\right) - q_0\bar{n}_0\cdot\bar{n}\right]d\tau\\
&= \int_0^t \partial_\tau[\zeta_\lambda\bar{n}_0\mathbb{T}(u,q_\sigma)\bar{n}]d\tau\\
&\quad - \int_0^t \left[\dot{\zeta}_\lambda\bar{n}_0\mathbb{T}(u,q_\sigma)\bar{n} + \sigma\bar{n}_0\zeta_\lambda\dot{\Delta}_{S_\tau}(\tau)\left(\xi + \int_0^\tau u(\tau')d\tau'\right) + q_\sigma\zeta_\lambda\partial_\tau(\bar{n}_0\cdot\bar{n})\right]d\tau\\
&\quad - \sigma\bar{n}_0\cdot\int_0^t \Delta_{S_\tau}(\tau)u_\lambda(\tau)d\tau.
\end{aligned}
$$

Next we introduce the differences: $u^{(s)}(\xi,t) = u_\lambda(\xi,t) - u'_\lambda(\xi,t), q^{(s)}(t) = q_\lambda(\xi,t) - q'_\lambda(\xi,t)$, where $w'(\xi,t) = w(\xi,t-s)$, $0 < s < t_0$. Therefore, we obtain the following equations:

(5.29)

$$
\begin{aligned}
&\eta u_t^{(s)} - \mu\nabla_u^2 u^{(s)} - \nu\nabla_u\nabla_u\cdot u^{(s)} = \nabla_u q^{(s)} - (\eta-\eta')u'_{\lambda t} + \mu(\nabla_u^2 - \nabla_{u'}^2)u'_\lambda\\
&\quad + \nu(\nabla_u\nabla_u - \nabla_{u'}\nabla_{u'})\cdot u'_\lambda + (\nabla_u - \nabla_{u'})q'_\lambda\\
&\quad + (\eta-\eta')u\dot{\zeta}_\lambda + \eta' u'(\dot{\zeta}_\lambda - \dot{\zeta}'_\lambda) \equiv E \text{ in } \Omega^T,\\
&\Pi_0\Pi\mathbb{D}_u(u^{(s)})\bar{n} = \Pi_0(\Pi\mathbb{D}_u(u'_\lambda)\bar{n} - \Pi'\mathbb{D}_{u'}(u'_\lambda)\bar{n}' \equiv F \quad \text{on } S^T,\\
&\bar{n}_0\mathbb{D}_u(u^{(s)})\bar{n} - \sigma\bar{n}_0\int_0^t \Delta_{S_\tau}(\tau)u^{(s)}(\tau)d\tau = q^{(s)}\bar{n}\cdot\bar{n}_0\\
&\quad + \bar{n}_0(\mathbb{D}_u(u'_\lambda)\bar{n} - \mathbb{D}_{u'}(u'_\lambda)\bar{n}') - \sigma\bar{n}_0\int_0^t (\Delta_{S_\tau}(\tau) - \Delta'_{S_\tau}(\tau))u'_\lambda d\tau\\
&\quad + \sigma\int_0^t (B(\tau) - B'(\tau'))d\tau + q'_\lambda\bar{n}_0\cdot(\bar{n}-\bar{n}') \equiv G_1 + \int_0^t G_2(\tau)d\tau \quad \text{on } S^T,\\
&u^{(s)}|_{t=0} = 0 \quad \text{in } \Omega.
\end{aligned}
$$

Let us introduce the notation: $\lambda \in (0,1), t_0 + \lambda < T, Q_\lambda = \Omega \times (t_0 + \lambda, T), G_\lambda = S \times (t_0 + \lambda, T)$.

From considerations in [34] we get

(5.30)
$$\|u^{(s)}\|_{4,Q_\lambda} \le c\left(\|E\|_{2,Q_{\lambda/2}} + \|F\|_{3-1/2,G_{\lambda/2}} + \|G_1\|_{3-1/2,G_{\lambda/2}} + \|G_2\|_{2-1/2,G_{\lambda/2}}\right).$$

Now we have to estimate the particular terms in the right-hand side of (5.30). Using the explicit form of E, F, G_1, G_2 we have

$$\|u^{(s)}\|_{4,Q_\lambda} \le c(A)As. \tag{5.31}$$

Hence we have

$$\begin{aligned}
&\| \|u(\cdot)\|_{4,\Omega} \|^2_{B^1_{2,\infty}(t_0+\lambda,T)} \\
&\quad = \sup_s \int_{t_0+\lambda}^T \frac{|\,\|u(t)\|_{4,\Omega} - \|u(t-s)\|_{4,\Omega}|^2}{s^2} dt + \int_{t_0+\lambda}^T \|u(t)\|^2_{4,\Omega} dt \\
&\quad \le \sup_s \int_{t_0+\lambda}^T \frac{\|u(t) - u(t-s)\|^2_{4,\Omega}}{s^2} dt + \int_{t_0+\lambda}^T \|u(t)\|^2_{4,\Omega} dt \le c(A)A.
\end{aligned}$$

Therefore, by imbedding theorems for Besov and Nikolskii spaces we have (see [13, Chap. 6.1])

$$B^\beta_{2,\infty}(0,T) \subset C([0,T]) \quad \text{for } \beta > \tfrac{1}{2}.$$

Hence (5.27) is shown. This concludes the proof.

Now we prove a result which guarantees a prolongation of the local solution. The result plays a crucial role in the proof of global existence. A similar result is shown in [33] and [35].

LEMMA 5.4. *Assume that there exists a local solution in $\mathcal{M}(t), t \le T$. Let $\bar{c}_1$ be a positive constant. Assume that*

$$\begin{aligned}
&\varphi(0) \le \gamma/2\bar{c}_1, \qquad \gamma \in (0, \tfrac{1}{2}], \\
&\tilde{\psi}(t) = \sup_{t' \le t} \psi(t') \le \delta_0, \qquad \left\| H(\cdot,0) + \frac{2}{R_0} \right\|^2_{2,S^1} \le d, \delta_0, \ d \in (0,1).
\end{aligned} \tag{5.32}$$

Then for sufficiently small γ, δ_0, and d we have

$$\varphi(t) \le \gamma/2\bar{c}_1, \qquad t \le T. \tag{5.33}$$

Proof. First we have to obtain an appropriate differential inequality which enables us to prove (5.33). Introducing the new quantity

$$\Phi_{00} = \Phi_0 + \left\| \int_0^t v d\tau \right\|^2_{4,S_t}, \tag{5.34}$$

and using (4.154) and (5.1) in (4.195) yields

(5.35)
$$\begin{aligned}\frac{d}{dt}\varphi + \Phi_{00} &\le cP(X)X(1+X^3)Y + c\psi_1(t)\\ &+ c\left(\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1} + \left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}\right)\\ &+ c\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^4_{2,S^1} + c\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\left\|\int_0^t v\,d\tau\right\|^2_{3,S_t}\\ &+ c\|R(\cdot,t)-R(\cdot,0)\|^2_{3,S^1}\left\|\int_0^t v\,d\tau\right\|^2_{4,S_t},\end{aligned}$$

where

(5.36)
$$\begin{aligned}\psi_1(t) = \|v\|^2_{0,\Omega_t} + \|p_\sigma\|^2_{0,\Omega_t} + \int_0^t \|v\|^2_{0,\Omega_\tau}\,d\tau\\ +\|R(\cdot,t)-R_0\|^2_{0,S^1} + \|R(\cdot,0)-R_0\|^2_{0,S^1}.\end{aligned}$$

The inequality (5.35) has been proved for the local solution. Therefore, in view of boundary conditions (4.1c), Lemmas 5.1 and 5.2, we have

(5.37)
$$\begin{aligned}&\|R(\cdot,t)-R(\cdot,0)\|^2_{4+1/2,S^1}\\ &\quad\le \|R(\cdot,t)-R_0\|^2_{4+1/2,S^1} + \|R(\cdot,0)-R_0\|^2_{4+1/2,S^1}\\ &\quad\le c(\|v\|^2_{4,\Omega_t} + \|p_\sigma\|^2_{3,\Omega_t}) + \|R(\cdot,t)-R_0\|^2_{0,S^1} + \|R(\cdot,0)-R_0\|^2_{4+1/2,S^1}\\ &\quad\equiv 0(\varepsilon_0+\varepsilon_1+\varepsilon_2).\end{aligned}$$

Using (5.37) with sufficiently small $\varepsilon_0, \varepsilon_1, \varepsilon_2$ in (5.35) we obtain

(5.38)
$$\begin{aligned}\frac{d}{dt}\varphi + \Phi_{00} &\le cP(X)X(1+X^3)Y + c\psi_1(t)\\ &+c\left(\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1} + \left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1}\right),\end{aligned}$$

where we have used the fact that $\|H(\cdot,0)+(2/R_0)\|^2_{2,S^1}$ is also small.

Employing the inequality

(5.39)
$$\begin{aligned}\|R(\cdot,t)-R(\cdot,0)\|^2_{4,S^1} &\le \|R(\cdot,t)-R_0\|^2_{4,S^1} + \|R(\cdot,0)-R_0\|^2_{4,S^1}\\ &\le \varepsilon(\|v\|^2_{4,\Omega_t} + \|p_\sigma\|^2_{3,\Omega_t}) + c\left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S^1} + c\psi_1(t),\end{aligned}$$

and introducing the new quantity

(5.40)
$$\Phi = \Phi_{00} + \left\|H(\cdot,0)+\frac{2}{R_0}\right\|^2_{2,S_t},$$

instead of (5.38), we get

$$\frac{d}{dt}\varphi + \Phi \le cP(X)X(1+X^3)Y + c\tilde{\psi}(t)c\left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{2,S^1}, \tag{5.41}$$

where we have used the fact that $\psi_1(t) \le c\tilde{\psi}(t)$.

Using $X(t) \le \varphi(t) + \int_0^t \Phi_0(\tau)d\tau$, $Y(t) \le \Phi_0(t) + \int_0^t \Phi_0(\tau)d\tau$,

$$\int_0^t \Phi_0(\tau)d\tau \le c\left(\tilde{\psi}(t) + \left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{2,S^1}\right) + \varphi(0) \le cA,$$

where the last inequality follows from (5.3), the first term on the right-hand side of (5.41) yields

$$P(X)X(1+X^3)Y \le cA\Phi + c\varphi(1+\varphi^3)\Phi_0 + cA^2,$$

where we have used that $\varphi(t) \le c\Phi(t)$. Hence, using the form of A, instead of (5.41), we get

$$\frac{d}{dt}\varphi + \frac{1}{2}\Phi \le \bar{c}_1\varphi(1+\varphi^3)\Phi + \bar{c}_1\left(\tilde{\psi}(t) + \left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{2,S^1}\right) + \bar{c}_1\varphi^2(0), \tag{5.42}$$

where $\bar{c}_1$ is a constant which bounds all previous constants.

Assume that $t_* = \inf\{t \in [0,T] : \varphi(t) > \gamma/2\bar{c}_1\}$. Consider (5.42) in the interval $[0,t_*]$. From the definition of t_* we have that $\varphi(t_*) = \gamma/2\bar{c}_1$. Then for $t \le t_*$ we have

$$\bar{c}_1\tilde{\psi}(t) + \bar{c}_1\left\|H(\cdot,0) + \frac{2}{R_0}\right\|^2_{2,S^1} + \bar{c}_1\varphi^2(0) \le \bar{c}_1\left(\delta_0 + d + \frac{\gamma^2}{4\bar{c}_1^2}\right).$$

Let us assume that γ, δ_0, d are so small that

$$\bar{c}_1\left(\delta_0 + d + \frac{\gamma^2}{4\bar{c}_1^2}\right) \le \frac{\bar{c}_2}{16\bar{c}_1}\gamma, \tag{5.43}$$

where $\bar{c}_2$ is a constant from the inequality

$$\bar{c}_2\varphi(t) \le \Phi(t). \tag{5.44}$$

Therefore, from (5.42) we obtain

$$\varphi_t(t_*) \le -\Phi\left(\frac{1}{2} - \bar{c}_1\left(\frac{\gamma}{2\bar{c}_1} + \frac{\gamma^3}{2\bar{c}_1^3}\right)\right) + \frac{\bar{c}_2}{16\bar{c}_1}\gamma.$$

So, knowing that (5.44) holds, we have

$$\varphi_t(t_*) \le -\frac{\bar{c}_2\gamma}{2\bar{c}_1}\left[\frac{1}{2} - \frac{\gamma}{2} - \frac{\gamma^3}{8}\right] + \frac{\bar{c}_2}{16\bar{c}_1}\gamma \le -\frac{\bar{c}_2\gamma}{2\bar{c}_1}\left(\frac{1}{4} - \frac{1}{64}\right) + \frac{\bar{c}_2\gamma}{16\bar{c}_1} < 0.$$

Hence $\varphi_t(t_*) < 0$, a contradiction. This concludes the proof.

Finally, we prove the main result of this paper.

THEOREM 5.5. *Assume $f = 0$ and relations* (2.35) *among the constant parameters $\mu, \nu, \sigma, M, A, \kappa, |S|, |\Omega|, \int_\Omega \rho_0 v_0^2 dx, \int_\Omega \rho_0^\kappa dx, p_0$ of* (1.1) *such that the quantities*

$$\int_{\Omega_t} \rho v^2 dx, \quad \psi_t - \psi_{t'}, \quad |S_t| - |S_{t'}|, |\Omega_t| - |\Omega_{t'}| \quad \forall t, t' \in \mathbb{R}^1_+, \tag{5.45}$$

and

$$B \equiv \frac{1}{2}\int_\Omega \rho_0 v_0^2 dx + \psi - \psi_* + p_0(|\Omega| - |\Omega_*|) + \sigma(|S| - |S_*|) \le \varepsilon_0$$

are sufficiently small and $\psi_t = (A/(\kappa - 1))\int \rho^\kappa dx, \psi = \psi_0, \psi_* = \min_t \psi_t, |\Omega_*| = \min_t |\Omega_t|, |S_*| = 4\pi R_*^2$, and $(4\pi/3)R_*^3 = |\Omega_*|$ (see also Lemmas 2.1 *and* 2.2 *and Remark* 2.7).

Assume that $\rho_0 \in H^3(\Omega)$, $v_0 \in H^6(\Omega)$ are such that

$$\int_\Omega \rho_0 v_0 \cdot (a + b \times x) dx = 0, \qquad \int_\Omega \rho_0 x dx = 0, \tag{5.46}$$

where a, b are arbitrarily constant vectors, and

$$\varphi(0) \le \alpha_1, \tag{5.47}$$

where $p_\sigma = p(\rho) - p_0 - q_0, q_0 = 2\sigma/R_0, R_0 = ((3/4\pi)|\Omega|)^{1/3}$ and α_1 is sufficiently small.

Assume that S_t is described by $|x| = R(\omega, t), \omega \in S^1$ (unit sphere) and the initial boundary $S = S_0$ belongs to $H^{4+1/2}$ and is very close to a ball, so that

$$\left\| H(\cdot, 0) + \frac{2}{R_0} \right\|^2_{2, S^1} \le \alpha_2, \tag{5.48}$$

where α_2 is sufficiently small.

Assume compatibility conditions (for more explanation see Remark 5.7)

$$D_s^\alpha \partial_t^i (\mathbb{T}(v, p)\bar{n} - \sigma H \bar{n} + p_0 \bar{n})|_{t=0, S} = 0, \qquad |\alpha| + i \le 2.$$

Then there exists a global solution of (1.1a–e) *such that $(v, p_\sigma) \in \mathcal{M}(t), S_t \in H^{4+1/2}, t \in \mathbb{R}^1_+$ and*

$$\varphi(t) \le \alpha_1, \qquad \left\| H(\cdot, t) + \frac{2}{R_0} \right\|_{2, S^1} \le \alpha_2. \tag{5.49}$$

Moreover,

$$\int_\Omega \rho v \cdot (a + b \times x) dx = 0, \qquad \int_{\Omega_t} \rho x dx = 0,$$

and

$$\frac{1}{2}\int_{\Omega_t} \rho v^2 dx + \psi_t - \psi_* + p_0(|\Omega_t| - |\Omega_*|) + \sigma(|S_t| - |S_*|) \le B \le \varepsilon_0.$$

Proof. In view of (5.47) and (5.48) from Lemma 5.1 it follows that there exists a local solution $(v, p_\sigma) \in \mathcal{M}(t), t \le T$, such that

$$\varphi(T) + \int_0^T \Phi_0(\tau)d\tau \le c_1(\alpha_1 + \alpha_2), \tag{5.50}$$

and T is the time of local existence.

Then for $t \le T$ we have

$$\begin{aligned}
|R(\cdot,t) - R_0| &\le |R(\cdot,0) - R_0| + |R(\cdot,t) - R(\cdot,0)| \\
&\le \varepsilon \left\| H(\cdot,0) + \frac{2}{R_0} \right\|^2_{0,S^1} + c\|R(\cdot,0) - R_0\|^2_{0,S^1} + c\left|\int_0^t u d\tau\right| \left(1 + \left|\int_0^t u d\tau\right|\right) \\
&\le c_2(\alpha_1 + \alpha_2),
\end{aligned}$$

where we have used that

$$|R(\cdot,t) - R(\cdot,0)| = ||x| - |\xi|| = |x^2 - \xi^2|/(|x| + |\xi|) \quad \text{and} \quad x = \xi + \int_0^t u d\tau,$$

$$|\nabla R(t)| \le |\nabla R(0)| + |\int_0^t u(\tau)d\tau| + |\int_0^t \nabla u(\tau)d\tau| \le c_3(\alpha_1 + \alpha_2).$$

Assume that $\alpha_1 + \alpha_2$ is so small that (2.46) holds with a sufficiently small δ. Then (2.47) is valid. Next, sufficient smallness of quantities in (5.45) implies assumptions of Lemma 2.2. Thus, in view of Lemma 2.4 and Remark 2.7 there exists small $\delta_3 = c\varepsilon_0$ such that

$$\|v\|^2_{0,\Omega_t} + \|R(\cdot,t) - R_0\|^2_{0,S^1} \le \alpha_3. \tag{5.51}$$

Next in view of Lemma 5.2 we have that

$$\|p_\sigma\|^2_{0,\Omega_t} \le \alpha_4 = c_4[c_1(\alpha_1 + \alpha_2)\delta' + c_5\alpha_3],$$

where $\delta' \in (0,1)$, so $\alpha_4 > \alpha_3$. Therefore,

$$\tilde{\psi}(t) \le c_6(\alpha_3 + \alpha_4) \equiv \alpha_0.$$

Then in view of Lemma 5.3 we have

$$\|v\|^2_{4,\Omega_t} \le c_7(\alpha_1 + \alpha_2), \tag{5.52}$$

so (5.50) and (5.52) imply

$$\|v\|^2_{4,\Omega_t} + \|p_\sigma\|^2_{3,\Omega_t} \le c_8(\alpha_1 + \alpha_2). \tag{5.53}$$

Therefore, from the boundary conditions we get that $S_t \in H^{4+1/2}$ and

$$\left\| H(\cdot,T) + \frac{2}{R_0} \right\|^2_{2,S^1} \le \varepsilon c_8(\alpha_1 + \alpha_2) + c(\varepsilon)\alpha_0, \qquad \varepsilon \in (0,1). \tag{5.54}$$

Knowing that

$$|R_T - R_0| \le c|\Omega_T - \Omega| \le c_9\alpha_0$$

we obtain

$$\left\| H(\cdot,T) + \frac{2}{R_T} \right\|^2_{2,S^1} \leq \varepsilon c_8(\alpha_1+\alpha_2) + c_{10}\alpha_0 \leq \alpha_2, \tag{5.55}$$

where the last inequality imposes a restriction on α_0 (it must be sufficiently small). Finally, Lemma 5.4 yields

$$\varphi(T) \leq \alpha_1. \tag{5.56}$$

Therefore (5.47) and (5.48) are satisfied for $t = T$.

Now we are in a position to extend the considerations for interval $[T, 2T]$. In view of (5.56) and (5.51) for $t = T$, Remark 3.2 implies local existence of solutions for $t \in [T, 2T]$ which is such that

$$\|u\|_{4,\Omega\times(T,2T)} + \|q_\sigma\|_{3,\Omega\times(T,2T)} + |q_\sigma|_{3,0,\infty,\Omega\times(T,2T)} \leq c_{11}(\alpha_1+\alpha_3),$$

where on the right-hand side we generally have the same bound as for $t \in [0,T]$. Therefore,

$$\left| \int_T^{2T} u d\tau \right| \leq c_{12}(\alpha_1+\alpha_3),$$

so the change of the shape of Ω_t is as small as for interval $[0,T]$. Hence the Korn inequalities and imbedding theorems necessary in the proof of (4.195) can be applied with the same constants. This follows that the same inequality (4.195) holds for $t \in [T, 2T]$. Continuing the considerations, we prove global existence. This concludes the proof.

THEOREM 5.6 (case with $p_0 = 0$). *Let the assumptions of Theorem* 5.5 *with* $p_0 = 0$ *be satisfied. Then there exists a global solution*

$$v, p_\sigma = p - 2\sigma/R_0 \in \mathcal{M}(t), \qquad S_t \in H^{4+1/2}t \in \mathbb{R}^1_+$$

such that (5.49) *holds.*

The proof is the same as in Theorem 5.5.

Remark 5.7. We express explicitly the compatibility conditions formulated in assumptions of Theorem 5.5. Let $i = 0$. Then they take the form

$$D_s^\alpha \left(\mathbb{D}(v_0)\bar{n}_0 - \left(p(\rho_0) - p_0 - \frac{2\sigma}{R_0} \right) \bar{n}_0 - \frac{\sigma}{\sqrt{g_0}} \partial_{s_\alpha} \sqrt{g_0} g_0^{\alpha\beta} \xi_\beta \right) |_S = 0, \qquad |\alpha| \leq 2, \tag{5.57}$$

where $\bar{n}_0 = \xi_1 \times \xi_2 / |\xi_1 \times \xi_2|, \xi_1 = \xi_{s_i}, i = 1,2, g_{0\alpha\beta} = \xi_\alpha \cdot \xi_\beta, g_0 = \det\{g_{0\alpha\beta}\}, g_0^{\alpha\beta}$ is the inverse matrix to $g_{0\alpha\beta}$.

For $i = 1$ we have

$$D_s^\alpha \left(\mathbb{D}(v_t(0))\bar{n}_0 + \mathbb{D}(v_0)\bar{n}_t(0) - p_t(0)\bar{n}_0 - \left(p(\rho_0) - p_0 - \frac{2\sigma}{R_0} \right) \bar{n}_t(0) \right.$$
$$\left. - \partial_t(\Delta_{S_t} x)|_{t=0} \right) |_S = 0, \qquad |\alpha| \leq 1,$$

where $v_t(0), p_t(0)$ are calculated from (1.1a, b) at $t = 0$,

$$\bar{n}_t(0) = \frac{v_{01} \times \xi_2 + \xi_1 \times v_{02}}{|\xi_1 \times \xi_2|} - \frac{\xi_1 \times \xi_2}{|\xi_1 \times \xi_2|^3}(v_{01} \times \xi_2 + \xi_1 \times v_{02}) \cdot (\xi_1 \times \xi_2),$$
$$\partial_t(\Delta_{S_t} x)|_{t=0} = \partial_t(\Delta_{S_t})|_{t=0}\xi + \Delta_S v_0$$

and coefficients of $\partial_t(\Delta_{S_t})|_{t=0}$ depend on $g_{0\alpha\beta}, \partial_t g_{\alpha\beta}|_{t=0} = v_{0\alpha} \cdot \xi_\beta + \xi_\alpha \cdot v_{0\beta}$ and their derivatives with respect to $s = (s_1, s_2)$. Finally, the case $i = 2$ gives

$$\Big(\mathbb{D}(v_{tt}(0))\bar{n}_0 + 2\mathbb{D}(v_t(0))\bar{n}_t(0) + \mathbb{D}(v_0)\bar{n}_{tt}(0) - p_{tt}(0)\bar{n}_0 - 2p_t(0)\bar{n}_t(0)$$
$$-\Big(p(\rho_0) - p_0 - \frac{2\sigma}{R_0}\Big)\bar{n}_{tt}(0) - \partial_t^2(\Delta_{S_t} x)|_{t=0}\Big)|_S = 0,$$

where $v_{tt}(0), p_{tt}(0), \bar{n}_{tt}(0)$ are calculated inductively by employing (1.1a, b) at $t = 0$ and

$$\partial_t^2(\Delta_{S_t} x)|_{t=0} = (\partial_t^2 \Delta_{S_t})|_{t=0}\xi + 2(\partial_t \Delta_{S_t})|_{t=0} v_0 + \Delta_S v_t(0),$$

where coefficients of $(\partial_t^2 \Delta_{S_t})|_{t=0}$ depend on $\partial_t^2 g_{\alpha\beta}|_{t=0} = v_{\alpha t}(0) \cdot \xi_\beta + 2v_{0\alpha} \cdot v_{0\beta} + \xi_\alpha \cdot v_{\beta t}, \partial_t g_{\alpha\beta}|_{t=0}$ and $g_{0\alpha\beta}$.

Acknowledgment. The author is very indebted to professor V. A. Solonnikov for very fruitful discussions during the preparation of this paper.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] G. ALLAIN, *Small-time existence for the Navier–Stokes equations with a free surface*, Appl. Math. Optim., 16 (1987), pp. 37–50.
[3] J. T. BEALE, *The initial value problem for the Navier–Stokes equations with a free boundary*, Comm. Pure Appl. Math., 31 (1980), pp. 359–392.
[4] ———, *Large time regularity of viscous surface waves*, Arch. Rational Mech. Anal., 84 (1984), pp. 307–352.
[5] O. V. BESOV, V. P. ILYIN, AND S. M. NIKOLSKII, *Integral Representations of Functions and Imbedding Theorems*, Nauka, Moscow, 1975 (in Russian); Scripta Series in Mathematics, Halsted Press Books, Washington D. C., Winston 1979. (In English.)
[6] O. A. LADYZHENSKAYA AND V. A. SOLONNIKOV, *On some problems of vector analysis and generalized formulations of boundary problems for Navier–Stokes equations*, Zap. Nauchn. Sem. LOMI, 59 (1976), pp. 81–116. (In Russian.)
[7] L. LANDAU AND E. LIFSCHITZ, *Mechanics of Continuum Media*, Nauka, Moscow, 1984 (in Russian); Pergamon Press, Oxford, 1959 (in English); Hydrodynamics, Nauka, Moscow, 1986. (In Russian.)
[8] 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.
[9] ———, *The initial value problem for the equations of motion of compressible viscous and heat-conductive fluids*, Proc. Japan. Acad. Ser. A, 55 (1979), pp. 337–342.
[10] ———, *The initial boundary value problem for the equations of motion of compressible viscous and heat-conductive fluid*, Univ. of Wisconsin, MRC Technical Summary Report 2237, 1981, preprint.
[11] ———, *The initial boundary value problems for the equations of motion of general fluids*, in Computing Methods in Applied Sciences and Engineering, R. V. Glowinski and J. L. Lions, eds., North-Holland, Amsterdam, 1982.
[12] ———, *Initial boundary value problems for the equations of motion of compressible viscous and heat-conductive fluids*, Comm. Math. Phys., 89 (1983), pp. 445–464.
[13] S. M. NIKOLSKII, *Approximation of Functions of Multiple Variables and Imbedding Theorems*, Nauka, Moscow, 1977. (In Russian.)
[14] T. NISHIDA, *Equations of fluid dynamics: free surface problems*, Comm. Pure Appl. Math., 39 (1986), pp. 221–238.

[15] K. PILECKAS AND W. M. ZAJACZKOWSKI, *A free boundary problem for stationary compressible Navier–Stokes equations*, Comm. Math. Phys., 129 (1990), pp. 169–204.
[16] P. SECCHI AND A. VALLI, *A free boundary problem for compressible viscous fluids*, J. Reine Angew. Math, 341 (1983), pp. 1–31.
[17] P. SECCHI, *On the uniqueness of motion of viscous gaseous stars*, Math. Meth. Appl. Sci., 13 (1990), pp. 391–404.
[18] ———, *On the motion of gaseous stars in the presence of radiation*, Comm. Partial Differential Equations, 15 (1990), pp. 185–204.
[19] ———, *On the Evolution Equations of Viscous Gaseous Stars*, Ann. Scuola Norm. Sup. Pisa Ser. IV, 18 (1991), pp. 295–318.
[20] V. A. SOLONNIKOV, *Free boundary problems and problems in noncompact domains for the Navier–Stokes equations*, in Proc. Intern. Congress Math., Berkeley, CA, 1986, pp. 1113–1122.
[21] ———, *On an initial-boundary value problem for the Stokes system which appears in free boundary problems*, Trudy Math. Inst. Steklov, 188 (1990), pp. 150–188. (In Russian.)
[22] ———, *On the solvability of the initial-boundary value problem for equations of motion of the viscous compressible fluid*, Zap. Nauchn. Sem. LOMI, 56 (1976), pp. 128–142. (In Russian.)
[23] ———, *On an unsteady flow of a finite mass of a liquid bounded by a free surface*, Zap. Nauchn. Sem. LOMI, 152 (1986), pp. 137–157 (in Russian); J. Soviet Math., 40 (1988), pp. 672–686. (In English.)
[24] ———, *On boundary problems for linear parabolic systems of differential equations of general type*, Trudy Mat. Inst. Steklov, 83 (1965) (in Russian); Proc. Steklov Inst. Math., 83 (1967). (In English.)
[25] ———, *Solvability of the evolution problem for an isolated mass of a viscous incompressible capillary liquid*, Zap. Nauchn. Sem. LOMI, 140 (1984), pp. 179–186 (in Russian); J. Soviet Math., 32 (1986), pp. 223–238. (In English.)
[26] ———, *On an unsteady motion of an isolated volume of a viscous incompressible fluid*, Izv. Akad. Nauk. SSSR Ser. Mat., 51 (1987), pp. 1065–1087. (In Russian.)
[27] ———, *Solvability of a problem on the motion of a viscous incompressible fluid bounded by a free surface*, Izv. Akad. Nauk SSSR Ser. Mat., 41 (1977), pp. 1388–1424 (in Russian); Math. USSR Izv., 11 (1977), pp. 1323–1358. (In English.)
[28] ———, *Estimates of solutions of an initial-boundary value problem for the linear nonstationary Navier–Stokes system*, Zap. Nauchn. Sem LOMI, 59 (1976), pp. 178–254 (in Russian); J. Soviet Math., 10 (1978), pp. 336–393. (In English.)
[29] ———, *On the solvability of the second initial-boundary value problem for the linear nonstationary Navier–Stokes system*, Zap. Nauch. Sem. LOMI, 69 (1977), pp. 200–218 (in Russian); J. Soviet Math., 10 (1978), pp. 141–193. (In English.)
[30] ———, *A priori estimates for second order parabolic equations*, Trudy Mat. Inst. Steklov., 70 (1964), pp. 133–212.
[31] V. A. SOLONNIKOV AND A. TANI, *Free boundary problem for a viscous compressible flow with surface tension*, Zap. Nauchn. Sem. LOMI, 182 (1990), pp. 142–148; also in Constantin Caratheodory: an International Tribute, M. Rassias, ed., Vol. 2, World Scientific, 1991, pp. 1270–1303.
[32] A. VALLI, *Periodic and stationary solutions for compressible Navier–Stokes equations via a stability method*, Ann. Scuola Norm. Super. Pisa, 4 (1983), pp. 607–647.
[33] A. VALLI AND W. M. ZAJACZKOWSKI, *Navier–Stokes equations for compressible fluids: global existence and qualitative properties of the solutions in the general case*, Comm. Math. Phys., 103 (1986), pp. 259–296.
[34] W. M. ZJACZKOWSKI, *On an initial-boundary value problem for the parabolic system which appears in free boundary problems for compressible Navier–Stokes equations*, Dissertations Math., 304 (1990).
[35] ———, *On nonstationary motion of a compressible viscous fluid bounded by a free surface*, Dissertations Math., 324 (1993).
[36] ———, *On local motion of a compressible viscous fluids bounded by a free surface*, in Partial Differential Equations, Vol. 27, Banach Center Publ., Warsaw, 1992.
[37] ———, *Existence of local solutions for free boundary problems for viscous compressible barotropic fluids*, Ann. Polon. Math., to appear.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 85–98, January 1994

002

SOLUTIONS FOR TWO-DIMENSIONAL SYSTEM FOR MATERIALS OF KORTEWEG TYPE*

HARUMI HATTORI† AND DENING LI‡

Abstract. Dunn and Serrin [*Arch. Rational Mech. Anal.*, 88 (1985), pp. 95–133] proposed the interstitial working term and modified the system of compressible fluids based on the Korteweg theory of capillarity. This term was introduced to overcome a difficulty: the higher-order terms of density are not compatible with the classical theory of thermodynamics. In this paper the existence of local solutions to the above system in the multidimensional case are discussed.

Key words. Korteweg material, linearization, local solution

AMS subject classifications. primary 35M10, 35Q35, 76N15; secondary 35K30

1. Introduction.

In order to model the capillarity effect of materials, Korteweg [5] formulated a constitutive equation for the Cauchy stress that includes density gradients. Specifically, he proposed a compressible fluid model in which the "elastic" or "equilibrium" portion of the Cauchy stress tensor $\mathbf{T}$ is given by

$$\mathbf{T} = (-p + \alpha\triangle\rho + \beta|\nabla\rho|^2)\mathbf{I} + \delta\nabla\rho\otimes\nabla\rho + \gamma\nabla\otimes\nabla\rho, \tag{1.1}$$

where

- $\rho = \rho(\mathbf{x}, t)$ is the density of the fluid at the point $\mathbf{x}$ at time t;
- $\nabla\rho$ and $\triangle\rho$ are the gradient and Laplacian of ρ with respect to $\mathbf{x}$;
- $(\mathbf{a}\otimes\mathbf{b})_{ij} = a_i b_j$ is the tensor product of $\mathbf{a}$ and $\mathbf{b}$;
- α, β, δ, and γ are functions of density ρ and temperature θ;
- $p = p(\rho, \theta)$ is the pressure.

This form of $\mathbf{T}$ is a special example of an elastic material of grade N ($N = 2$ in this case). The difficulty with these higher-grade models is that they are, in general, incompatible with conventional thermodynamics. In order to remedy this difficulty, Dunn and Serrin [2] proposed the concept of interstitial working w. It turns out that w must have the form

$$w = \mathbf{w}\cdot\mathbf{n},$$

where $\mathbf{n}$ is the outer unit normal to the boundary of the domain in which the usual set of integral balance laws is postulated and $\mathbf{w}$ is called the interstitial work flux representing spatial interactions of longer range.

Employing this interstitial working into the balance of energy equation, they derived the following set of equations for the conservation of mass, the balance of

* Received by the editors July 10, 1992; accepted for publication (in revised form) April 6, 1993.
† Mathematics Department, West Virginia University, Morgantown, West Virginia 26506. The work of this author was supported in part by Army grant DAAL 03-89-G-0088.
‡ Mathematics Department, West Virginia University, Morgantown, West Virginia, 26506. The work of this author was supported in part by Office of Naval Research grant N00014-91-J-1291.

linear momentum, the balance of energy, and the Clausius–Duhem inequality:

$$\text{(1.2)}\qquad \begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ \rho \dfrac{D\mathbf{u}}{Dt} = \nabla \cdot \mathbf{T}, \\ \rho \dfrac{D\varepsilon}{Dt} = \mathbf{T} \cdot \mathbf{L} - \nabla \cdot \mathbf{q} + \nabla \cdot \mathbf{w}, \\ \rho\theta \dfrac{D\eta}{Dt} + \nabla \cdot \mathbf{q} - \dfrac{\mathbf{q} \cdot \nabla\theta}{\theta} \geq 0, \end{cases}$$

where $Df/Dt = f_t + \mathbf{u} \cdot \nabla f$ and we have the following:

- $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ is the velocity of fluid;
- $\theta = \theta(\mathbf{x}, t)(> 0)$ is the absolute temperature;
- $\varepsilon = \varepsilon(\mathbf{x}, t)$ is the specific internal energy per unit mass;
- $\eta = \eta(\mathbf{x}, t)$ is the specific entropy per unit mass;
- $\mathbf{T} = \mathbf{T}(\mathbf{x}, t)$ is the Cauchy stress tensor;
- $\mathbf{q} = \mathbf{q}(\mathbf{x}, t)$ is the heat flux vector;
- $\mathbf{L} = \nabla \mathbf{u}$.

Remark 1.1. In Dunn and Serrin "∇" denotes differentiation with respect to the particle and "grad" denotes differentiation with respect to the point. We do not make such a distinction in this paper. They both mean the differentiation with respect to the point.

In terms of the Helmholtz free energy $\psi = \varepsilon - \theta\eta$, the last inequality in (1.2) can be rewritten as

$$\text{(1.3)}\qquad \rho\left(\frac{D\psi}{Dt} + \eta\frac{D\theta}{Dt}\right) - \mathbf{T} \cdot \mathbf{L} - \nabla \cdot \mathbf{w} + \frac{\mathbf{q} \cdot \mathbf{g}}{\theta} \leq 0.$$

They consider the materials of Korteweg type in which the constitutive relations are given by

$$\text{(1.4)}\qquad \begin{aligned} \varepsilon &= \varepsilon(\rho, \theta, \mathbf{d}, \mathbf{S}, \mathbf{g}, \mathbf{L}), \\ \eta &= \eta(\rho, \theta, \mathbf{d}, \mathbf{S}, \mathbf{g}, \mathbf{L}), \\ \mathbf{T} &= \mathbf{T}(\rho, \theta, \mathbf{d}, \mathbf{S}, \mathbf{g}, \mathbf{L}), \\ \mathbf{q} &= \mathbf{q}(\rho, \theta, \mathbf{d}, \mathbf{S}, \mathbf{g}, \mathbf{L}), \\ \mathbf{w} &= \mathbf{w}(\rho, \theta, \mathbf{d}, \mathbf{S}, \mathbf{g}, \mathbf{L}), \\ \psi &= \psi(\rho, \theta, \mathbf{d}, \mathbf{S}, \mathbf{g}, \mathbf{L}), \end{aligned}$$

where $\mathbf{d} = \nabla\rho$, $\mathbf{S} = \mathbf{S}^T = \nabla^2\rho$, and $\mathbf{g} = \nabla\theta$. Using the Clausius–Duhem inequality (1.3), it is shown that $\mathbf{S}$, $\mathbf{g}$, and $\mathbf{L}$ drop out of ε, η, and ψ and indeed that

$$\text{(1.5)}\qquad \begin{aligned} \psi &= \psi(\rho, \theta, \mathbf{d}), \\ \eta &= -\psi_\theta(\rho, \theta, \mathbf{d}), \\ \varepsilon &= \psi(\rho, \theta, \mathbf{d}) - \theta\psi_\theta(\rho, \theta, \mathbf{d}). \end{aligned}$$

Furthermore, they have proved that for a given Helmholtz free energy $\psi(\rho, \theta, \mathbf{d})$ the following forms of $\mathbf{w}$ and $\mathbf{T}$,

$$\text{(1.6)}\qquad \begin{aligned} \mathbf{w} &= \rho\dot{\rho}\psi_{\mathbf{d}} + \bar{\mathbf{w}}, \\ \mathbf{T} &= (-\rho^2\psi_\rho + \rho\mathbf{d} \cdot \psi_{\mathbf{d}} + \rho^2\nabla \cdot \psi_{\mathbf{d}})\mathbf{I} - \rho\mathbf{d} \otimes \psi_{\mathbf{d}}, \end{aligned}$$

are compatible with (1.3). Here $\rho^2\psi_\rho(\rho,\theta,0)$ is the pressure and $\bar{\mathbf{w}}$ is the "static" portion of the interstitial work flux $\mathbf{w}$. They have shown that if the material possesses a center of symmetry, $\bar{\mathbf{w}}=0$. In what follows, we consider the materials that possess the center of symmetry. They also have observed that the classical forms of viscosity and conductivity tensors are compatible.

In this paper we consider the existence of a unique, local, smooth solution in the two-dimensional isothermal motion of the Korteweg type materials, where the viscous effect is also included. The three-dimensional case can be discussed similarly. The restriction to isothermal motions is a special one, and it would be important to remove it in future work. In what follows, we state the assumptions on the Helmholtz free energy and derive the system that we shall discuss. We assume that the Helmholtz free energy is given by

$$\psi = F(\rho) + \frac{\nu}{2\rho}(\rho_x^2+\rho_y^2), \tag{1.7}$$

where F is a smooth function of ρ and ν is a positive constant. This choice is to make the terms appearing in (1.2) as simple as possible, yet reflect the effect of the higher-order terms of ρ.[1]

With the choice of the Helmholtz free energy given in (1.7) and with $\lambda=-\frac{1}{3}\mu$, the system then becomes

$$\begin{cases} \rho_t+(\rho u)_x+(\rho v)_y=0, \\ (\rho u)_t+(\rho u^2)_x+(\rho uv)_y=(T_{11})_x+(T_{12})_y, \\ (\rho v)_t+(\rho uv)_x+(\rho v^2)_y=(T_{21})_x+(T_{22})_y, \end{cases} \tag{1.8}$$

where u and v are the x and y component of the velocity and

$$\begin{aligned} \mathbf{T} &= \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \\ &= \Big\{-p+\frac{\nu}{2}(\rho_x^2+\rho_y^2)+\nu\rho\triangle\rho\Big\}\mathbf{I}-\nu\begin{pmatrix} \rho_x^2 & \rho_x\rho_y \\ \rho_x\rho_y & \rho_y^2 \end{pmatrix}+\mathbf{V}, \end{aligned} \tag{1.9}$$

$$p=\rho^2F'(\rho), \tag{1.10}$$

$$\mathbf{V}=\begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}=\mu\{(\nabla\mathbf{u})+(\nabla\mathbf{u})^T-\tfrac{2}{3}(\nabla\mathbf{u})\mathbf{I}\}. \tag{1.11}$$

Here $\mathbf{I}$ is the unit rank-two tensor, and superscript T denotes the transpose of a tensor. Since we discuss the existence of a local solution, we do not need the monotonicity of the pressure on ρ. Further computation simplifies the $\nabla\cdot\mathbf{T}$ term

$$\nabla\cdot\mathbf{T}=-\nabla p+\nu\rho\nabla(\triangle\rho)+\nabla\cdot\mathbf{V}. \tag{1.12}$$

In this paper we discuss the local existence for the initial value problem of (1.8) with the initial data given by

$$(\rho,u,v)(x,y,0)=(\rho_0,u_0,v_0)(x,y). \tag{1.13}$$

[1] Another reasonable choice is to change the last term in (1.7) with $(\nu/2)(\rho_x^2+\rho_y^2)$. Although this choice may be physically more realistic, mathematically it is more cumbersome to handle. For example, the expression for $\nabla\cdot\mathbf{T}$ becomes more complicated; therefore, we do not discuss this case.

We assume that the initial data satisfy

$$(\rho_0 - \bar{\rho_0}, u_0, v_0) \in H^k(R^2), \qquad \bar{\rho_0} \geq \delta > 0,$$

where $k \geq 4$ and $\bar{\rho_0} > 0$ is a positive constant. Denote by $\|\cdot\| \equiv \|\cdot\|_0$ the L^2 norm and by $\|\cdot\|_k$ the kth order Sobolev norm. Set

$$|||w|||_{0,T}^2 \equiv \sup_{0\leq t\leq T} \left(\|w(t)\|^2 + \|\nabla\rho(t)\|^2\right) + \int_0^T \left(\|\nabla u(t)\|^2 + \|\nabla v(t)\|^2\right) dt \tag{1.14}$$

and

$$|||w|||_k^2 = \sum_{|j|\leq k} |||\partial_{x,y}^j w|||^2,$$

where $w \equiv (\rho, u, v)$. Then the main theorem of this paper can be stated as follows.

THEOREM 1.1. *For any initial data (ρ_0, u_0, v_0) such that the condition $\rho_0 \geq \delta > 0$ is satisfied and $(\rho_0 - \bar{\rho_0}, u_0, v_0) \in H^k(R^2)$ $(k \geq 4)$, where $\bar{\rho_0} > 0$ is a constant, there exists a $T > 0$ such that in $t \in [0, T]$, the Cauchy problem (1.8) and (1.13) has a unique solution (ρ, u, v) such that $\rho - \bar{\rho_0} \in L^\infty([0,T]; H^{k+1}(R^2))$, $(u, v) \in L^\infty([0,T]; H^k(R^2))$, and*

$$|||w|||_k^2 \leq C_k \|w_0\|_k^2 + \|\rho_0\|_{k+1}^2. \tag{1.15}$$

Since the linearized problem of (1.8) and (1.13) is not of any classical type, the existence of solutions is not known even for the linearized problem. We prove the existence of solutions for the linearized problem by establishing an energy estimate for the dual problem and then using the dual argument.

For one-dimensional problems, the effects of the higher-order derivatives of density (or in the context of elasticity, the higher-order derivatives of strain) have been discussed extensively in phase transition problems where the pressure or the stress is a nonmonotone function of the density or strain. In compressible fluids, Serrin [11], [12] reconsidered the Korteweg theory and has shown the existence of steady profile connecting two different phases. In [13] and [14], Slemrod considered the existence of travelling wave solutions connecting two different phases. In elasticity, Andrews and Ball [1] discussed the existence and the asymptotic behavior of solutions in the hard loading case. Concerning the dynamical aspects of the soft loading case, Hattori and Mischaikow [4] proved the existence of global solutions and a global compact attractor in H^1, examined the dynamic stability and the bifurcations of stationary solutions, and demonstrated the connecting orbit problems in the semiflow. Sprekels and Zheng [15] discussed the existence of solutions to the equations of a Ginzburg–Landau theory for structural phase transition in shape memory alloys. The system that they discussed was derived by Falk [3]. It is interesting to note that his system has the same term as the interstitial working term, although he derived it independently. This motivates our study of multidimensional problems.

In the one space-dimensional case, the problem is solved by introducing Lagrangian coordinates, and the system (1.8) reduces to scalar equations for the velocity u of higher order. However, in the higher-dimensional space, this kind of reduction is not available. The appearance of higher-order derivatives of ρ in the momentum equations of (1.8) makes the system unsymmetric. Therefore, to obtain the a priori

estimate for the linearized problem, we have repeatedly used the first equation in (1.8).

This paper consists of four sections. The second section discusses the linearization of (1.8), (1.13), and the a priori estimate for smooth solutions for the linearized problem. Section 3 establishes the existence of solutions for the linearized problem and in §4 we discuss the existence of local solutions for (1.8) and (1.13) and complete the proof of Theorem 1.1.

2. Linearized problem and a priori estimate. The linearized equations for the perturbation $\dot{w} \equiv (\dot{\rho}, \dot{u}, \dot{v})$ of (1.8) at a given solution $w \equiv (\rho, u, v)$ can be written as follows:

(2.1)
$$\begin{cases} (\partial_t + u\partial_x + v\partial_y)\dot{\rho} + \rho(\dot{u}_x + \dot{v}_y) + \ell_1(\dot{w}) = \dot{f}_1, \\ \rho(\partial_t + u\partial_x + v\partial_y)\dot{u} + p'(\rho)\dot{\rho}_x - \nu\rho\triangle\dot{\rho}_x - \mu\triangle\dot{u} - \frac{\mu}{3}(\dot{u}_{xx} + \dot{v}_{xy}) + \ell_2(\dot{w}) = \dot{f}_2, \\ \rho(\partial_t + u\partial_x + v\partial_y)\dot{v} + p'(\rho)\dot{\rho}_y - \nu\rho\triangle\dot{\rho}_y - \mu\triangle\dot{v} - \frac{\mu}{3}(\dot{u}_{xy} + \dot{v}_{yy}) + \ell_3(\dot{w}) = \dot{f}_3. \end{cases}$$

Here $\ell_j(\cdot)$ $(j = 1, 2, 3)$ denote the linear functions of the arguments with the coefficients of ℓ_1 depending upon ∇w and the coefficients of ℓ_2, ℓ_3 depending upon ∇w and $\nabla^2(u, v)$.

Consider the Cauchy problem of (2.1) with initial data:

$$\dot{\rho}(x, y, 0) = \dot{\rho}_0(x, y), \quad \dot{u}(x, y, 0) = \dot{u}_0(x, y), \quad \dot{v}(x, y, 0) = \dot{v}_0(x, y). \tag{2.2}$$

Equations (2.1) can be rewritten in the following matrix form:

$$\dot{L}(w)\dot{w} \equiv \begin{pmatrix} \dot{L}_1(w)\dot{w} \\ \dot{L}_2(w)\dot{w} \\ \dot{L}_3(w)\dot{w} \end{pmatrix} \equiv \partial_t\dot{w} + A_1\partial_x\dot{w} + A_2\partial_y\dot{w} + (T_1 + T_2)\dot{w} + \ell(\dot{w}) = \dot{f}, \tag{2.3}$$

where A_1, A_2 are coefficient matrices of the first-order space derivative terms and $T_1 + T_2$ is an operator matrix involving derivatives of order two or higher:

$$A_1 = \begin{pmatrix} u & \rho & 0 \\ p'(\rho)\rho^{-1} & u & 0 \\ 0 & 0 & u \end{pmatrix}, \qquad A_2 = \begin{pmatrix} v & 0 & \rho \\ 0 & v & 0 \\ p'(\rho)\rho^{-1} & 0 & v \end{pmatrix}, \tag{2.4}$$

$$T_1\dot{w} = -\nu\triangle \begin{pmatrix} 0 \\ \nabla\dot{\rho} \end{pmatrix}, \qquad T_2\dot{w} = -\mu\rho^{-1} \begin{pmatrix} 0 \\ \triangle\dot{u} + \frac{1}{3}(\dot{u}_{xx} + \dot{v}_{xy}) \\ \triangle\dot{v} + \frac{1}{3}(\dot{u}_{xy} + \dot{v}_{yy}) \end{pmatrix}. \tag{2.5}$$

Let $\langle \cdot, \cdot \rangle$ denote the L^2 inner product in the $(x, y) \in R^2$. Denote $\|\cdot\| \equiv \|\cdot\|_0$, the corresponding norm, and $\|\cdot\|_k$, the kth order Sobolev norm.

Let β_0 be a constant such that the variables $w = (\rho, u, v)$ in the coefficients of (2.3) satisfy

$$\sup_{t,x,y} \left(\rho^{-1} + |\rho_t| + \sum_{|j| \le 2} |D^j_{x,y} w| \right) \le \beta_0. \tag{2.6}$$

Let C_0 denote the constant that depends only upon β_0. Then we have the following zero-order energy estimate.

THEOREM 2.1. *The smooth solution* $\dot{w} \in C_0^\infty([0,T] \times R^2)$ *of* (2.1) *and* (2.2) *satisfies the estimate*

$$\partial_t(\|\dot{w}\|^2 + \|\nabla\dot{\rho}\|^2) + \|\nabla\dot{u}\|^2 + \|\nabla\dot{v}\|^2 \le C_0(\|\dot{w}\|^2 + \|\nabla\dot{\rho}\|^2 + \|\dot{f}\|_0^2 + \|\dot{f}_1\|_1^2) \tag{2.7}$$

and

$$|||\dot{w}|||_{0,T}^2 \le C_0(T)\left(\|\dot{w}_0\|^2 + \|\dot{\rho}_0\|_1^2 + \int_0^T (\|\dot{f}\|^2 + \|\dot{f}_1\|_1^2)dt\right). \tag{2.8}$$

Here the norm $|||\dot{w}|||_{0,T}$ *is defined as follows*:

$$|||\dot{w}|||_{0,T}^2 \equiv \sup_{0\le t\le T}\left(\|\dot{w}(t)\|^2 + \|\nabla\dot{\rho}(t)\|^2\right) + \int_0^T \left(\|\nabla\dot{u}(t)\|^2 + \|\nabla\dot{v}(t)\|^2\right)dt. \tag{2.9}$$

Proof. Define the diagonal matrix

$$A_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \rho & 0 \\ 0 & 0 & \rho \end{pmatrix}.$$

Taking L^2 inner products of (2.3) with vector $A_0\dot{w}$ gives

$$\langle \partial_t\dot{w}, A_0\dot{w}\rangle + \langle (T_1+T_2)\dot{w}, A_0\dot{w}\rangle \le C_0(\|\dot{f}\| + \|\dot{w}\| + \|\nabla\dot{w}\|)\|\dot{w}\|. \tag{2.10}$$

Obviously,

$$\langle \partial_t\dot{w}, A_0\dot{w}\rangle \ge \tfrac{1}{2}\partial_t\|\sqrt{A_0}\dot{w}\|^2 - C_0\|\dot{w}\|^2. \tag{2.11}$$

From the first equation in (2.1),

$$\rho(\dot{u}_x + \dot{v}_y) = -(\partial_t + u\partial_x + v\partial_y)\dot{\rho} - \ell_1(\dot{w}) + \dot{f}_1,$$

and standard integration by parts gives

$$\begin{aligned} \langle T_1\dot{w}, A_0\dot{w}\rangle &= \langle \nu\triangle\dot{\rho}, \rho(\dot{u}_x + \dot{v}_y) + \nabla\rho\cdot(\dot{u},\dot{v})^t\rangle \\ &= \langle \nu\triangle\dot{\rho}, -(\partial_t + u\partial_x + v\partial_y)\dot{\rho} - \ell_1(\dot{w}) + \dot{f}_1 + \nabla\rho\cdot(\dot{u},\dot{v})^t\rangle \\ &\ge \nu\langle\nabla\dot{\rho}, \partial_t\nabla\dot{\rho}\rangle - C_0\|\nabla\dot{\rho}\|(\|\dot{w}\| + \|\nabla\dot{w}\| + \|\dot{f}_1\|_1) \\ &\ge \tfrac{\nu}{2}\partial_t\|\nabla\dot{\rho}\|^2 - C_0\epsilon^{-1}\|\nabla\dot{\rho}\|^2 - \epsilon\|\nabla\dot{w}\|^2 - C_0(\|\dot{w}\|^2 + \|\dot{f}_1\|_1^2) \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} \langle T_2\dot{w}, A_0\dot{w}\rangle &= -\mu\langle\triangle\dot{u} + \tfrac{1}{3}(\dot{u}_{xx} + \dot{v}_{xy}), \dot{u}\rangle - \mu\langle\triangle\dot{v} + \tfrac{1}{3}(\dot{u}_{xy} + \dot{v}_{yy}), \dot{v}\rangle \\ &\ge \mu(\|\nabla\dot{u}\|^2 + \|\nabla\dot{v}\|^2) - C_0\|\nabla\dot{w}\|\|\dot{w}\|. \end{aligned} \tag{2.13}$$

Combining (2.10)–(2.13) and noticing that $\|A_0\dot{w}\| \sim \|\dot{w}\|$, $\|A_0\nabla\dot{w}\| \sim \|\nabla\dot{w}\|$ by (2.6), we obtain the estimate (2.7) by taking $\epsilon \ll 1$.

By the Gronwall theorem, we have from (2.7) that

$$\|\dot{w}(t)\|^2 + \|\nabla\dot{\rho}(t)\|^2 \le e^{C_0t}\left(\|\dot{w}_0\|^2 + \|\nabla\dot{\rho}_0\|^2\right) + \int_0^t e^{C_0(t-s)}(\|\dot{f}(s)\|^2 + \|\dot{f}_1(s)\|_1^2)ds. \tag{2.14}$$

Replacing the terms $\|\dot{w}(t)\|^2 + \|\nabla\dot{\rho}(t)\|^2$ on the right of (2.7) and integrating in t from 0 to T, we obtain (2.8).

To derive higher-order estimates, we denote $\|\cdot\|_k$ the kth-order Sobolev norm in $(x, y) \in R^2$ and

$$\|\dot{w}\|_k^2 \equiv \sum_{|j|\le k} \|\partial_{x,y}^j \dot{w}\|^2.$$

Let β_k be a constant such that the variables $w = (\rho, u, v)$ in the coefficients of (2.3) satisfy

$$\sup_{t,x,y} \left(\rho^{-1} + \sum_{|j|\le 2} |D_{x,y}^j w| + |\rho_t| \right) + \|w\|_k^2 \le \beta_k. \tag{2.15}$$

Let C_k denote the constant that depends only upon β_k. Then we have the following kth order energy estimate.

THEOREM 2.2. *For integer $k \ge 4$, the smooth solution $\dot{w} \in C_0^\infty([0,T] \times R^2)$ of (2.1) and (2.2) satisfies the estimate*

$$\begin{aligned} &\partial_t(\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2) + \|\dot{u}\|_{k+1}^2 + \|\dot{v}\|_{k+1}^2 \\ &\quad \le C_k \left(\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2 + \|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1} \right) \end{aligned} \tag{2.16}$$

and

$$\|\dot{w}\|_k^2 \le C_k(T) \left(\|\dot{w}_0\|_k^2 + \|\dot{\rho}_0\|_{k+1}^2 \right) + C_k(T) \int_0^T (\|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1}^2) dt. \tag{2.17}$$

Proof. Applying the operator ∇^j to (2.3), we have

$$\dot{L}(w)\nabla^j \dot{w} = \nabla^j \dot{f} - [\nabla^j, \dot{L}]\dot{w}. \tag{2.18}$$

Taking the L^2 inner product of (2.18) with $A_0 \nabla^j \dot{w}$, we obtain

$$\begin{aligned} &\partial_t \|\nabla^j \dot{w}\|^2 + \|\nabla^{j+1}\dot{u}\|^2 + \|\nabla^{j+1}\dot{v}\|^2 + \langle A_0 \nabla^j \dot{w}, T_1 \nabla^j \dot{w} \rangle \\ &\quad \le C_j \left(\|\dot{w}\|_j^2 + \|\dot{\rho}\|_{j+1}^2 + \|\dot{f}\|_{j-1}^2 + |\langle A_0 \nabla^j \dot{w}, [\nabla^j, \dot{L}]\dot{w} \rangle| \right). \end{aligned} \tag{2.19}$$

First consider

$$\langle A_0 \nabla^j \dot{w}, T_1 \nabla^j \dot{w} \rangle = \nu \langle \triangle \nabla^j \dot{\rho}, \rho \nabla^j (\dot{u}_x + \dot{v}_y) \rangle + \nu \langle \triangle \nabla^j \dot{\rho}, \nabla \rho \nabla^j (\dot{u}, \dot{v})^t \rangle.$$

From the first equation in (2.18), we have

$$\rho \nabla^j (\dot{u}_x + \dot{v}_y) = -(\partial_t + u\partial_x + v\partial_y)\nabla^j \dot{\rho} - \ell_1(\nabla^j \dot{w}) - \nabla^j \dot{f}_1 - [\nabla^j, \dot{L}_1(w)]\dot{w}.$$

Noticing that $[\nabla^j, \dot{L}_1(w)]$ is an operator of order j, hence

$$\begin{aligned} \langle A_0 \nabla^j \dot{w}, T_1 \nabla^j \dot{w} \rangle \ge &-\nu \langle \triangle \nabla^j \dot{\rho}, \rho \nabla^j \dot{\rho}_t \rangle \\ &- \epsilon \left(\|\nabla^{j+1}\dot{u}\|^2 + \|\nabla^{j+1}\dot{v}\|^2 \right) - C_j \left(\|\nabla^{j+1}\dot{\rho}\|^2 + \|\dot{f}_1\|_{j+1}^2 \right). \end{aligned}$$

Since

$$-\langle \triangle \nabla^j \dot{\rho}, \rho \nabla^j \dot{\rho}_t \rangle \geq \tfrac{1}{2}\partial_t \|\nabla^{j+1}\dot{\rho}\|^2 - C_j \|\nabla^{j+1}\dot{\rho}\| \|\nabla^j \dot{\rho}_t\|$$
$$\geq \tfrac{1}{2}\partial_t \|\nabla^{j+1}\dot{\rho}\|^2 - C_j(\|\nabla^{j+1}\dot{\rho}\|^2 + \|\dot{f}_1\|_j^2) - \epsilon\left(\|\nabla^{j+1}\dot{u}\| + \|\nabla^{j+1}\dot{u}\|\right)$$

by the first equation in (2.3), we have

(2.20)
$$\langle A_0 \nabla^j \dot{w}, T_1 \nabla^j \dot{w} \rangle \geq \tfrac{1}{2}\partial_t \|\nabla^{j+1}\dot{\rho}\|^2$$
$$- \epsilon\left(\|\nabla^{j+1}\dot{u}\|^2 + \|\nabla^{j+1}\dot{v}\|^2\right) - C_j(\|\nabla^{j+1}\dot{\rho}\|^2 + \|\dot{f}_1\|_{j+1}^2).$$

To discuss the terms involving the commutator $[\nabla^j, \dot{L}]$, we notice the structure of the operator $\dot{L}$ in (2.3) and the terms in $[\nabla^j, \dot{L}]\dot{w}$ having the form

$$\nabla^{j-i} F(w) \nabla^{2+i} \dot{w}, \qquad 0 \leq i \leq j-1,$$

where F denotes smooth functions of its argument. According to Nirenberg's inequality [7], [9], we have

$$\|\nabla^{j-i} F(w) \nabla^{2+i} \dot{w}\|_0 \leq C\left(|\nabla F(w)|_{C^0} \|\nabla^2 \dot{w}\|_{j-1} + \|\nabla F(w)\|_{j-1} |\nabla^2 \dot{w}|_{C^0}\right), \tag{2.21}$$

where $|\cdot|_{C^0}$ denotes the usual maximum norm of continuous functions. Consequently,

$$|\langle A_0 \nabla^j \dot{w}, [\nabla^j, \dot{L}]\dot{w} \rangle| \leq C_0 \|\nabla^j \dot{w}\| \|\dot{w}\|_{j+1} + C_j |\nabla^2 \dot{w}|_{C^0} \|\nabla^j \dot{w}\|. \tag{2.22}$$

Replacing $|\langle A_0 \nabla^j \dot{w}, [\nabla^j, \dot{L}]\dot{w} \rangle|$ in (2.19) by (2.22) and summing up for all $j \leq k$, we have

(2.23)
$$\partial_t(\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2) + \|\dot{u}\|_{k+1}^2 + \|\dot{v}\|_{k+1}^2$$
$$\leq C_k\left(\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2 + \|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1} + |\nabla^2 \dot{w}|_{C^0}^2\right).$$

For $k \geq 4$, $H^k(R^2) \subset C^2(R^2)$; therefore, we obtain (2.16).

Applying the Gronwall theorem to (2.16), we find

(2.24)
$$\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2$$
$$\leq C_k\left(\|\dot{w}_0\|_k^2 + \|\dot{\rho}_0\|_{k+1}^2 + \|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1}^2 + \int_0^t e^{C(t-s)}(\|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1}^2) ds\right).$$

Replacing $\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2$ in (2.16) and integrating in t over $[0,T]$, we obtain (2.17). This concludes the proof of Theorem 2.2.

Remark 2.1. From (2.24) and (2.3), it is easy to obtain the corresponding estimates for $\partial_t^j \dot{w}$.

3. Existence of solution for linearized problem. By the energy estimate (2.17) and the continuation method, in order to prove the existence of the solution for the problem (2.1) and (2.2), we need only to show the existence of a solution for $f \in C_0^\infty([0,T] \times R^2)$, $\dot{w}_0 = 0$ with the lower-order terms in (2.1) omitted.

In the following, we use the dual method [6], [10] to prove the existence of the following problem in $[0,T]$:

$$L\dot{w} \equiv \partial_t \dot{w} + B_1' \partial_x \dot{w} + B_2' \partial_y \dot{w} + (T_1 + T_2)\dot{w} = \dot{f}, \tag{3.1}$$

$$\dot{w}(x,y,0)=0. \tag{3.2}$$

In (3.1), the operators T_1, T_2 are defined in (2.5), and the matrices B_1, B_2 are obtained by omitting the lower-order terms of A_1, A_2 in (2.4):

$$B_1=\begin{pmatrix} u & \rho & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}, \qquad B_2=\begin{pmatrix} v & 0 & \rho\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}. \tag{3.3}$$

The adjoint operator L^* for (5.1) is defined by

$$\langle L\dot{w},\dot{\phi}\rangle=\langle \dot{w},L^*\dot{\phi}\rangle.$$

To prove the existence of weak solution $\dot{w}\in L^2([0,T],H^{k+1})$ for (3.1) and (3.2), we need to establish the energy estimates of negative order for the operator L^*. Since the operator L is not symmetric, we first derive the classical energy estimate for

$$L^*\dot{\phi}=-\partial_t\dot{\phi}-B_1^*\partial_x\dot{\phi}-B_2^*\partial_y\dot{\phi}+(T_1^*+T_2^*)\dot{\phi}=\dot{g}, \tag{3.4}$$

$$\dot{\phi}(x,y,T)=0. \tag{3.5}$$

THEOREM 3.1. *The solutions of* (3.4) *and* (3.5) *satisfy the following estimate*:

$$\|\dot{\phi}(t)\|^2+\|\dot{\phi}_{2,3}(t)\|_2^2\le C\left(\|\dot{g}(t)\|^2+\int_0^T\|\dot{g}(\tau)\|^2d\tau\right). \tag{3.6}$$

Proof. Explicitly, the operator L^* can be written as follows:

$$\begin{cases} -(\partial_t+\partial_x u+\partial_y v)\dot{\phi}_1+\nu\triangle(\partial_x\dot{\phi}_2+\partial_y\dot{\phi}_3)=\dot{g}_1,\\ -\partial_t\dot{\phi}_2-\partial_x(\rho\dot{\phi}_1)-(\mu\triangle+\frac{\mu}{3}\partial_{xx})(\rho^{-1}\dot{\phi}_2)-\frac{\mu}{3}\partial_{xy}(\rho^{-1}\dot{\phi}_3)=\dot{g}_2,\\ -\partial_t\dot{\phi}_3-\partial_y(\rho\dot{\phi}_2)-(\mu\triangle+\frac{\mu}{3}\partial_{xx})(\rho^{-1}\dot{\phi}_3)-\frac{\mu}{3}\partial_{xy}(\rho^{-1}\dot{\phi}_2)=\dot{g}_3. \end{cases} \tag{3.7}$$

Taking the inner product of the second and third equations in (3.7) with $\dot{\phi}_2,\dot{\phi}_3$ and integrating by parts, we have

$$-\partial_t\|\dot{\phi}_{2,3}\|^2+\mu\|\rho^{-\frac{1}{2}}\nabla\dot{\phi}_{2,3}\|^2\le C(\|\dot{\phi}\|^2+\|\dot{g}_{2,3}\|^2). \tag{3.8}$$

Then, taking the inner product of the second and third equations in (3.7) with $-\triangle\dot{\phi}_2,-\triangle\dot{\phi}_3$ and integrating by parts, we have

$$\begin{aligned} &-\tfrac{1}{2}\partial_t\|\nabla\dot{\phi}_{2,3}\|^2+\mu\|\rho^{-\frac{1}{2}}\triangle\dot{\phi}_{2,3}\|^2-\langle\rho\dot{\phi}_1,\triangle(\partial_x\dot{\phi}_2+\partial_y\dot{\phi}_3)\rangle\\ &\quad\le C(\|\dot{\phi}_{2,3}\|^2+\|\nabla\dot{\phi}_{2,3}\|^2+\|\dot{g}_{2,3}\|^2). \end{aligned} \tag{3.9}$$

From the first equation of (3.7),

$$\triangle(\partial_x\dot{\phi}_2+\partial_y\dot{\phi}_3)=\nu^{-1}[(\partial_t+\partial_x u+\partial_y v)\dot{\phi}_1+\dot{g}_1],$$

and

$$-\langle\rho\dot{\phi}_1,\partial_t\dot{\phi}_1\rangle\ge-\tfrac{1}{2}\partial_t\|\rho^{-\frac{1}{2}}\dot{\phi}_1\|^2-C\|\dot{\phi}_1\|^2,$$
$$|\langle\rho\dot{\phi}_1,(\partial_x u+\partial_y v)\dot{\phi}_1\rangle|\le C\|\dot{\phi}_1\|^2;$$

therefore, we obtain the following estimate from (3.9):

$$-\partial_t \left(\|\dot{\phi}_1\|^2 + \|\nabla \dot{\phi}_{2,3}\|^2 \right) + \|\rho^{-\frac{1}{2}} \triangle \dot{\phi}_{2,3}\|^2 \leq C \left(\|\dot{\phi}\|^2 + \|\nabla \dot{\phi}_{2,3}\|^2 + \|\dot{g}\|^2 \right). \tag{3.10}$$

Combining (3.8) and (3.10), we obtain

$$-\partial_t \left(\|\dot{\phi}\|^2 + \|\nabla \dot{\phi}_{2,3}\|^2 \right) + \|\rho^{-\frac{1}{2}} \triangle \dot{\phi}_{2,3}\|^2 \leq C \left(\|\dot{\phi}\|^2 + \|\nabla \dot{\phi}_{2,3}\|^2 + \|\dot{g}\|^2 \right). \tag{3.11}$$

Applying the Gronwall inequality and noticing $\dot{\phi} = 0$ at $t = T$, we get (3.6) in Theorem 3.1.

Next we derive the negative norm estimate for the solution of (3.3) and (3.4). Let Λ denote the operator with symbol

$$\lambda(\xi) = \sqrt{1 + |\xi|^2},$$

where $\xi = (\xi_1, \xi_2)$ is the dual variable of (x, y). Now we are going to establish the following estimate.

THEOREM 3.2. *For any $s \in R$, the solutions of* (3.4) *and* (3.5) *satisfy the following estimate*:

$$\|\Lambda^s \dot{\phi}(t)\|^2 + \|\Lambda^{s+2} \dot{\phi}_{2,3}(t)\|^2 \leq C \left(\|\Lambda^s \dot{g}(t)\|^2 + \int_0^T \|\Lambda^s \dot{g}(\tau)\|^2 d\tau \right). \tag{3.12}$$

Proof. Let $L^*_{2,3}$ denote the second and third components of the operator L^*. Consider the inner product

$$\langle \Lambda^s \dot{\phi}_{2,3}, \Lambda^s \dot{g}_{2,3} \rangle = \langle \Lambda^s \dot{\phi}_{2,3}, L^*_{2,3} \Lambda^s \dot{\phi} \rangle + \langle \Lambda^s \dot{\phi}_{2,3}, [\Lambda^s, L^*_{2,3}] \dot{\phi} \rangle.$$

Similar to (3.8), we have

$$\begin{aligned} &-\partial_t \|\Lambda^s \dot{\phi}_{2,3}\|^2 + \|\rho^{-\frac{1}{2}} \nabla \Lambda^s \dot{\phi}_{2,3}\|^2 \\ &\quad \leq C \left(\|\Lambda^s \dot{\phi}\|^2 + \|\Lambda^s \dot{g}_{2,3}\|^2 + |\langle \Lambda^s \dot{\phi}_{2,3}, [\Lambda^s, L^*_{2,3}] \dot{\phi} \rangle| \right). \end{aligned} \tag{3.13}$$

Since the commutator operator $[\Lambda^s, L^*_{2,3}]$ is of order $s + 1$,

$$|\langle \Lambda^s \dot{\phi}_{2,3}, [\Lambda^s, L^*_{2,3}] \dot{\phi} \rangle| \leq C \|\Lambda^{s+1} \dot{\phi}_{2,3}\| \|\Lambda^s \dot{\phi}\| \leq \epsilon \|\Lambda^{s+1} \dot{\phi}_{2,3}\|^2 + C(\epsilon) \|\Lambda^s \dot{\phi}\|^2.$$

Therefore, we obtain the following from (3.13):

$$-\partial_t \|\Lambda^s \dot{\phi}_{2,3}\|^2 + \|\rho^{-\frac{1}{2}} \nabla \Lambda^s \dot{\phi}_{2,3}\|^2 \leq C \left(\|\Lambda^s \dot{\phi}\|^2 + \|\Lambda^s \dot{g}_{2,3}\|^2 \right). \tag{3.14}$$

Then consider the inner product

$$-\langle \triangle \Lambda^s \dot{\phi}_{2,3}, \Lambda^s \dot{g}_{2,3} \rangle = -\langle \triangle \Lambda^s \dot{\phi}_{2,3}, L^*_{2,3} \Lambda^s \dot{\phi} \rangle - \langle \triangle \Lambda^s \dot{\phi}_{2,3}, [\Lambda^s, L^*_{2,3}] \dot{\phi} \rangle.$$

Similar to (3.9), we have

$$\begin{aligned} &-\tfrac{1}{2} \partial_t \|\nabla \Lambda^s \dot{\phi}_{2,3}\|^2 + \mu \|\rho^{-\frac{1}{2}} \triangle \Lambda^s \dot{\phi}_{2,3}\|^2 - \langle \rho \Lambda^s \dot{\phi}_1, \triangle (\partial_x \Lambda^s \dot{\phi}_2 + \partial_y \Lambda^s \dot{\phi}_3) \rangle \\ &\quad \leq C \left(\|\Lambda^s \dot{\phi}_{2,3}\|^2 + \|\nabla \Lambda^s \dot{\phi}_{2,3}\|^2 + \|\Lambda^s \dot{g}_{2,3}\|^2 + |\langle \triangle \Lambda^s \dot{\phi}_{2,3}, [\Lambda^s, L^*_{2,3}] \dot{\phi} \rangle| \right). \end{aligned} \tag{3.15}$$

From the first equation in (3.7),

$$\triangle\Lambda^s(\partial_x\dot\phi_2+\partial_y\dot\phi_3)=\nu^{-1}\Lambda^s[(\partial_t+\partial_x u+\partial_y v)\dot\phi_1+\Lambda^s\dot g_1],$$

and

$$-\langle\rho\Lambda^s\dot\phi_1,\Lambda^s[(\partial_t+\partial_x u+\partial_y v)\dot\phi_1]\rangle\geq-\tfrac12\partial_t\|\sqrt\rho\Lambda^s\dot\phi_1\|^2-C\|\Lambda^s\dot\phi_1\|^2,$$

hence

(3.16)
$$-\langle\rho\Lambda^s\dot\phi_1,\triangle(\partial_x\Lambda^s\dot\phi_2+\partial_y\Lambda^s\dot\phi_3)\rangle\geq-\frac{1}{2\nu}\partial_t\|\sqrt\rho\Lambda^s\dot\phi_1\|^2-C\left(\|\Lambda^s\dot\phi_1\|^2+\|\Lambda^s\dot g_1\|^2\right).$$

Because $[\Lambda^s,L^*_{2,3}]$ is an operator of order $s+1$ with respect to $\dot\phi_{2,3}$ and an operator of order s with respect to $\dot\phi_1$,

$$|\langle\triangle\Lambda^s\dot\phi_{2,3},[\Lambda^s,L^*_{2,3}]\dot\phi\rangle|\leq\epsilon\|\triangle\Lambda^s\dot\phi_{2,3}\|^2+C(\epsilon)\left(\|\Lambda^{s+1}\dot\phi_{2,3}\|^2+\|\Lambda^s\dot\phi\|^2\right). \tag{3.17}$$

From (3.15)–(3.17), we have

$$\begin{aligned}&-\partial_t\left(\|\nabla\Lambda^s\dot\phi_{2,3}\|^2+\partial_t\|\sqrt\rho\Lambda^s\dot\phi_1\|^2\right)+\|\rho^{-\frac12}\triangle\Lambda^s\dot\phi_{2,3}\|^2\\&\qquad\leq C\left(\|\Lambda^s\dot\phi\|^2+\|\nabla\Lambda^s\dot\phi_{2,3}\|^2+\|\Lambda^s\dot g\|^2\right).\end{aligned} \tag{3.18}$$

Combining (3.14) and (3.18), we obtain

$$\begin{aligned}&-\partial_t\left(\|\Lambda^s\dot\phi\|^2+\|\nabla\Lambda^s\dot\phi_{2,3}\|^2\right)+\|\rho^{-\frac12}\triangle\Lambda^s\dot\phi_{2,3}\|^2\\&\qquad\leq C_s\left(\|\Lambda^s\dot\phi\|^2+\|\nabla\Lambda^s\dot\phi_{2,3}\|^2+\|\Lambda^s\dot g\|^2\right).\end{aligned} \tag{3.19}$$

Applying the Gronwall inequality to (3.19) concludes the proof of the Theorem 3.2.

From Theorem 3.2, it is standard to derive the existence of a differentiable weak solution for (3.1) and (3.2). Since, for any large integer k,

$$\left|\int_0^T\langle\dot f,\dot\phi\rangle dt\right|=\left|\int_0^T\langle\Lambda^k\dot f,\Lambda^{-k}\dot\phi\rangle dt\right|\leq\int_0^T\|\Lambda^k\dot f\|\|\Lambda^{-k}\dot\phi\|dt.$$

Applying (3.12) for $s=-k$, we have

$$\left|\int_0^T\langle\dot f,\dot\phi\rangle dt\right|\leq C\int_0^T\|\Lambda^k\dot f\|dt\int_0^T\|\Lambda^{-k}L^*\dot\phi\|dt.$$

Therefore,

$$\int_0^T\langle\dot f,\dot\phi\rangle dt$$

defines a bounded linear functional of $L^*\dot\phi$ in the space $L^2([0,T],H^{-k}(R^2))$. By the Hahn–Banach extension theorem and the Riesz representation theorem, we obtain a weak solution $w\in L^2([0,T],H^k(R^2))$ such that

$$\int_0^T\langle\dot f,\dot\phi\rangle dt=\int_0^T\langle w,L^*\dot\phi\rangle dt\quad\forall\dot\phi\in C_0^\infty([0,T)\times R^2). \tag{3.20}$$

Since differentiable weak solutions satisfy the equation in the classical sense, we have proved the following existence theorem.

THEOREM 3.3. *For all* $\dot{f} \in C([0,T], H^k(R^2))$, $\dot{f}_1 \in C([0,T], H^{k+1}(R^2))$, $\dot{w}_0 \in H^k(R^2)$ *and* $\dot{\rho}_0 \in H^{k+1}(R^2)$, *the Cauchy problem* (2.1) *and* (2.2) *has a unique solution* $\dot{w}$ *such that its norm* $|||\dot{w}|||$ *is bounded and satisfies the estimate* (2.17).

4. Local existence of solution for the nonlinear problem. In this section, we shall establish the local existence of classical solutions for the Cauchy problem of (1.8) and (1.13). Consider the initial value problem

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y = (T_{11})_x + (T_{12})_y, \\ (\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y = (T_{21})_x + (T_{22})_y, \end{cases} \tag{4.1}$$

$$\rho(x,y,0) = \rho_0(x,y), \quad u(x,y,0) = u_0(x,y), \quad v(x,y,0) = v_0(x,y). \tag{4.2}$$

We have the following theorem.

THEOREM 4.1. *For any initial data* (ρ_0, u_0, v_0) *such that the condition* $\rho_0 \geq \delta > 0$ *is satisfied and* $(\rho_0 - \bar{\rho}_0, u_0, v_0) \in H^k(R^2)$ $(k \geq 4)$, *where* $\bar{\rho}_0 > 0$ *is a constant, there exists a* $T > 0$ *such that for* $t \in [0,T]$, *the Cauchy problem* (4.1) *and* (4.2) *has a unique solution* (ρ, u, v) *such that* $\rho - \bar{\rho}_0 \in L^\infty([0,T]; H^{k+1}(R^2))$, $(u,v) \in L^\infty([0,T]; H^k(R^2))$, *and*

$$|||w|||_k^2 \leq C_k \|w_0\|_k^2 + \|\rho_0\|_{k+1}^2.$$

Remark 4.1. The solution in Theorem 4.1 is the classical solution, i.e., all the corresponding derivatives in (4.1) and (4.2) exist and are continuous and satisfy (4.1) and (4.2) in the classical sense.

Remark 4.1 is shown as follows. From

$$\rho \in L^\infty(0,T; H^5), \qquad (u,v) \in L^\infty(0,T; H^4), \tag{4.3}$$

we have, by the Sobolev imbedding theorem,

$$\nabla^3 \rho \in L^\infty(0,T; C^0), \qquad \nabla^2(u,v) \in L^\infty(0,T; C^0). \tag{4.4}$$

From (4.1), we obtain

$$\partial_t \rho \in L^\infty(0,T; H^3), \qquad \partial_t^2(u,v) \in L^\infty(0,T; H^2). \tag{4.5}$$

Therefore,

$$\partial_t \nabla^3 \rho \in L^\infty(0,T; H^0), \qquad \partial_t \nabla^2(u,v) \in L^\infty(0,T; H^0). \tag{4.6}$$

Combining (4.4) and (4.6), we have, by the trace theorem [8],

$$\nabla^3 \rho \in C(0,T; C^0), \qquad \nabla^2(u,v) \in C(0,T; C^0). \tag{4.7}$$

Using (4.1) again, we have

$$\partial_t^2 \rho \in L^\infty(0,T; H^1), \qquad \partial_t^2(u,v) \in L^\infty(0,T; H^0). \tag{4.8}$$

Combining (4.5) and (4.8), we have

$$\partial_t \rho \in C(0,T;C^0), \qquad \partial_t(u,v) \in C(0,T;C^0). \tag{4.9}$$

This concludes the proof of the remark.

As in §2, denote the quasilinear differential operator in (4.1) as

$$\mathcal{L}(w)w \equiv \begin{cases} (\partial_t + u\partial_x + v\partial_y)\rho + \rho(u_x + v_y) = 0, \\ (\partial_t + u\partial_x + v\partial_y)u + p'(\rho)\rho^{-1}\partial_x\rho - \nu\triangle\partial_x\rho - \mu\rho^{-1}\triangle u \\ \qquad\qquad - \frac{\mu}{3\rho}(\partial_{xx}u + \partial_{xy}v) = 0, \\ (\partial_t + u\partial_x + v\partial_y)v + p'(\rho)\rho^{-1}\partial_y\rho - \nu\triangle\partial_y\rho - \mu\rho^{-1}\triangle v \\ \qquad\qquad - \frac{\mu}{3\rho}(\partial_{xy}u + \partial_{yy}v) = 0. \end{cases}$$

Let $\tilde{w}_0(x,y,t)$ be the unique bounded solution for the Cauchy problem

$$\partial_t w - \triangle w = 0, \qquad w(x,y,0) = w_0(x,y). \tag{4.10}$$

It is readily checked that the solution $\tilde{w}_0(x,y,t)$ satisfies the following estimate:

$$\int_0^T \|\partial_t \tilde{w}_0\|_k^2 dt + \int_0^T \|\tilde{w}_0\|_{k+2}^2 dt \le C\|w_0\|_k^2. \tag{4.11}$$

Let $w \equiv \tilde{w}_0 + \dot{w}$, the problem (4.1) and (4.2) can be rewritten as the following Cauchy problem for the new unknown functions $\dot{w}$:

$$\begin{cases} \dot{L}(\dot{w})\dot{w} = -\mathcal{L}(\tilde{w}_0)\tilde{w}_0 \equiv \dot{f}, \\ \dot{w}(x,y,0) = 0, \end{cases} \tag{4.12}$$

where $\dot{L}(\dot{w})\dot{w} \equiv \mathcal{L}(\tilde{w}_0 + \dot{w})\dot{w} + (\mathcal{L}(\tilde{w}_0 + \dot{w}) - \mathcal{L}(\tilde{w}_0))\tilde{w}_0$ is the quasilinear differential operator of $\dot{w}$, whose linearization has the same structure as $\dot{L}$ in §2.

Since by (4.11), $\dot{f}_1 \in L^2(0,T;H^{k+1})$, and $\dot{f}_{2,3} \in L^2(0,T;H^k)$, it is obvious that Theorem 4.1 is equivalent to the following.

THEOREM 4.2. *Under the condition of Theorem* 4.1, *for any* $\dot{f}_1 \in L^2(0,T;H^{k+1})$ *and* $\dot{f}_{2,3} \in L^2(0,T;H^k)$, *there exists a* $T > 0$ *such that in* $t \in [0,T]$, *the Cauchy problem* (4.12) *has a unique solution* $\dot{w}$ *such that* $\dot{\rho} \in L^\infty([0,T];H^{k+1}(R^2))$, $(\dot{u},\dot{v}) \in L^\infty([0,T];H^k(R^2))$, *satisfying*

$$|||\dot{w}|||_k^2 \le C_k \int_0^T (\|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1}^2) dt. \tag{4.13}$$

Proof. Theorem 4.2 is proved by iteration. Let $\dot{w}_0(x,y,t) = 0$ and $\dot{w}_j(x,y,t)$ $(j = 1,2,\cdots)$ be defined as the unique solution of the following linear Cauchy problem:

$$\dot{L}(\dot{w}_{j-1})\dot{w}_j = \dot{f}, \qquad \dot{w}_j(x,y,0) = 0. \tag{4.14}$$

Choose $T \ll 1$ such that (2.15) is satisfied for $\tilde{w}_0$ in $[0,T]$. By the Sobolev imbedding theorem [9], we need only to show that for $T, \delta > 0$ sufficiently small, we have the successive solutions $\dot{w}_j (j = 1,2,\ldots)$, satisfying

$$|||\dot{w}_j|||_k^2 \le \delta \le \beta_k, \tag{4.15}$$

$$|||\dot{w}_j - \dot{w}_{j-1}|||_{k-2} \leq \tfrac{1}{2}|||\dot{w}_{j-1} - \dot{w}_{j-2}|||_{k-2}. \tag{4.16}$$

Assume (4.15) and (4.16) to be true for j and smaller indices. From the energy estimate (2.17) for the linearized problem, we have

$$|||\dot{w}_{j+1}|||_k^2 \leq C_k \int_0^T (\|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1}^2)dt.$$

On the other hand, $\dot{w}_{j+1} - \dot{w}_j$ satisfies the homogeneous initial data and the following equation:

$$\dot{L}(\dot{w}_j)(\dot{w}_{j+1} - \dot{w}_j) = \left(\dot{L}(\dot{w}_{j-1}) - \dot{L}(\dot{w}_j)\right)\dot{w}_j. \tag{4.17}$$

Applying (2.17) of the order $k-2$, we have

$$|||\dot{w}_{j+1} - \dot{w}_j|||_{k-2}^2 \leq C_{k-2}\delta \int_0^T |||\dot{w}_j - \dot{w}_{j-1}|||_{k-2}^2 dt. \tag{4.18}$$

Choosing δ such that $C_{k-2}\delta < \frac{1}{2}$, we obtain (4.16). This finishes the proof of Theorem 4.2.

REFERENCES

[1] G. ANDREWS AND J. M. BALL, *Asymptotic behavior and change of phase in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 44 (1982), pp. 306–341.

[2] J. E. DUNN AND J. SERRIN, *On the thermodynamics of interstitial working*, Arch. Rational Mech. Anal., 88 (1985), pp. 95–133.

[3] F. FALK, *Landau theory and martensitic phase transitions*, J. Physique, 43 (1983), pp. 3–15.

[4] H. HATTORI AND K. MISCHAIKOW, *A dynamical systems approach to a phase transition problem*, J. Differential Equations, 94 (1991), pp. 340–378.

[5] D. J. KORTEWEG, *Sur la forme que prennent les équations des movement des fluides si l'on tient comple des forces capillaires par des variations de densité*, Arch. Neerl. Sci. Exactes. Nat. Ser. II, 6 (1901), pp. 1–24.

[6] D. LI, *The general initial-boundary value problems of linear hyperbolic-parabolic coupled system*, Chinese Ann. Math. Ser. B, 7 (1986), pp. 408–424.

[7] ———, *The nonlinear initial-boundary value problem and the existence of multi-dimensional shock wave for quasilinear hyperbolic-parabolic coupled systems*, Chinese Ann. Math. Ser. B, 8 (1987), pp. 252–280.

[8] J. L. LIONS, *Quelques méthodes de résolutions des problèmes aux limites nonlinéaires*, Dunod, Gauthier–Villars, Paris, 1969.

[9] L. NIRENBERG, *On elliptic partial differential equations*, Ann. Scuola Norm. Sup. Pisa, 13 (1959), pp. 116–162.

[10] R. SAKAMOTO, *Mixed problems for hyperbolic equations, part* II, J. Math. Kyoto Univ., 10(1970), pp. 403–417.

[11] J. SERRIN, *The form of interfacial surfaces in Korteweg's theory of phase equilibria*, Quart. J. Appl. Math., 41 (1983), pp. 357–364.

[12] ———, *Phase transition and interfacial layers for van der Waals fluids*, in Proceedings of SAFA IV Conference, Recent Methods in Nonlinear Analysis and Applications, A. Camfora, S. Rionero, C. Sbordone, C. Trombetti, eds., Ligouri, Naples, 1980.

[13] M. SLEMROD, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Rat. Mech. Anal., 81 (1983), pp. 301–315.

[14] ———, *Dynamic phase transitions in a van der Walls fluid*, J. Differential Equations, 52 (1984), pp. 1–23.

[15] J. SPREKELS AND S. ZHENG, *Global solutions to the equations of a Ginsburg–Landau theory for structural phase transitions in shape memory alloys*, Phys. D., 39 (1989), pp. 59–76.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 99–134, January 1994

003

INSTABILITY OF PLANAR INTERFACES IN REACTION-DIFFUSION SYSTEMS*

MASAHARU TANIGUCHI† AND YASUMASA NISHIURA‡

Abstract. Instability of planar front solutions to reaction-diffusion systems in two space dimensions is studied. Let ε denote the width of interface. Then the planar front solution—or a solution having an internal transition layer which is flat—loses its stability when the length of interface along the tangential direction exceeds $O(\varepsilon^{1/2})$. The wavelength of the fastest growth is of $O(\varepsilon^{1/3})$ which is inherent in the system and determined by the nonlinearity and diffusion coefficients. Complete asymptotic characterization of these quantities as $\varepsilon \to 0$ is given by the analysis of what is called the singular dispersion relation derived from the linearized eigenvalue problem. The numerical computations also confirm that the theoretically predicted fastest growth wavy pattern actually arises from a randomly perturbed planar front.

Key words. stability, interface, singular perturbation, reaction diffusion system

AMS subject classifications. 35B25, 35B32, 35K57

1. Introduction and main results.

A variety of dissipative structures are created by symmetry breaking through successive bifurcations. The resulting patterns often have a separation boundary (or interfacial region) between two stable physical or chemical states. The simplest geometry of such interfaces is planar, and there is an extensive literature concerning "transition from planar to wavy patterns" in various fields. Especially in solidification problems, the classical Mullins–Sekerka instability is well known [MS], where the main issue is to determine the finite (or infinite) band of unstable wave numbers and find the fastest growth wavelength. They solved a free boundary problem of a Stefan type model where the interface has no thickness. However, there are several drawbacks to this approach, especially in the relation between linearized stability and nonlinear one which is in general unclear in this framework from a mathematical point of view (see, for instance, [St]). This motivates us to take another approach, namely, to adopt reaction-diffusion equations where the interface has small but positive thickness and there are no free boundaries. The phase field model (see, for instance, [Ca]) and activator-inhibitor systems, which we treat here, are the most well-known examples. The main objective of this paper is to describe rigorously the transition from planar to wavy patterns for the reaction-diffusion system (1.1).

Since our model system (1.1) is semilinear parabolic, it has several important advantages: global existence (in time) of solutions is easily obtained and the linearized stability implies a nonlinear one; it is an appropriate framework to study symmetry breaking bifurcations, since dynamical system theory can be applied to it; numerically it is quite easy to track the behavior of interface, since location of an interface is defined as a contour.

Despite this, there are very few rigorous results for asymptotic characterizations of reaction-diffusion systems. The main reason is that it is not an easy task to char-

*Received by the editors June 30, 1992; accepted for publication (in revised form) February 4, 1993.

†Department of Mathematical Sciences, University of Tokyo, Hongo, Tokyo 113, Japan.

‡Division of Mathematics and Informatics, Faculty of Integrated Arts and Sciences, Hiroshima University, Higashi-Hiroshima 724, Japan.

acterize asymptotic forms as the thickness of an interface tends to zero, since both planar solutions and eigenfunctions of the linearized eigenvalue problem have singularities at the layer position in the above limit. In fact, for some problems, sharp interface models where the width of interface equals zero, are easier to handle; see [Ch], [HNM], and [GGI].

However, as far as the stability problem is concerned, the SLEP method originated in Nishiura and Fujii [NF] turns out to be quite useful to resolve the difficulties mentioned above. They decomposed the linearized eigenvalue problem into singular and nonsingular parts like the Lyapunov–Schmidt reduction in bifurcation theory and gave an asymptotic characterization of the singular part. This method has been developed by [NM], [N1], and [N2], and enables us to control critical eigenvalues which are crucial to the study of stability and bifurcation; see also [Sa] and [GJ].

The aim of this paper is the following. First, we derive a fundamental relation of eigenvalues and wave numbers called the *singular dispersion relation* from the linearized eigenvalue problem at a planar front solution; second, by using this relation, we determine the size of instability region in wave number space; third, we characterize the asymptotic behavior of the fastest growth wavelength as well as the associated eigenvalue when the width of interface tends to zero; finally, we study numerically to what extent the linearized stability analysis is valid to predict qualitatively the final wavy patterns starting from a perturbed planar front.

The idea of [NF] is available to resolve the first two problems. For the third problem, however, a more subtle analysis is necessary to determine such wavelength, because the eigenfunction which characterizes such wavelength has no useful limit when the interface becomes sharp. It should be noted that our theory is based on the analysis of linear stability, thus the theoretically predicted fastest growth wavelength—namely that of $O(\varepsilon^{1/3})$—is valid only when the perturbation is small. Nonetheless, our numerical computations show that this wavelength is dominant even for largely deformed wavy patterns (see Fig. 2).

In [OMK], they obtained interesting bifurcation curves for special activator-inhibitor models. They could solve explicitly the linearized problem, since piecewise-linear nonlinearity was employed. However, they did not treat the third problem mentioned above.

The model system takes the following form:

$$
\begin{aligned}
&\tau \frac{\partial u}{\partial t} = \varepsilon \Delta u + \frac{1}{\varepsilon} f(u,v) && \\
&\qquad\qquad && (t,x,y) \in \mathbb{R}^+ \times \Omega, \\
&\frac{\partial v}{\partial t} = D \Delta v + g(u,v) && \\
&\frac{\partial u}{\partial \nu} = 0 = \frac{\partial v}{\partial \nu} && (t,x,y) \in \mathbb{R}^+ \times \partial\Omega,
\end{aligned}
\tag{1.1}
$$

where Ω is a rectangle in the (x,y)-plane:

$$
\Omega = (0,1) \times (0,\ell), \qquad \ell > 0, \tag{1.2}
$$

τ, ε and D are positive constants, $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, and ν denotes the unit outer normal vector. Assumptions for f and g will be stated at the end of this section. Let $\mathfrak{U}(\varepsilon) = (u(x,\varepsilon), v(x,\varepsilon))$ be a one-dimensional steady state solution (see Fig. 1), which

is defined in §2; then it is apparent that $\bar{\mathfrak{U}}(\varepsilon)$ defined by

$$\bar{\mathfrak{U}}(\varepsilon) \equiv (\bar{u}(x,y,\varepsilon), \bar{v}(x,y,\varepsilon)), \tag{1.3}$$

$$\begin{cases} \bar{u}(x,y,\varepsilon) \equiv u(x,\varepsilon) \\ \bar{v}(x,y,\varepsilon) \equiv v(x,\varepsilon) \end{cases} \quad \text{for any} \quad (x,y) \in \Omega \tag{1.4}$$

is a two-dimensional stationary solution of (1.1). The location of interface Γ^ε of $\bar{\mathfrak{U}}(\varepsilon)$ defined by

$$\Gamma^\varepsilon \equiv \{(x,y) \in \Omega; \bar{u}(x,y,\varepsilon) = h_0(\bar{v}(x,y,\varepsilon))\} \tag{1.5}$$

forms a straight line through $x = x^\varepsilon$, where x^ε denotes the layer position of $\mathfrak{U}(\varepsilon)$ (see Theorem 2.1).

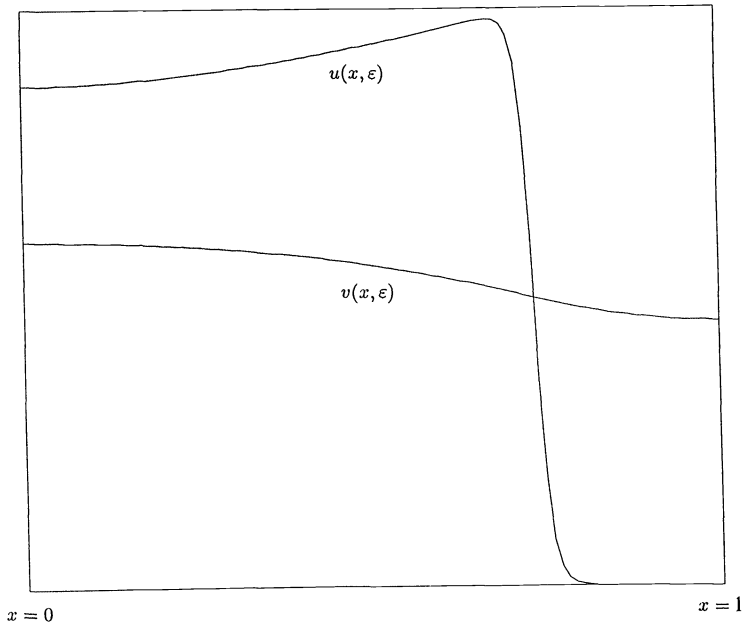


FIG. 1. *The graph of* $\mathfrak{U}(\varepsilon)$.

In order to study the stability properties of $\bar{\mathfrak{U}}(\varepsilon)$, we resort to the linearized principle, namely, we determine the spectral distribution of the following linearized eigenvalue problem at $\bar{\mathfrak{U}}(\varepsilon)$:

$$\begin{aligned} \begin{pmatrix} \mathfrak{L}(\varepsilon) & f_v(\bar{u},\bar{v}) \\ g_u(\bar{u},\bar{v}) & \mathfrak{M}(\varepsilon) \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix} &= \lambda \begin{pmatrix} \varepsilon\tau & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix} \quad \text{in } \Omega, \\ \frac{\partial w}{\partial \nu} = 0 = \frac{\partial z}{\partial \nu} \quad &\text{on } \partial\Omega, \end{aligned} \tag{1.6}$$

where $\mathfrak{L}(\varepsilon) \equiv \varepsilon^2\Delta + f_u(\bar{u},\bar{v})$ and $\mathfrak{M}(\varepsilon) \equiv D\Delta + g_v(\bar{u},\bar{v})$. Under the hypothesis given in Theorem 2.1 for (f,g) and parameters (τ, D), our results can be summarized as follows (more precise statements are given in §4).

MAIN THEOREM. (a) (Instability result.) *There exists* $\varepsilon_1 = \varepsilon_1(f, g, D, \tau) > 0$ *such that the following stability criterion holds for any fixed* $\varepsilon \in (0, \varepsilon_1)$:

$$\bar{\Omega}(\varepsilon) \text{ is stable if } \ell < \omega(\varepsilon)\varepsilon^{\frac{1}{2}}, \text{ and is unstable if } \ell > \omega(\varepsilon)\varepsilon^{\frac{1}{2}}, \tag{1.7}$$

where $\omega(\varepsilon)$ *is a positive function satisfying*

$$\lim_{\varepsilon \downarrow 0} \omega(\varepsilon) = \pi \widehat{\zeta_0}(0)^{-\frac{1}{2}} > 0. \tag{1.8}$$

(b) (The fastest growth wavelength.) *Let* $\lambda_{\max}(\varepsilon)$ *be the eigenvalue of* (1.6) *with the largest real part. Then, for any fixed* $\varepsilon \in (0, \varepsilon_1)$, $\lambda_{\max}(\varepsilon)$ *exists and is a real number satisfying*

$$\lim_{\varepsilon \downarrow 0} \lambda_{\max}(\varepsilon) = \widehat{\zeta_0}(0)/\tau > 0. \tag{1.9}$$

The eigenspace associated with $\lambda_{\max}(\varepsilon)$ *is the linear hull of a finite number of eigenfunctions, each of which can be expressed by*

$$\begin{aligned} w_0(x, y, \varepsilon) &= w_{m_0(\varepsilon)}(x) \cos \frac{m_0(\varepsilon)\pi y}{\ell}, \\ z_0(x, y, \varepsilon) &= z_{m_0(\varepsilon)}(x) \cos \frac{m_0(\varepsilon)\pi y}{\ell}, \end{aligned} \tag{1.10}$$

where $w_{m_0(\varepsilon)}(x)$, $z_{m_0(\varepsilon)}(x)$ *are smooth functions given by* (4.6), *and* $m_0(\varepsilon)$ *is some positive integer that satisfies*

$$\left| \varepsilon^{\frac{1}{3}} m_0(\varepsilon) - \frac{\ell}{\pi} \sqrt[3]{\frac{c_1^* c_2^*}{4D}} \right| < \delta(\varepsilon). \tag{1.11}$$

Here $\delta(\varepsilon)$ *is a positive function satisfying* $\delta(\varepsilon) \to 0$ *as* $\varepsilon \downarrow 0$, *and* c_1^*, c_2^* *are positive constants given by* (2.11). *The associated fastest growth wavelength is thus given by*

$$\mu_0(\varepsilon) = \frac{2\ell}{m_0(\varepsilon)} = 2\pi \sqrt[3]{\frac{4D}{c_1^* c_2^*}} \varepsilon^{\frac{1}{3}} + o(\varepsilon^{\frac{1}{3}}). \tag{1.12}$$

Note that the principal part of the right-hand side does not depend on ℓ.

REMARK 1.1. In the case when f is a piecewise linear function, part (a) was obtained by [OMK].

In the singular dispersion relation (see (3.25)), the term $\varepsilon\kappa^2$ corresponding to surface tension exerted along the tangential direction cannot be negligible, because it becomes dominant for deformations of short wavelengths (large wave numbers). On the other hand, from (1.11), we see that the most unstable wave number (see (4.5)) tends to ∞ as $\varepsilon \downarrow 0$. Dealing with the behavior of eigenfunctions for large wave numbers requires a delicate analysis. Also, it should be noted that although the eigenfunctions $(w_{m_0(\varepsilon)}(x), z_{m_0(\varepsilon)}(x))$ associated with $\lambda_{\max}(\varepsilon)$ have finite limits in appropriate function spaces as $\varepsilon \downarrow 0$, the limit of $z_{m_0(\varepsilon)}(x)$ turns out trivial. Hence it is not an easy task to extract useful information to obtain (1.11) from their limits. In general, when we deal with (1.6) in a multidimensional domain, we cannot expect

any "useful" limits of the eigenfunctions with the largest growth rate as $\varepsilon \downarrow 0$. This makes a sharp contrast with one-dimensional cases, in which the configuration of the limiting eigenfunction gives crucial information about the stability. In this paper, we always study (1.6), keeping ε positive, and get asymptotic expansions of eigenvalues for a small but positive ε.

From the Main Theorem, it is anticipated that a wavy pattern having the wavelength (1.12) starts to grow most rapidly at the first stage when a small perturbation is added to $\bar{\mathfrak{U}}(\varepsilon)$. We check this numerically for the data:

$$\begin{gathered} f(u,v) = u^2(\tfrac{3}{2} - u) - \tfrac{1}{2}uv, \qquad g(u,v) = \tfrac{3}{4}uv - \tfrac{1}{10}v - \tfrac{2}{5}v^2 \\ D = 0.4, \quad \varepsilon = 0.0075, \quad \ell = 2.46, \quad \tau = 1. \end{gathered} \tag{1.13}$$

In view of (A2), (A3), (2.11) and Remark 2.3, this immediately leads to

$$\begin{gathered} v^* = 1, \quad h_-(v^*) = 0, \quad h_0(v^*) = \tfrac{1}{2}, \quad h_+(v^*) = 1, \\ J'(v^*) = -\tfrac{1}{4}, \quad \gamma = \sqrt[4]{72}, \quad c_1^* = \tfrac{1}{4}\gamma, \quad c_2^* = \tfrac{3}{4}\gamma, \\ \Phi(\eta) = 1/\{1 + \exp(-\eta/\sqrt{2})\}. \end{gathered} \tag{1.14}$$

By numerical computation for the one-dimensional solution $\mathfrak{U}(\varepsilon)$, we have

$$\int_0^{x^*} g(U,V)dx = 0.117\ldots,$$

which together with (1.8) and (2.13) implies that $\omega(\varepsilon)\varepsilon^{1/2} = 0.35\ldots$. Hence, it follows from the Main Theorem (a) and (1.13) that $\bar{\mathfrak{U}}(\varepsilon)$ is unstable. Substituting (1.14) into (1.12), we see that the most unstable wavelength is equal to $1.232\ldots$, which implies that 4-mode wavy pattern should in general be selected after a short transition period. Here note that any mode higher than or equal to 8 does not grow from Theorem (a) and $\ell/8 < \omega(\varepsilon)\varepsilon^{1/2}$. Fig. 2 shows the evolution of a planar interface (see (1.5)) after we add a small bump on it (a) or give a random perturbation (b). The predicted wavelength (1.12) is seen to be dominant not only at the initial stage but also at the final stage in which the wavy patterns are fully developed. The final two states coincide almost exactly up to the phase shift.

On the other hand, suppose we put a pure sinusoidal perturbation of 5-mode on the planar front as in Fig. 3(c). The solution settles down to a wavy pattern of 5-mode, not 4-mode. This process is rather robust. In fact, even if we add a small bump to the previous perturbation as in Fig. 3(d), the same final pattern will eventually arise as before. This suggests the coexistence of stable wavy patterns (at least 4- and 5-modes) for the parameter values (1.13). Nevertheless, our numerical computations confirm that in most cases the solutions converge to the 4-mode pattern. Thus the theoretically predicted wavy pattern seems to have quite a large basin of attraction.

Now we state the assumptions for f and g.

(A1) f, g are smooth functions of u, v defined on some open set $\mathcal{O}$ in $\mathbb{R}^2$.

(A2) The nullcline $\{(u,v) \in \mathcal{O}; f(u,v) = 0\}$ is sigmoidal and consists of three curves C_i $(i = -, 0, +)$. We have $C_i \equiv \{(u,v) \in \mathcal{O}; u = h_i(v), v = I_i\}$, where h_i is a smooth function on an interval I_i $(i = -, 0, +)$. Let $\underline{v}$ (respectively, $\overline{v}$) be the minimum (respectively, maximum) of I_- (respectively, I_+), then we have $h_-(v) < h_0(v) < h_+(v)$ for any $v \in I^* \equiv (\underline{v}, \overline{v})$.

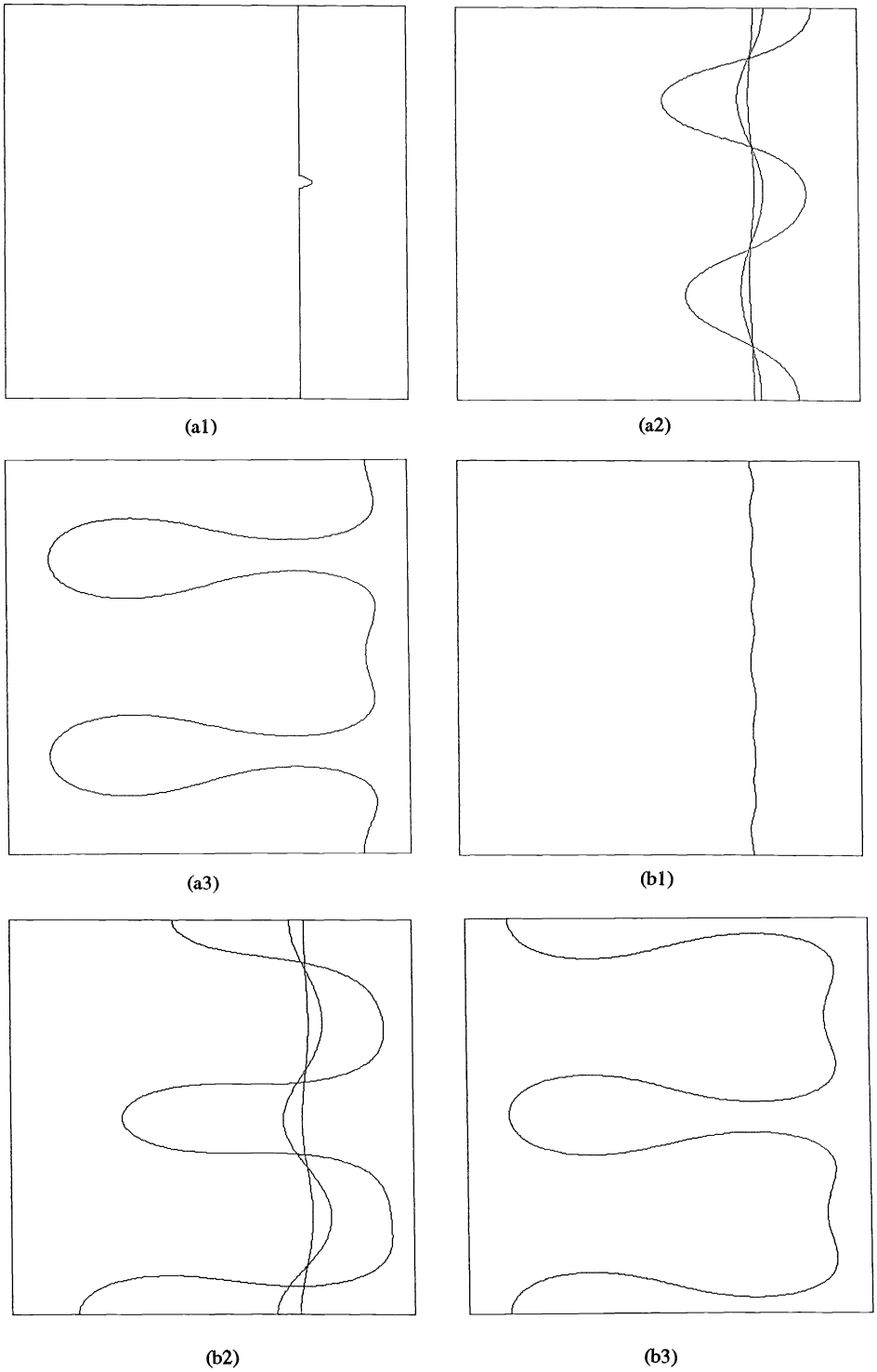


FIG. 2. *The evolution of interfaces from planar ones* I. (a) *A small projection is added to* Γ^ϵ; (b) Γ^ϵ *is perturbed randomly.*

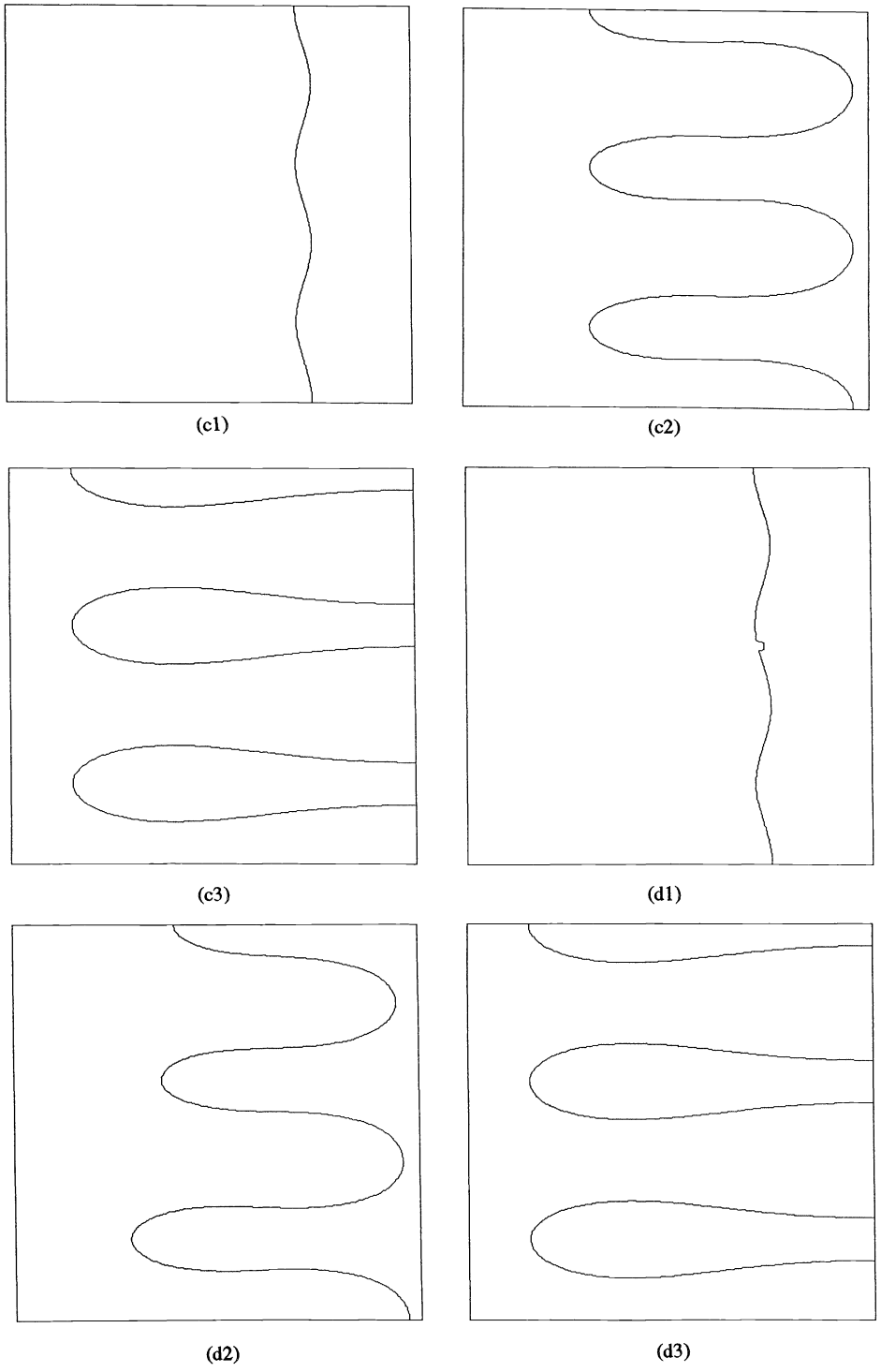


FIG. 3. *The evolution of interfaces from planar ones* II. (a) Γ^{ϵ} *is perturbed by a pure sinusoidal perturbation;* (b) *a small projection is added in addition.*

(A3) Let $J(v) \equiv \int_{h_-(v)}^{h_+(v)} f(s,v)ds$ for $v \in I^*$, then there exists $v^* \in I^*$ such that $J(v^*) = 0$, $J'(v^*) < 0$.

(A4) The nullcline of g intersects transversally with that of f. Let the intersection point on C_i, if it exists, be denoted by $P_i = (h_i(v_i), v_i)$, for $i = -, 0, +$. Then we assume that $v_- < v^* < v_+$.

(A5) (a) $f_u < 0$ on $R_- \cup R_+$, where $R_\pm$ are defined by

$$R_- = \{(h_-(v), v); v_- < v \leq v^*\}, \qquad R_+ = \{(h_+(v), v); v^* \leq v < v_+\};$$

(b) $g|_{R_-} < 0 < g|_{R_+}$;

(c) $(f_u g_v - f_v g_u)|_{R_- \cup R_+} > 0$, $g_v|_{R_- \cup R_+} \leq 0$.

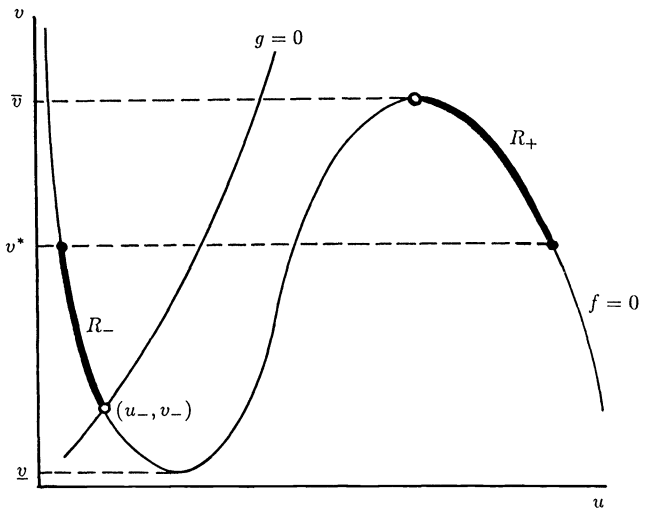


FIG. 4. *The nullclines of f and g.*

Fig. 4 shows typical functional forms of nullclines of f and g. It should be noted that, in order to satisfy (A1)–(A5), it is not necessary to assume that f and g intersect in this manner (*monostable type*).

The outline of this paper is as follows. In §2, we summarize several results for the Sturm–Liouville eigenvalue problem as well as the existence of a one-dimensional layered solution. In §3, we derive the singular dispersion relation (3.25) from the linearized eigenvalue problem at a planar front solution $\bar{\mathfrak{U}}(\varepsilon)$. In §4, we prove the main result Theorem 4.1, and show that the Main Theorem is a direct consequence of this theorem.

We shall use the following notation throughout this paper:

$I = (0,1)$.

$\mathbb{C}_a = \{\lambda \in \mathbb{C}; \Re\lambda \geq -a\}$ for $a > 0$.

$\mathcal{B}(X,Y)$ = the set of bounded linear operators from X to Y, where X and Y are Banach spaces.

$(\cdot,\cdot)_{L^2(I)}$=the inner product in $L^2(I)$.
$H^1(I)$, $H^{-1}(I)$ = the usual Sobolev space and its dual space.
${}_{H^1(I)}\langle\cdot,\cdot\rangle_{H^{-1}(I)}$ = the inner product between the above spaces.
o, O = the usual symbols of Landau.

2. Preliminaries. We start by summarizing the results of existence and stability of monolayered equilibrium solutions to (1.1) on a finite interval I:

$$\begin{aligned} \varepsilon\tau u_t &= \varepsilon^2 u_{xx} + f(u,v), \\ v_t &= Dv_{xx} + g(u,v), \\ u_x &= 0 = v_x, \end{aligned} \qquad \begin{aligned} &x \in I, \quad t > 0, \\ &\\ &x \in \partial I, \quad t > 0. \end{aligned} \tag{2.1}$$

We state the existence and stability of monolayered equilibrium solutions of (2.1) as follows.

THEOREM 2.1. *Assume f, g satisfy* (A1)–(A5). *Then there exists $D_* > 0$ and for any $D > D_*$, there exists $\tau_* > 0$ such that, for any fixed $(D,\tau) \in (D_*,\infty)\times(\tau_*,\infty)$,* (2.1) *has a stable equilibrium solution $\mathfrak{U}(\varepsilon) = (u(x,\varepsilon), v(x,\varepsilon))$ for arbitrary $\varepsilon \in (0,\varepsilon_0)$, where $\varepsilon_0 = \varepsilon_0(f,g,D,\tau) > 0$. $\{\mathfrak{U}(\varepsilon); \varepsilon \in (0,\varepsilon_0)\}$ is bounded in $C(\bar I)\times C^2(\bar I)$. Moreover there exists $x^* \in I$ and $V \in C^1(\bar I)$ which is a monotone decreasing function such that*

$$u(\cdot,\varepsilon) \to U \quad \text{in} \quad C([0,x^*-\iota]\cup[x^*+\iota,1]), \tag{2.2a}$$

$$v(\cdot,\varepsilon) \to V \quad \text{in} \quad C^1(\bar I), \tag{2.2b}$$

as $\varepsilon \downarrow 0$, where ι is an arbitrary positive number and

$$U(x) = \begin{cases} h_+(V(x)) & \text{if} \quad x \in [0,x^*], \\ h_-(V(x)) & \text{if} \quad x \in (x^*,1]. \end{cases} \tag{2.3}$$

Proof. See Appendix 1 in [NF] for the existence; see also [F], [Ito], [MTH], and [Sa]. For the stability, see [NF] and [NM].

REMARK 2.1. D_* is given in [NF, Prop. 1.1], and τ_* is expressed by (4.14) in §4.2.

REMARK 2.2. The location of the *interfacial point* $x = x^\varepsilon$ of $\mathfrak{U}(\varepsilon)$ is defined by

$$u(x^\varepsilon,\varepsilon) = h_0(v(x^\varepsilon,\varepsilon)). \tag{2.4}$$

The upper estimate for the distance between x^ε and x^* is given by $|x^\varepsilon - x^*| = O(\varepsilon)$. (See Sakamoto [Sa] for the proof.)

In order to solve (1.6), it is important to know the spectral behavior of the following Sturm–Liouville problem:

$$L(\varepsilon)[\phi] = \zeta\phi \quad \text{in } I, \qquad \phi_x = 0 \quad \text{on } \partial I, \tag{2.5}$$

where $L(\varepsilon)$ is a selfadjoint operator defined by

$$L(\varepsilon) \equiv \varepsilon^2 \frac{d^2}{dx^2} + f_u(u(x,\varepsilon), v(x,\varepsilon)). \tag{2.6}$$

Let $\{\phi_i(\varepsilon)\}_{i\ge 0}$ be the complete orthonormal set in $L^2(I)$ consisting of the eigenfunctions of $L(\varepsilon)$, and let $\{\zeta_i(\varepsilon)\}_{i\ge 0}$ be the associated eigenvalues. It is clear that $\{\zeta_i(\varepsilon)\}$ are real numbers ($\zeta_0(\varepsilon) > \zeta_1(\varepsilon) \ge \zeta_2(\varepsilon) \ge \cdots$). More precisely we have the following results.

PROPOSITION 2.1. (1) *There exists a constant* $\zeta_* > 0$ *such that*

$$\zeta_i(\varepsilon) < -\zeta_* < 0 < \zeta_0(\varepsilon) \qquad (i = 1, 2, 3, \dots) \tag{2.7}$$

is satisfied for any $\varepsilon \in (0, \varepsilon_0)$, *where* ε_0 *is the same one as in Theorem* 2.1. *Moreover* $\widehat{\zeta}_0(\varepsilon) \equiv \varepsilon^{-1}\zeta_0(\varepsilon)$ *converges to a positive constant* $\widehat{\zeta}_0(0)$ *as* $\varepsilon \downarrow 0$.

(2) $\{\varepsilon^{-1/2}\phi_0(\varepsilon)\}_{0<\varepsilon<\varepsilon_0}$ *is bounded in* $L^1(I)$, *and there exist positive constants* β, B *such that*

$$|\phi_0(x,\varepsilon)| \le B\varepsilon^{-\frac{1}{2}} \exp(-\beta|x - x^*|/\varepsilon) \quad \text{for } x \in I. \tag{2.8}$$

(3) *Let* h_0, h_1 *and* h_2 *be the functions defined by*

$$h_0(x,\varepsilon) = \varepsilon^{-\frac{1}{2}}\phi_0(x,\varepsilon), \tag{2.9a}$$

$$h_1(x,\varepsilon) = -\varepsilon^{-\frac{1}{2}} f_v(u(x,\varepsilon), v(x,\varepsilon))\phi_0(x,\varepsilon), \tag{2.9b}$$

$$h_2(x,\varepsilon) = \varepsilon^{-\frac{1}{2}} g_u(u(x,\varepsilon), v(x,\varepsilon))\phi_0(x,\varepsilon). \tag{2.9c}$$

Then they satisfy

$$h_i(x,\varepsilon) \longrightarrow c_i^*\delta(x - x^*) \quad \text{in } H^{-1}(I), \tag{2.10}$$

as $\varepsilon \downarrow 0$, *where* c_i^* *is a positive constant* $(i = 0, 1, 2)$, *and* $\delta(x - x^*)$ *is Dirac's* δ*-function at* x^*. *We have*

$$c_1^* = -\gamma J'(v^*), \qquad c_2^* = \gamma\{g(h_+(v^*), v^*) - g(h_-(v^*), v^*)\}, \tag{2.11}$$

where γ *is a positive constant* (*see* (2.13)).

Proof. See [NF], [NM], and [Sa] for the proof.

COROLLARY 2.1. *Under the same notation as in Proposition* 2.1, *we have*

$$|h_i(x,\varepsilon)| \le B\varepsilon^{-1} \exp(-\beta|x - x^*|/\varepsilon) \quad \text{for } x \in I, \tag{2.12a}$$

$$\int_I |h_i(x,\varepsilon)|dx \le 2B/\beta, \tag{2.12b}$$

for $i = 0, 1, 2$. *In particular, we have* $\|h_i(x,\varepsilon)\|_{H^{-1}(I)} < B_*$ *for some* $B_* = B_*(f, g, D) > 0$.

Proof of Corollary 2.1. We obtain the conclusions from (2.8), (2.9), and the fact that $\{\mathfrak{U}(\varepsilon); \varepsilon \in (0, \varepsilon_0)\}$ is bounded in $C(\bar{I}) \times C^2(\bar{I})$. We replaced B by a larger one, if necessary. Since $h_i(\varepsilon)$ is bounded in $L^1(I)$ uniformly for ε, it is uniformly bounded also in $H^{-1}(I)$ $(i = 1, 2)$. □

REMARK 2.3. According to [NF], γ and $\widehat{\zeta}_0(0)$ in Proposition 2.1 are explicitly given as follows. Let $\Phi(\eta)$ be a monotone increasing function in $C^\infty(\mathbb{R})$ defined by the unique solution of

$$\frac{d^2\Phi}{d\eta^2} + f(\Phi(\eta), v^*) = 0, \quad \Phi(0) = h_0(v^*), \quad \Phi(\pm\infty) = h_\pm(v^*),$$

then they are given by

$$\begin{aligned} \gamma &= \|d\Phi/d\eta\|_{L^2(I)}^{-1}, \\ \widehat{\zeta}_0(0) &= -\gamma^2 J'(v^*) D^{-1} \int_0^{x^*} g(U(x), V(x))dx. \end{aligned} \tag{2.13}$$

3. Derivation of the singular dispersion relation for the planar front. The aim of this section is to derive the singular dispersion relation (3.25) of eigenvalues and wave numbers for the planar front. Because of the singular nature of the transition layer as $\varepsilon \downarrow 0$, we immediately encounter several difficulties to solve (1.6), namely

(1) The highest order term of $\mathfrak{L}(\varepsilon)$ degenerates when $\varepsilon \downarrow 0$;

(2) $\bar{\mathfrak{U}}(\varepsilon)$ and hence the coefficients of (1.6) become discontinuous at the layer position when $\varepsilon \downarrow 0$.

The method given by Nishiura and Fujii [NF] is one of the useful tools for resolving these difficulties, and in fact we can derive (3.25) with the aid of this. However, we have another subtle problem which is inherent in higher-dimensional case, namely the dependency of eigenvalues on the wave number along the interfacial direction. It turns out that the fastest growth wave number depends on ε and tends to infinity when $\varepsilon \downarrow 0$. A careful analysis is needed to describe this dependency, and will be discussed in §4.

It is convenient to use a complete orthonormal system $\{Y_m\}_{m=0}^{\infty}$ in $L^2(0,\ell)$, where

$$Y_m(y) = \begin{cases} \ell^{-1/2} & \text{for } m = 0, \\ \sqrt{2}\ell^{-1/2}\cos(m\pi y/\ell) & \text{for } m > 0. \end{cases} \tag{3.1}$$

For (w,z) in (1.6), we set

$$w_m(x) = \int_0^\ell w(x,y)Y_m(y)dy, \qquad z_m(x) = \int_0^\ell z(x,y)Y_m(y)dy, \tag{3.2}$$

for $x \in I$, $(m = 0,1,2,\cdots)$. Then (w,z) is expanded as follows:

$$w(x,y) = \sum_{m=0}^{\infty} w_m(x)Y_m(y), \qquad z(x,y) = \sum_{m=0}^{\infty} z_m(x)Y_m(y) \tag{3.3}$$

in $L^2(\Omega)$. This decomposes (1.6) into the following countably many eigenvalue problems for $(w_m(x), z_m(x))$ and $\lambda \in \mathbb{C}$ with $m \in \bar{\mathbb{N}}$:

$$(L(\varepsilon) - \varepsilon^2\kappa^2)w_m + f_v(\bar{u},\bar{v})z_m = \varepsilon\tau\lambda w_m, \tag{3.4a}$$

$$\left(D\frac{d^2}{dx^2} + g_v(\bar{u},\bar{v}) - D\kappa^2\right)z_m + g_u(\bar{u},\bar{v})w_m = \lambda z_m, \tag{3.4b}$$

on I, subject to the zero flux boundary conditions

$$\frac{dw_m}{dx}(0) = 0 = \frac{dw_m}{dx}(1), \qquad \frac{dz_m}{dx}(0) = 0 = \frac{dz_m}{dx}(1), \tag{3.4c}$$

where $\kappa \equiv m\pi/\ell$ and

$$L(\varepsilon) \equiv \varepsilon^2\frac{d^2}{dx^2} + f_u(\bar{u},\bar{v}). \tag{3.5}$$

Since the planar front $\bar{\mathfrak{U}}(\varepsilon) = (\bar{u},\bar{v})$ defined by (1.3) and (1.4) is independent of y, so are $f_v(\bar{u},\bar{v}), g_v(\bar{u},\bar{v}), \ldots$. It is obvious that $L(\varepsilon)$ defined above is the same as (2.6).

REMARK 3.1. If (3.4) has a solution $\lambda \in \mathbb{C}$ and $(w_m(x), z_m(x)) \not\equiv (0,0)$, then $(w_m(x)Y_m(y), w_m(x)Y_m(x))$ satisfies (1.6) with the same λ. On the contrary, if (1.6) has an eigenvalue $\lambda \in \mathbb{C}$ and an eigenfunction (w, z), then $(w_m(x), z_m(x))$ defined by (3.2) satisfies (3.4) with the same λ for any $m \in \bar{\mathbb{N}}$. Moreover, $(w_m(x), z_m(x))$ is nontrivial for some $m \in \bar{\mathbb{N}}$.

This remark allows us to study (3.4) for each $m \in \bar{\mathbb{N}}$ instead of studying the eigenvalues and eigenfunctions of (1.6). Hereafter we regard κ as a continuous valuable in $[0, \infty)$, although κ takes only discrete values. We begin with the following fact.

PROPOSITION 3.1. *Let $\lambda \in \mathbb{C}$ be an eigenvalue of* (3.4) *for some $\kappa \in [0, \infty)$. There exists a positive constant $\lambda_* = \lambda_*(f, g, D)$, and for any given $\delta > 0$, there is $\bar{\varepsilon}(\delta)$ such that either*

$$\Re\lambda < -\lambda_* < 0 \quad \text{or} \quad |\varepsilon^2\kappa^2 + \varepsilon\tau\lambda| < \delta$$

holds for any $\varepsilon \in (0, \bar{\varepsilon}(\delta))$.

Proof. The proof of this lemma can be carried out by the same argument as in [NF,Prop.2.1] with no essential changes. We omit the proof. □

It suffices to study the behavior of eigenvalues in $\mathbb{C}_{\lambda_*}$ because the others have nothing to do with the instability of $\bar{\mathfrak{U}}(\varepsilon)$. Therefore we may assume, by virtue of Proposition 3.1, that there is a positive function $\bar{\delta}(\varepsilon)$ ($\bar{\delta}(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$) such that

$$|\varepsilon^2\kappa^2 + \varepsilon\tau\lambda| < \bar{\delta}(\varepsilon), \tag{3.6}$$

for any $\varepsilon \in (0, \varepsilon_0)$, where $\varepsilon_0 = \varepsilon_0(f, g, D, \tau) > 0$ is the same one as in Theorem 2.1. We note that $\bar{\delta}(\varepsilon)$ is independent of (κ, λ). From (3.6), we have, for small $\varepsilon > 0$,

$$|\zeta_i(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda| \geq \zeta_*/2 > 0 \qquad (i = 1, 2, 3, \cdots), \tag{3.7}$$

which guarantees the existence of (3.9) below. We make some preparations in order to derive the relation (3.25).

First we introduce two operators from $L^2(I)$ to $L^2(I)$ as follows:

$$R(\varepsilon, \kappa, \lambda) \equiv -g_v(\bar{u}, \bar{v}) - g_u(\bar{u}, \bar{v})(L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)^\dagger(-f_v(\bar{u}, \bar{v})\cdot), \tag{3.8a}$$

$$S(\varepsilon, \kappa, \lambda) \equiv -g_u(\bar{u}, \bar{v})(L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)^{2\dagger}(-f_v(\bar{u}, \bar{v})\cdot), \tag{3.8b}$$

when $\varepsilon > 0$, where

$$(L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)^\dagger \equiv \sum_{i=1}^{\infty} \frac{(\cdot, \phi_i(\varepsilon))_{L^2(I)}}{\zeta_i(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda}\phi_i(\varepsilon), \qquad L^2(I) \to L^2(I), \tag{3.9}$$

and when $\varepsilon = 0$, we set

$$R(0, \kappa, \lambda) = (f_u g_v - f_v g_u)/(-f_u)\big|_{\substack{u=U(x)\\v=V(x)}}, \tag{3.10a}$$

$$S(0, \kappa, \lambda) = f_v g_u / {f_u}^2\big|_{\substack{u=U(x)\\v=V(x)}}. \tag{3.10b}$$

Note that the right-hand sides of (3.10) are independent of (κ, λ), and that the right-hand side of (3.10a) is a strictly positive function on I, by virtue of (2.3) and (A5) in §1. Next we define a sesquilinear form $B(\varepsilon, \kappa, \lambda) : H^1(I) \times H^1(I) \to \mathbb{C}$ as follows:

$$B(\varepsilon, \kappa, \lambda)(z^1, z^2) \equiv D\,(z^1_x, z^2_x)_{L^2(I)} + ((R(\varepsilon, \kappa, \lambda) + D\kappa^2 + \lambda)\,z^1, z^2)_{L^2(I)}, \tag{3.11}$$

for $z^1, z^2 \in H^1(I)$. We also define an operator $T(\varepsilon, \kappa, \lambda) : H^1(I) \to H^{-1}(I)$ by

$$T(\varepsilon, \kappa, \lambda)z \equiv -Dz_{xx} + (R(\varepsilon, \kappa, \lambda) + D\kappa^2 + \lambda)z, \tag{3.12}$$

for $z \in H^1(I)$. Applying the standard Lax–Milgram theorem, we have the following lemma.

LEMMA 3.1. *There exist* $\varepsilon_* > 0$, $\delta_* > 0$ $(0 < \delta_* < \zeta_*/2)$ *such that* (i)–(v) *hold true for* $\varepsilon \in [0, \varepsilon_*)$, $\kappa \in [0, \infty)$ *and* $\lambda \in \mathbb{C}_{\lambda_*}$ *which satisfy*

$$|\varepsilon^2\kappa^2 + \varepsilon\tau\lambda| < \delta_*. \tag{3.13}$$

(i) $R(\varepsilon, \kappa, \lambda)$, $S(\varepsilon, \kappa, \lambda)$ *are uniformly bounded linear operators from* $L^2(I)$ *to* $L^2(I)$ *for* $(\varepsilon, \kappa, \lambda)$.

(ii) $B(\varepsilon, \kappa, \lambda)$ *is a bounded and coercive sesquilinear form on* $H^1(I)$.

(iii) $T(\varepsilon, \kappa, \lambda)$ *belongs to* $\mathcal{B}(H^1(I), H^{-1}(I))$, *and has an inverse operator denoted by* $K(\varepsilon, \kappa, \lambda)$. $K(\varepsilon, \kappa, \lambda)$ *is a uniformly bounded linear operator from* $H^{-1}(I)$ *to* $H^1(I)$ *for* $(\varepsilon, \kappa, \lambda)$.

(iv) *If we assume* (3.6) *in addition, we have*

$$K(\varepsilon, \kappa, \lambda) \to K(0, \kappa, \lambda) \quad \text{in} \quad \mathcal{B}(H^{-1}(I), H^1(I)),$$

as $\varepsilon \downarrow 0$. *The convergence is uniform for* (κ, λ) *in any compact subset of* $[0, \infty) \times \mathbb{C}_{\lambda_*}$.

(v) $K(\varepsilon, \kappa, \lambda)$ *depends continuously on* ε, *analytically on* $\lambda \in \overset{\circ}{\mathbb{C}}_{\lambda_*}$ *and real-analytically on* $\kappa > 0$ *in* $\mathcal{B}(H^{-1}(I), H^1(I))$, *respectively, and satisfies*

(3.14a)
$$\frac{\partial K}{\partial \lambda}(\varepsilon, \kappa, \lambda) = -K(\varepsilon, \kappa, \lambda)K(\varepsilon, \kappa, \lambda) - \varepsilon\tau K(\varepsilon, \kappa, \lambda)S(\varepsilon, \kappa, \lambda)K(\varepsilon, \kappa, \lambda),$$

(3.14b)
$$\frac{\partial K}{\partial \kappa}(\varepsilon, \kappa, \lambda) = -2\kappa\left\{D \cdot K(\varepsilon, \kappa, \lambda)K(\varepsilon, \kappa, \lambda) + \varepsilon^2 K(\varepsilon, \kappa, \lambda)S(\varepsilon, \kappa, \lambda)K(\varepsilon, \kappa, \lambda)\right\}.$$

Proof. We first note that $\bar{\mathfrak{U}} = (\bar{u}, \bar{v})$ can be regarded as a function of x, and is uniformly bounded in $C(\bar{I}) \times C^2(\bar{I})$ for ε. Using this fact, (2.7) and (3.13), we can see the right-hand sides of (3.8), (3.10) belong to $\mathcal{B}(L^2(I), L^2(I))$ and satisfy the conclusions of (i). The proofs of (ii)–(v) can be carried out by the same arguments as in [NF, Lemma 3.1] by making use of (3.13). □

We set $\varepsilon_1 = \varepsilon_1(f, g, D, \tau) > 0$ such that we have

$$\varepsilon_1 < \min\{\varepsilon_*, \varepsilon_0\}, \qquad 0 < \bar{\delta}(\varepsilon) < \delta_* \quad \text{for any} \quad \varepsilon \in (0, \varepsilon_1). \tag{3.15}$$

We have, from (3.4a),

$$(L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)w_m = -f_v(\bar{u}, \bar{v})z_m, \tag{3.16}$$

for any fixed $\varepsilon \in (0, \varepsilon_1)$. Multiplying the both sides of (3.16) by $\phi_i(\varepsilon)$, and integrating over I, we obtain

$$(\zeta_i(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)(w_m, \phi_i(\varepsilon))_{L^2(I)} = (-f_v(\bar{u}, \bar{v})z_m, \phi_i(\varepsilon))_{L^2(I)}. \tag{3.17}$$

In view of (3.7), (3.9) and (3.17), we see that it is necessary that $w_m(x)$ should have the form:

$$w_m = \alpha_0\phi_0(\varepsilon) + (L - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)^\dagger(-f_v(\bar{u}, \bar{v})z_m) \tag{3.18}$$

for some $\alpha_0 \in \mathbb{C}$. The solvability of (3.16) will be discussed in the proof of Lemma 3.2. In what follows we shall derive necessary conditions so that λ may be an eigenvalue of (3.4). Inserting (3.18) into (3.4b), we obtain from (3.12)

$$T(\varepsilon,\kappa,\lambda)z_m = \alpha_0 g_u(\bar{u},\bar{v})\phi_0(\varepsilon), \tag{3.19}$$

which yields

$$z_m = \alpha_0 K(\varepsilon,\kappa,\lambda)(g_u(\bar{u},\bar{v})\phi_0(\varepsilon)). \tag{3.20}$$

Putting $\alpha = \varepsilon^{1/2}\alpha_0$ and using (2.9), we rewrite (3.18) and (3.20) as follows.

$$z_m = \alpha K(\varepsilon,\kappa,\lambda)h_2(\varepsilon), \tag{3.21a}$$

$$w_m = \alpha\varepsilon^{-\frac{1}{2}}\phi_0(\varepsilon) + (L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)^\dagger(-f_v(\bar{u},\bar{v})z_m). \tag{3.21b}$$

Multiplying both sides of (3.4a) by $\phi_0(\varepsilon)$ and integrating over I, we find

$$(\zeta_0(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)(w_m,\phi_0(\varepsilon))_{L^2(I)} = (-f_v(\bar{u},\bar{v})z_m,\phi_0(\varepsilon))_{L^2(I)}. \tag{3.22}$$

From (3.21b), we have $(w_m,\phi_0(\varepsilon))_{L^2(I)} = \alpha\varepsilon^{-1/2}$. Hence we see from (2.9) and (3.21a) that (3.22) can be rewritten as

$$\alpha(\widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa^2 - \tau\lambda) = \alpha(K(\varepsilon,\kappa,\lambda)h_2(\varepsilon),h_1(\varepsilon))_{L^2(I)}. \tag{3.23}$$

Without loss of generality, we can assume (w_m,z_m) to be nontrivial, which implies $\alpha \neq 0$. Thus we obtain from (3.23)

$$\widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa^2 - \tau\lambda = (K(\varepsilon,\kappa,\lambda)h_2(\varepsilon),h_1(\varepsilon))_{L^2(I)}. \tag{3.24}$$

LEMMA 3.2. *For any fixed $\varepsilon \in (0,\varepsilon_1)$, $\lambda \in \mathbb{C}_{\lambda_*}$ is an eigenvalue of* (3.4) *if and only if, it satisfies*

$$F(\varepsilon,\kappa,\lambda) = 0, \tag{3.25}$$

where

$$F(\varepsilon,\kappa,\lambda) \equiv \widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa^2 - \tau\lambda - H(\varepsilon,\kappa,\lambda), \tag{3.26}$$

$$H(\varepsilon,\kappa,\lambda) \equiv (K(\varepsilon,\kappa,\lambda)h_2(\varepsilon),h_1(\varepsilon))_{L^2(I)}. \tag{3.27}$$

The associated eigenfunction of (3.4) *is represented by* (3.21). *We call* (3.25) *the singular dispersion relation for the planar front.*

Proof. Tracing back the arguments before Lemma 3.2, we see that all what we have to do is to show the solvability of (3.16) starting from (3.25) and (3.21a). Let us consider the homogeneous equation:

$$(L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)\phi = 0 \quad \text{in } I, \qquad \phi_x = 0 \quad \text{on } \partial I. \tag{3.28}$$

In the case where (3.28) has no nontrivial solution, we can solve (3.16) as follows.

$$\begin{aligned} w_m =& (L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)^{-1}(-f_v(\bar{u},\bar{v})z_m) \\ =& \frac{(-f_v(\bar{u},\bar{v})z_m,\phi_0(\varepsilon))_{L^2(I)}}{\zeta_0(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda}\phi_0(\varepsilon) + (L(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda)^\dagger(-f_v(\bar{u},\bar{v})z_m). \end{aligned} \tag{3.29}$$

From (3.21a) and (3.25), we have

$$(\text{The first term of (3.29)}) = \alpha\varepsilon^{-\frac{1}{2}}\phi_0(\varepsilon).$$

Putting together the above two equalities, we obtain (3.18). Next we consider the case where (3.28) has an nontrivial solution. In view of (3.7), we see that it is $\phi_0(\varepsilon)$. We can assume that

$$\zeta_0(\varepsilon) - \varepsilon^2\kappa^2 - \varepsilon\tau\lambda = 0,$$

which, together with (3.25) yields

$$H(\varepsilon, \kappa, \lambda) = 0.$$

Hence we obtain, from (3.21a),

$$\begin{aligned}(-f_v(\bar{u}, \bar{v})z_m, \phi_0(\varepsilon))_{L^2(I)} =& \alpha(K(\varepsilon, \kappa, \lambda)h_2(\varepsilon), -f_v(\bar{u}, \bar{v})\phi_0(\varepsilon))_{L^2(I)} \\ =& \alpha\varepsilon^{\frac{1}{2}}(K(\varepsilon, \kappa, \lambda)h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)} \\ =& \alpha\varepsilon^{\frac{1}{2}}H(\varepsilon, \kappa, \lambda) = 0,\end{aligned}$$

which equals to the orthogonality condition in (3.16). This completes the proof. □

4. Proof of the Main Theorem. The aim of this section is to give a proof of the Main Theorem in §1. We first state a key result Theorem 4.1 in §4.1, which immediately leads us to the Main Theorem. In §4.2 and §4.3, we study the precise behaviors of solutions of (3.25) with respect to κ, which are the main ingredients of §4.4, where we present a proof of Theorem 4.1. More precisely, a priori bound for solutions of (3.25) is given in §4.2. In §4.3 we parameterize the eigenvalues by κ of the form $\lambda = \widetilde{\lambda}(\varepsilon, \kappa)$ with the aid of the standard implicit function theorem. We also show in §4.3, that λ associated with sufficiently small κ and sufficiently large κ is negative-valued. In §4.4 we characterize $\kappa(\varepsilon)$ in Theorem 4.1 by $\left(\partial\widetilde{\lambda}/\partial\kappa\right)(\varepsilon, \kappa(\varepsilon)) = 0$. Although it is just a necessary condition for $\kappa(\varepsilon)$, it turns out to be sufficient to control the behavior of $\kappa(\varepsilon)$ when $\varepsilon \downarrow 0$ (see (4.2)).

4.1. Asymptotic behaviors of the fastest growth wave numbers and their eigenvalues. The following is a key result in this section.

THEOREM 4.1. *Under the same assumptions for D and τ as in Theorem 2.1, there exists $\varepsilon_1 = \varepsilon_1(f, g, D, \tau) > 0$ such that (1), (2), and (3) hold true for any fixed $\varepsilon \in (0, \varepsilon_1)$.*

(1) *Let $\lambda \in \mathbb{C}_{\lambda_*}$ satisfy (3.25), then λ is real.*

(2) *There exist $\underline{\kappa}(\varepsilon)$, $\overline{\kappa}(\varepsilon)$ $(0 < 2\underline{\kappa}(\varepsilon) < \overline{\kappa}(\varepsilon) < \infty)$ such that, if $\lambda \in (-\lambda_*, \infty)$ satisfies (3.25) with some $\kappa \notin [\underline{\kappa}(\varepsilon), \overline{\kappa}(\varepsilon)]$, then $\lambda < 0$.*

(3) *Let $\mathcal{S}$ be the set of (κ, λ) satisfying (3.25) and $\lambda \geq 0$, then we have*

$$\mathcal{S} = \{(\kappa, \lambda); \kappa \in [\underline{\kappa}(\varepsilon), \overline{\kappa}(\varepsilon)], \lambda = \widetilde{\lambda}(\varepsilon, \kappa)\},$$

where $\widetilde{\lambda}(\varepsilon, \cdot)$ is a real-valued function that fulfills

(i) $\widetilde{\lambda}(\varepsilon, \cdot) \in C^\infty[\underline{\kappa}(\varepsilon), \overline{\kappa}(\varepsilon)]$;

(ii) $\widetilde{\lambda}(\varepsilon, \underline{\kappa}(\varepsilon)) = 0 = \widetilde{\lambda}(\varepsilon, \overline{\kappa}(\varepsilon))$;

(iii) $\widetilde{\lambda}(\varepsilon, \kappa) > 0$ *for any* $\kappa \in (\underline{\kappa}(\varepsilon), \overline{\kappa}(\varepsilon))$

with the asymptotic limits

$$\lim_{\varepsilon\downarrow 0}\varepsilon\overline{\kappa}(\varepsilon)^2=\widehat{\zeta}_0(0)\quad\text{and}\quad\lim_{\varepsilon\downarrow 0}\underline{\kappa}(\varepsilon)=\underline{\kappa}(0)\in(0,\infty). \tag{4.1}$$

Moreover, let $\kappa(\varepsilon)$ be any local maximizer of $\widetilde{\lambda}(\varepsilon,\cdot)$ in $(\underline{\kappa}(\varepsilon),\overline{\kappa}(\varepsilon))$; then we have, as $\varepsilon\downarrow 0$,

$$\varepsilon\kappa(\varepsilon)^3\to\frac{c_1^*c_2^*}{4D}, \tag{4.2}$$

$$\widetilde{\lambda}(\varepsilon,\kappa)\to\widehat{\zeta}_0(0)/\tau\quad\textit{uniformly for }\kappa\in(\kappa(\varepsilon)-M,\kappa(\varepsilon)+M), \tag{4.3}$$

where M is any positive constant. Note that the asymptotic characterizations (4.2) *and* (4.3) *do not depend on the choice of a local maximizer.*

Proof of Main Theorem (a). Let $\underline{m}(\varepsilon)$, $\overline{m}(\varepsilon)$ be defined by

$$\underline{m}(\varepsilon)=\ell\underline{\kappa}(\varepsilon)/\pi,\qquad\overline{m}(\varepsilon)=\ell\overline{\kappa}(\varepsilon)/\pi.$$

We have $0<2\underline{m}(\varepsilon)<\overline{m}(\varepsilon)<\infty$. By virtue of Remark 3.1, Lemma 3.2, and Theorem 4.1, we have the following alternative.

Case 1. If $[\underline{m}(\varepsilon),\overline{m}(\varepsilon)]$ includes no positive integer, then (1.6) has no eigenvalue whose real part is nonnegative.

Case 2. If $(\underline{m}(\varepsilon),\overline{m}(\varepsilon))$ includes a positive integer, then (1.6) has at least one positive real eigenvalue.

Let us put $\omega(\varepsilon)\equiv\pi\varepsilon^{-1/2}\overline{\kappa}(\varepsilon)^{-1}$. We have

$$\omega(\varepsilon)\varepsilon^{\frac{1}{2}}=\pi\overline{\kappa}(\varepsilon)^{-1}=\ell/\overline{m}(\varepsilon). \tag{4.4}$$

When $\ell<\omega(\varepsilon)\varepsilon^{1/2}$, we have $\overline{m}(\varepsilon)<1$ from (4.4). Hence Case 1 holds in this case, which implies that $\overline{\mathfrak{U}}(\varepsilon)$ is stable. When $\ell>\omega(\varepsilon)\varepsilon^{1/2}$, we have $\overline{m}(\varepsilon)>1$ from (4.4). Putting $\overline{m}(\varepsilon)=N+\nu$ (N is a positive integer; $0<\nu\le 1$), we get

$$\underline{m}(\varepsilon)<\overline{m}(\varepsilon)/2=(N+\nu)/2\le N,$$

which implies that a positive integer N belongs to $(\underline{m}(\varepsilon),\overline{m}(\varepsilon))$. We see that Case 2 holds, and hence $\overline{\mathfrak{U}}(\varepsilon)$ turns out to be unstable in this case. □

Proof of Main Theorem (b). From Main Theorem (a), we see that Case 2 holds and hence $\overline{\mathfrak{U}}(\varepsilon)$ becomes unstable for sufficiently small ε. We set

$$\overline{\lambda}(\varepsilon,m)\equiv\widetilde{\lambda}(\varepsilon,m\pi/\ell),$$

for $m\in\mathbb{N}\cap(\underline{m}(\varepsilon),\overline{m}(\varepsilon))$. In view of Theorem 4.1, we see that there exist a finite number of positive integers which maximize $\overline{\lambda}(\varepsilon,\cdot)$. Let us denote arbitrary one of them by $m_0(\varepsilon)$. Then $\widetilde{\lambda}(\varepsilon,\cdot)$ must have at least one local maximizer in $((m_0(\varepsilon)-1)\pi/\ell,(m_0(\varepsilon)+1)\pi/\ell)$. Let $\kappa(\varepsilon)$ be any one of them. We can apply Theorem 4.1 for this $\kappa(\varepsilon)$. Then it follows that

$$\varepsilon\kappa(\varepsilon)^3\to\frac{c_1^*c_2^*}{4D}\quad(\text{as}\quad\varepsilon\downarrow 0),\qquad|m_0(\varepsilon)-\ell\kappa(\varepsilon)/\pi|<1$$

are satisfied, which yields (1.11). We note here that (1.11) holds true regardless of our way of selecting $m_0(\varepsilon)$. We put

$$\kappa_0(\varepsilon) = m_0(\varepsilon)\pi/\ell. \tag{4.5}$$

Then we have $|\kappa_0(\varepsilon) - \kappa(\varepsilon)| < \pi/\ell$. Using (4.3), we obtain

$$\widetilde{\lambda}(\varepsilon, \kappa_0(\varepsilon)) \to \widehat{\zeta}_0(0)/\tau,$$

as $\varepsilon \downarrow 0$ which, together with $\lambda_{\max}(\varepsilon) = \widetilde{\lambda}(\varepsilon, \kappa_0(\varepsilon))$, yields (1.9). Next, from (3.21), we have

$$z_{m_0(\varepsilon)} = \alpha K(\varepsilon, \kappa_0(\varepsilon), \lambda_{\max}(\varepsilon))h_2(\varepsilon), \tag{4.6a}$$

$$w_{m_0(\varepsilon)} = \alpha\varepsilon^{-\frac{1}{2}}\phi_0(\varepsilon) + (L(\varepsilon) - \varepsilon^2\kappa_0(\varepsilon)^2 - \varepsilon\tau\lambda_{\max}(\varepsilon))^{\dagger}(-f_v(\bar{u}, \bar{v})z_{m_0(\varepsilon)}), \tag{4.6b}$$

where α is an arbitrary constant in $\mathbb{R}$. Since $(w_{m_0(\varepsilon)}, z_{m_0(\varepsilon)})$ satisfies (3.4), the standard bootstrap argument implies that $w_{m_0(\varepsilon)}$ and $z_{m_0(\varepsilon)}$ are C^∞-functions on I. At last we obtain (1.12) immediately from (1.11), which completes the proof of Main Theorem (b). □

REMARK 4.1. Using Lemma 3.1 (iv) and (2.9c), we have the limit of (4.6a) in $H^1(I)$ as $\varepsilon \downarrow 0$. Let the limit be denoted by z^*. From (4.35b) and (2.9c), we obtain

$$\|\text{the right-hand side of (4.6a)}\|_{L^\infty(I)} = O(\kappa_0(\varepsilon)^{-\frac{1}{2}}),$$

which, together with the fact $\kappa_0(\varepsilon) \to \infty$ as $\varepsilon \downarrow 0$, implies that z^* is a trivial function. Hence it follows, from (2.9a), that $w_{m_0(\varepsilon)}(x)$ converges to $\alpha c_0^* \delta(x - x^*)$ in $H^{-1}(I)$ as $\varepsilon \downarrow 0$.

4.2. A priori bound for eigenvalues in $\mathbb{C}_{\lambda_*}$. In this subsection, we show the uniformly boundedness of $\lambda \in \mathbb{C}_{\lambda_*}$ satisfying the singular dispersion relation (3.25).

We introduce a Sturm–Liouville operator:

$$T_0 \equiv -D\frac{d^2}{dx^2} + (f_u g_v - f_v g_u)/(-f_u)|_{\substack{u=U(x)\\v=V(x)}} \tag{4.7}$$

subject to the zero flux boundary condition. Let $\{\psi_n, \gamma_n\}_{n=0}^\infty$ be the complete orthonormal set and the eigenvalues associated with T_0. It is clear from (2.3) and (A.5) in §1 that the eigenvalues are strictly positive. It follows from the general theory of the Sturm–Liouville problems (see for instance [CH]), that $\{\psi_n\}_{n=0}^\infty$ is bounded in $C(\bar{I})$ and $\gamma_n = O(n^2)$ (as $n \uparrow \infty$). In view of Lemma 3.1 and (3.12), we have

$$K(0, \kappa, \lambda)\delta(x - x^*) = \sum_{n=0}^{\infty} \frac{\psi_n(x^*)\psi_n(x)}{\gamma_n + D\kappa^2 + \lambda} \quad \text{in} \quad C(\bar{I}). \tag{4.8}$$

In what follows we assume $\lambda_* < \gamma_0$ with replacing λ_* in Proposition 3.1 by a smaller one, if necessary. In order to treat (3.25), the following properties of H (see (3.27)) are useful.

LEMMA 4.1. *Under the same hypotheses of Lemma* 3.1, *we have* (i)–(iii).

(i) $H(\varepsilon,\kappa,\lambda)$ *is a continuous and uniformly bounded function of* $(\varepsilon,\kappa,\lambda)$ *with*

$$\widehat{\zeta}_0(0) - H(0,0,0) < 0. \tag{4.9}$$

(ii) $H(\varepsilon,\kappa,\lambda)$ *depends analytically on* $\lambda \in \overset{\circ}{\mathbb{C}}_{\lambda_*}$ *and real analytically on* $\kappa > 0$, *respectively. We have*

$$\frac{\partial H}{\partial \lambda}(\varepsilon,\kappa,\lambda) = -I(\varepsilon,\kappa,\lambda) - \varepsilon\tau J(\varepsilon,\kappa,\lambda), \tag{4.10a}$$

$$\frac{\partial H}{\partial \kappa}(\varepsilon,\kappa,\lambda) = -2\kappa\{D\cdot I(\varepsilon,\kappa,\lambda) + \varepsilon^2\cdot J(\varepsilon,\kappa,\lambda)\}, \tag{4.10b}$$

where $I(\varepsilon,\kappa,\lambda), J(\varepsilon,\kappa,\lambda)$ *are uniformly bounded functions defined by*

$$I(\varepsilon,\kappa,\lambda) \equiv (K(\varepsilon,\kappa,\lambda)K(\varepsilon,\kappa,\lambda)h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}, \tag{4.11a}$$

$$J(\varepsilon,\kappa,\lambda) \equiv (K(\varepsilon,\kappa,\lambda)S(\varepsilon,\kappa,\lambda)K(\varepsilon,\kappa,\lambda)h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}. \tag{4.11b}$$

(iii) *We have*

$$\begin{aligned} H(0,\kappa,\lambda) &= c_1^*c_2^*\cdot {}_{H^1(I)}\langle K(0,\kappa,\lambda)\delta(x-x^*), \delta(x-x^*)\rangle_{H^{-1}(I)} \\ &= \sum_{n=0}^{\infty} \frac{c_1^*c_2^*|\psi_n(x^*)|^2}{\gamma_n + D\kappa^2 + \lambda}, \end{aligned} \tag{4.12}$$

$$\begin{aligned} I(0,\kappa,\lambda) &= c_1^*c_2^*\cdot {}_{H^1(I)}\langle K(0,\kappa,\lambda)K(0,\kappa,\lambda)\delta(x-x^*), \delta(x-x^*)\rangle_{H^{-1}(I)} \\ &= c_1^*c_2^*\cdot (K(0,\kappa,\lambda)\delta(x-x^*), K(0,\kappa,\bar{\lambda})\delta(x-x^*))_{L^2(I)} \\ &= \sum_{n=0}^{\infty} \frac{c_1^*c_2^*|\psi_n(x^*)|^2}{(\gamma_n + D\kappa^2 + \lambda)^2}. \end{aligned} \tag{4.13}$$

Proof. Basically the above properties of H are the consequences from those of K in Lemma 3.1 except (4.9), (4.12), and (4.13). The inequality (4.9) comes from the stability of $\mathfrak{U}(\varepsilon)$ when $\tau = 1/\varepsilon$, which was proved in [NF] (see also [NM, Thm. 2.2]). Next, let $\varepsilon \downarrow 0$ in (3.27) and (4.11a), then by virtue of (2.10), (4.8) and the property of $K(\varepsilon,\kappa,\lambda)$ in Lemma 3.1 (iv), we obtain (4.12) and (4.13), respectively. □

Now we can give τ_*, which appeared in §§2 and 3 and in the assumptions of Theorem 4.1. We put

$$\begin{aligned} \tau_* &= c_1^*c_2^*\,\|K(0,0,-\lambda_*)\delta(x-x^*)\|_{L^2(I)}^2 \\ &= \sum_{n=0}^{\infty} \frac{c_1^*c_2^*|\psi_n(x^*)|^2}{(\gamma_n - \lambda_*)^2}. \end{aligned} \tag{4.14}$$

Here we used (4.8) to get the second equality of (4.14). Since $\lambda_* = \lambda_*(f,g,D)$, we can see that τ_* depends only on f, g and D. Our standing assumption for τ throughout this paper is given by

$$\tau_* < \tau. \tag{4.15}$$

Recalling (3.15), Lemma 4.1 is valid for any $\varepsilon \in (0,\varepsilon_1)$. Then we have the following a priori estimate for λ.

PROPOSITION 4.1. *There exists* $M_1 = M_1(f, g, D) > 0$ *such that, for any fixed* $\varepsilon \in (0, \varepsilon_1)$, *we have*

$$|\lambda| < M_1 \tag{4.16}$$

for any eigenvalue λ *of* (3.25) *in* $\mathbb{C}_{\lambda_*}$.

Proof. We take the real part and imaginary parts of (3.25), respectively,

$$\begin{aligned} \widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa^2 - \tau\Re\lambda - \Re H(\varepsilon, \kappa, \lambda) = 0, \\ -\tau\Im\lambda - \Im H(\varepsilon, \kappa, \lambda) = 0, \end{aligned}$$

which yields

$$\begin{aligned} \tau\Re\lambda = \widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa^2 - \Re H(\varepsilon, \kappa, \lambda) \le |\widehat{\zeta}_0(\varepsilon)| + |H(\varepsilon, \kappa, \lambda)|, \\ \tau|\Im\lambda| \le |H(\varepsilon, \kappa, \lambda)|. \end{aligned}$$

The conclusion follows from Proposition 2.1, Lemma 4.1 (i), and (4.15). □

REMARK 4.2. In view of the proof of Proposition 4.1, we see that the assumption (4.15) is too strong to obtain a priori bound, in fact, it suffices to assume that τ is strictly bounded away from zero. However, as we will see in §4.3, we need the condition (4.15) to guarantee that the eigenvalues of (3.25) in $\mathbb{C}_{\lambda_*}$ are real and hence there does not occur the instability of Hopf type.

4.3. Parameterization of eigenvalues in $\mathbb{C}_{\lambda_*}$. The objective of this subsection is to show that the distribution of eigenvalues of singular dispersion relation (3.25) is, as a function of κ, like a parabolic curve qualitatively. More precisely we shall prove the following:

(1) Every eigenvalue in $\mathbb{C}_{\lambda_*}$ must be real.

(2) All the eigenvalues in $\mathbb{C}_{\lambda_*}$ can be parameterized by κ. Especially nonnegative ones, which interest us most, are located on an ε-dependent interval $[\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$ with $\lim_{\varepsilon\downarrow 0} \overline{\kappa}_+(\varepsilon) = \infty$ (see Prop. 4.3). The eigenvalues outside of this interval, namely those for small κ or large κ, are negative.

We begin with the study of the asymptotic behavior of $K(\varepsilon, \kappa, \lambda)$ of (3.25) for large κ, which, combined with the term $-\varepsilon\kappa^2$, determines the behavior of eigenvalues.

We make a decomposition defined by

$$K(\varepsilon, \kappa, \lambda) = \widehat{K}(\kappa) + \overline{K}(\varepsilon, \kappa, \lambda), \tag{4.17}$$

for $\varepsilon \in [0, \varepsilon_*)$, $\kappa > 0$ and $\lambda \in \mathbb{C}_{\lambda_*}$, where the first and the second terms are defined below. We first introduce a sesquilinear form on $H^1(I)$:

$$\widehat{B}(\kappa)(z^1, z^2) \equiv D\,(z^1_x, z^2_x)_{L^2(I)} + D\kappa^2 (z^1, z^2)_{L^2(I)}, \tag{4.18}$$

for $z^1, z^2 \in H^1(I)$ and $\kappa > 0$. We also define an operator given by

$$\widehat{T}(\kappa) z \equiv \left(-D\frac{d^2}{dx^2} + D\kappa^2 \right) z, \tag{4.19}$$

for $z \in H^1(I)$ and $\kappa > 0$. Since $\widehat{B}(\kappa)$ is bounded and coercive, by using Lax–Milgram's theorem again, we see that $\widehat{T}(\kappa)$ has an inverse operator $\widehat{K}(\kappa)$, which is a bounded linear operator from $H^{-1}(I)$ to $H^1(I)$. In a loose way, it can be represented by

$$\widehat{K}(\kappa) = \left(-D\frac{d^2}{dx^2} + D\kappa^2\right)^{-1}, \qquad H^{-1}(I) \to H^1(I). \tag{4.20}$$

The second term of (4.17) is simply defined by

$$\overline{K}(\varepsilon, \kappa, \lambda) \equiv K(\varepsilon, \kappa, \lambda) - \widehat{K}(\kappa). \tag{4.21}$$

It turns out later that $\widehat{K}(\kappa)$ plays a dominant role for the study of the asymptotic behavior of $K(\varepsilon, \kappa, \lambda)$ as $\kappa \uparrow \infty$. The following Green's function for $\widehat{T}(\kappa)$ with the Neumann boundary condition is useful:

$$G(x, \xi, \kappa) = \begin{cases} \dfrac{\cosh \kappa x \cosh \kappa(1-\xi)}{D\kappa \sinh \kappa} & (0 \le x \le \xi \le 1), \\ \dfrac{\cosh \kappa(1-x) \cosh \kappa\xi}{D\kappa \sinh \kappa} & (0 \le \xi < x \le 1). \end{cases} \tag{4.22}$$

The following lemma can be checked by a direct calculation from (4.22).

LEMMA 4.2. *There are positive constants $M_2 = M_2(D)$ and $d = d(D)$ such that we have*

$$\max_{x,\xi \in \bar{I}} |G(x, \xi, \kappa)| < M_2 \cdot \kappa^{-1} \qquad (\kappa > 1), \tag{4.23}$$

$$\max_{x,\xi \in \bar{I}} \left|\frac{\partial G}{\partial \xi}(x, \xi, \kappa)\right| < M_2 \qquad (\kappa > 1), \tag{4.24}$$

$$\max_{x \in \bar{I}} \|G(x, \cdot, \kappa)\|_{L^2(I)} < M_2 \cdot \kappa^{-3/2} \qquad (\kappa > 1), \tag{4.25}$$

$$\max_{x \in \bar{I}} \|G(x, \cdot, \kappa)\|_{H^1(I)} < M_2 \cdot \kappa^{-1/2} \qquad (\kappa > 1), \tag{4.26}$$

$$\|G(x^*, \cdot, \kappa)\|^2_{L^2(I)} = \frac{1}{4D^2\kappa^3} + o(\exp(-d\kappa)) \quad (\text{as } \kappa \uparrow \infty), \tag{4.27}$$

where x^* is the position of the layer (see Theorem 2.1).

Since $H^1(I)$ can be imbedded into $L^2(I)$ or $L^\infty(I)$ continuously, we can regard an element of $\mathcal{B}(H^{-1}(I), H^1(I))$ as that of $\mathcal{B}(H^{-1}(I), L^2(I))$ or $\mathcal{B}(H^{-1}(I), L^\infty(I))$.

LEMMA 4.3. *There exist positive constants $M_i = M_i(D)$ $(i = 3, 4)$ such that we have*

$$\|\widehat{K}(\kappa)\|_{\mathcal{B}(H^{-1}(I), L^2(I))} < M_3 \cdot \kappa^{-1} \qquad (\kappa > 1), \tag{4.28a}$$

$$\|\widehat{K}(\kappa)\|_{\mathcal{B}(H^{-1}(I), L^\infty(I))} < M_3 \cdot \kappa^{-1/2} \qquad (\kappa > 1), \tag{4.28b}$$

and moreover, if $\{\Theta(\kappa)\}_{\kappa > 1}$ is a set of functions satisfying $\|\Theta(\kappa)\|_{L^2(I)} < \kappa^{-1}$, then we have

$$\|\widehat{K}(\kappa)\Theta(\kappa)\|_{L^2(I)} < M_4 \cdot \kappa^{-3} \qquad (\kappa > 1), \tag{4.29a}$$

$$\|\widehat{K}(\kappa)\Theta(\kappa)\|_{L^\infty(I)} < M_4 \cdot \kappa^{-5/2} \qquad (\kappa > 1). \tag{4.29b}$$

Proof. See Appendix 1.

LEMMA 4.4. *Let us assume in addition to the hypotheses of Lemma* 3.1, *that* $|\lambda| < M_1$ *is satisfied, where* M_1 *is the same one as in Proposition* 4.1. *Then there exists* $\kappa_* = \kappa_*(f, g, D) > 1$ *such that we have*

$$\overline{K}(\varepsilon, \kappa, \lambda) = \widehat{K}(\kappa)\overline{R}(\varepsilon, \kappa, \lambda)\widehat{K}(\kappa) \quad \text{for} \quad \kappa > \kappa_*, \tag{4.30}$$

where $\overline{R}(\varepsilon, \kappa, \lambda)$ *is a bounded linear operator in* $L^2(I)$. *There exists* $M_5 = M_5(f, g, D) > 0$ *such that*

$$\|\overline{R}(\varepsilon, \kappa, \lambda)\|_{\mathcal{B}(L^2(I), L^2(I))} < M_5, \tag{4.31}$$

$$\|\overline{K}(\varepsilon, \kappa, \lambda)\|_{\mathcal{B}(H^{-1}(I), L^2(I))} < M_5 \cdot \kappa^{-3}, \tag{4.32a}$$

$$\|\overline{K}(\varepsilon, \kappa, \lambda)\|_{\mathcal{B}(H^{-1}(I), L^\infty(I))} < M_5 \cdot \kappa^{-5/2}, \tag{4.32b}$$

hold for $\kappa > \kappa_*$.

Proof. See Appendix 2.

LEMMA 4.5. *We assume that* $\kappa > \kappa_*$ *in addition to the hypothesis of Lemma* 4.1, *where* κ_* *is the same one as in Lemma* 4.4. *Then there is* $M_6 = M_6(f, g, D) > 0$ *such that we have*

$$|H(\varepsilon, \kappa, \lambda)| < M_6 \cdot \kappa^{-1/2}, \tag{4.33}$$

$$|I(\varepsilon, \kappa, \lambda)| < M_6 \cdot \kappa^{-5/2}, \tag{4.34a}$$

$$|J(\varepsilon, \kappa, \lambda)| < M_6 \cdot \kappa^{-5/2}. \tag{4.34b}$$

Proof. Using (4.28) and (4.32), we obtain, from (4.17),

$$\|K(\varepsilon, \kappa, \lambda)\|_{\mathcal{B}(H^{-1}(I), L^2(I))} < (M_3 + M_5) \cdot \kappa^{-1}, \tag{4.35a}$$

$$\|K(\varepsilon, \kappa, \lambda)\|_{\mathcal{B}(H^{-1}(I), L^\infty(I))} < (M_3 + M_5) \cdot \kappa^{-1/2}, \tag{4.35b}$$

for $\kappa > \kappa_*$. In view of (3.27), we have

$$|H(\varepsilon, \kappa, \lambda)| \le \|K(\varepsilon, \kappa, \lambda)h_2(\varepsilon)\|_{L^\infty(I)} \cdot \|h_1(\varepsilon)\|_{L^1(I)}. \tag{4.36}$$

From Corollary 2.1, $h_i(\varepsilon)$ is uniformly bounded in $L^2(I)$ and also in $H^{-1}(I)$ $(i = 1, 2)$. Using this fact and (4.35b), we obtain (4.33) from (4.36). From (4.35a) and the boundedness of $S(\varepsilon, \kappa, \lambda)$ in Lemma 3.1, we have

$$\|K(\varepsilon, \kappa, \lambda)h_2(\varepsilon)\|_{L^2(I)} = O(\kappa^{-1}), \tag{4.37a}$$

$$\|S(\varepsilon, \kappa, \lambda)K(\varepsilon, \kappa, \lambda)h_2(\varepsilon)\|_{L^2(I)} = O(\kappa^{-1}). \tag{4.37b}$$

Making use of (4.29b) and (4.32b), we obtain, from (4.17),

$$\|K(\varepsilon, \kappa, \lambda)\Theta(\kappa)\|_{L^\infty(I)} = O(\kappa^{-5/2}), \tag{4.38}$$

where $\{\Theta(\kappa)\}$ is a set of functions satisfying $\|\Theta(\kappa)\|_{L^2(I)} < \kappa^{-1}$. Regarding each term of (4.37) to be $\Theta(\kappa)$, we obtain from (4.37) and (4.38)

$$\|K(\varepsilon, \kappa, \lambda)K(\varepsilon, \kappa, \lambda)h_2(\varepsilon)\|_{L^\infty(I)} = O(\kappa^{-5/2}),$$

$$\|K(\varepsilon, \kappa, \lambda)S(\varepsilon, \kappa, \lambda)K(\varepsilon, \kappa, \lambda)h_2(\varepsilon)\|_{L^\infty(I)} = O(\kappa^{-5/2}).$$

In view of the definitions of $I(\varepsilon, \kappa, \lambda)$ and $J(\varepsilon, \kappa, \lambda)$, we obtain (4.34) from the above inequalities and the uniformly boundedness of $h_1(\varepsilon)$ in $L^1(I)$. □

Before studying the κ-parameterization of solutions of (3.25), the following observation is basic in characterizing the behaviors of eigenvalues in $\mathbb{C}_{\lambda_*}$, which allows us to consider only real solutions of (3.25).

PROPOSITION 4.2. *Let* $\lambda \in \mathbb{C}_{\lambda_*}$ *satisfy* (3.25) *with some* $\varepsilon \in (0, \varepsilon_1)$ *and* $\kappa \in [0, \infty)$, *then* λ *must be a real number.*

Proof. See Appendix 3. □

It is crucial to know the location of zero eigenvalues of (3.25) and their dependency on ε to clarify the stability properties of the planar front. In fact, there exist two zeros of (3.25) at $\underline{\kappa}(\varepsilon)$ and $\overline{\kappa}(\varepsilon)$, where $\underline{\kappa}(\varepsilon)$ remains finite and $\overline{\kappa}(\varepsilon) \to \infty$ as $\varepsilon \downarrow 0$. It turns out that instability occurs on the band region $(\underline{\kappa}(\varepsilon), \overline{\kappa}(\varepsilon))$, and outside of this interval the associated eigenvalues become nonpositive. The smaller one $\underline{\kappa}(\varepsilon)$ can be controlled by studying the *reduced problem* of (3.25) (namely the limiting one of (3.25) as $\varepsilon \downarrow 0$). While more careful analysis is needed to handle the larger one $\overline{\kappa}(\varepsilon)$, and hence we will discuss it in §4.4. Nevertheless we can show that the eigenvalues of (3.25) become negative for sufficiently large κ as well as for small κ. Moreover we can parameterize all the nonnegative eigenvalues by κ, which we are most concerned with. Putting $\lambda = 0$ in (3.25), we have

$$F_0(\varepsilon, \kappa) = 0, \tag{4.39}$$

for $\varepsilon \in [0, \varepsilon_1)$ and $\kappa \in [0, \infty)$, where

$$F_0(\varepsilon, \kappa) \equiv F(\varepsilon, \kappa, 0) = \widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa^2 - H(\varepsilon, \kappa, 0). \tag{4.40}$$

Recalling that (3.26) is well defined up to $\varepsilon = 0$ from Lemma 4.1, we see from (4.40) and the property of H in Lemma 4.1 that

$$\frac{\partial F_0}{\partial \kappa}(0, \kappa) = 2\kappa D \cdot c_1^* c_2^* \cdot \|K(0, \kappa, 0)\delta(x - x^*)\|^2_{L^2(I)} > 0, \tag{4.41}$$

for $\kappa > 0$. On the other hand, by using (4.9) and (4.33), we have, from (4.40),

$$F_0(0, \kappa)|_{\kappa=0} < 0, \qquad F_0(0, \kappa)|_{\kappa=\infty} > 0. \tag{4.42}$$

From (4.41), (4.42) and the continuity of $F_0(0, \cdot)$, we see that there is a unique $\kappa = \underline{\kappa}(0)$ such that

$$F_0(0, \underline{\kappa}(0)) = \widehat{\zeta}_0(0) - H(0, \underline{\kappa}(0), 0) = 0. \tag{4.43}$$

Thanks to (4.43) and (4.41), we can apply the standard implicit function theorem to (4.39) at $(\varepsilon, \kappa) = (0, \underline{\kappa}(0))$ yielding a unique $\underline{\kappa}(\varepsilon)$ with

$$F_0(\varepsilon, \underline{\kappa}(\varepsilon)) = 0, \tag{4.44a}$$

$$\underline{\kappa}(\varepsilon) \to \underline{\kappa}(0) \in (0, \infty) \quad (\text{as} \quad \varepsilon \downarrow 0), \tag{4.44b}$$

$$F_0(\varepsilon, \kappa) < 0 \quad \text{for any} \quad \kappa \in [0, \underline{\kappa}(\varepsilon)), \tag{4.44c}$$

for any $\varepsilon \in (0, \varepsilon_1)$. We replaced ε_1 in Proposition 4.2 by a smaller one, if necessary.

In view of (3.26) and (4.33), we easily see that when $\kappa = O(\varepsilon^{-1/2})$ the term $-\varepsilon\kappa^2$ becomes $O(1)$ and hence it is enough to compete with $\widehat{\zeta}_0(\varepsilon)$ to determine the sign of the eigenvalue λ. This motivates us to take the following special scaling of κ given by

$$\overline{\kappa}_+(\varepsilon) \equiv \sigma \cdot \varepsilon^{-\frac{1}{2}}, \tag{4.45}$$

where σ is a constant satisfying

$$\widehat{\zeta}_0(0) < \sigma^2 < \widehat{\zeta}_0(0) + \tfrac{1}{2}\tau\lambda_*. \tag{4.46}$$

Replacing ε_1 by a smaller one again, we have the following two lemmas.

LEMMA 4.6. *For any fixed* $\varepsilon \in (0, \varepsilon_1)$, *it holds that*

$$\frac{\partial F}{\partial \lambda}(\varepsilon, \kappa, \lambda) < -\frac{1}{2}(\tau - \tau_*) < 0 \tag{4.47}$$

for $(\kappa, \lambda) \in [0, \infty) \times [-\lambda_*, M_1)$ *satisfying* (3.6).

LEMMA 4.7. *We have, for* $\varepsilon \in (0, \varepsilon_1)$ *and* $\kappa \in [\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$,

$$F(\varepsilon, \kappa, \lambda)|_{\lambda=-\lambda_*} \geq \tfrac{1}{2}\lambda_*(\tau - \tau_*) > 0,$$
$$F(\varepsilon, \kappa, \lambda)|_{\lambda=M_1} < 0.$$

Proof of Lemma 4.6. Let us assume the contrary; then we have a sequence $\{(\varepsilon_i, \kappa_i, \lambda_i)\}_{i=1}^{\infty}$ such that $\varepsilon_i \to 0$ as $i \uparrow \infty$, and each $(\varepsilon_i, \kappa_i, \lambda_i)$ satisfies (3.6), (3.25), and

$$\frac{\partial F}{\partial \lambda}(\varepsilon_i, \kappa_i, \lambda_i) \geq -\frac{1}{2}(\tau - \tau_*).$$

Using (3.26) and Lemma 4.1, we rewrite it as follows:

$$-\tau + I(\varepsilon_i, \kappa_i, \lambda_i) + \varepsilon_i \tau J(\varepsilon_i, \kappa_i, \lambda_i) \geq -\tfrac{1}{2}(\tau - \tau_*). \tag{4.48}$$

First, we consider the case in which $\{\kappa_i\}$ is unbounded. Without loss of generality, we can assume that $\kappa_i \to \infty$ as $i \uparrow \infty$. Let $i \uparrow \infty$ in (4.48). Then from (4.34) we obtain the inequality: $-\tau \geq -\frac{1}{2}(\tau - \tau_*)$, which contradicts the positiveness of τ. Next, we consider the case in which $\{\kappa_i\}$ is bounded. Since $\{\lambda_i\}$ is bounded, we can take a subsequence such that $\kappa_i \to \kappa_0 \in [0, \infty)$, and $\lambda_i \to \lambda_0 \in [-\lambda_*, M_1]$. By virtue of Lemma 4.1 and the definition of τ_*, we have

$$\lim_{i \uparrow \infty} \{I(\varepsilon_i, \kappa_i, \lambda_i) + \varepsilon_i \tau J(\varepsilon_i, \kappa_i, \lambda_i)\}$$
$$= \sum_{n=0}^{\infty} \frac{c_1^* c_2^* |\psi_n(x^*)|^2}{(\gamma_n + D\kappa_0^2 + \lambda_0)^2} < \sum_{n=0}^{\infty} \frac{c_1^* c_2^* |\psi_n(x^*)|^2}{(\gamma_n - \lambda_*)^2} = \tau_*.$$

Letting $i \uparrow \infty$ in (4.48) and using the above inequality, we have $-\tau + \tau_* \geq -\frac{1}{2}(\tau - \tau_*)$, which contradicts (4.15). We complete the proof of Lemma 4.6. □

Proof of Lemma 4.7. We can obtain the second inequality from the definition of $F(\varepsilon, \kappa, \lambda)$ and the boundedness of $H(\varepsilon, \kappa, \lambda)$ in Lemma 4.1, and (4.15), if we replace M_1 by a larger one, if necessary. It is enough for us to prove the first inequality. Let us assume the contrary, then there exists $\{(\varepsilon_i, \kappa_i)\}_{i=0}^{\infty}$ such that we have $\varepsilon_i \to 0$ as $i \uparrow \infty$, $\kappa_i \in [\underline{\kappa}(\varepsilon_i), \overline{\kappa}_+(\varepsilon_i))$ and

$$F(\varepsilon_i, \kappa_i, -\lambda_*) < \tfrac{1}{2}\lambda_*(\tau - \tau_*), \tag{4.49}$$

which is equivalent to

$$\widehat{\zeta}_0(\varepsilon_i) - \varepsilon_i \kappa_i^2 + \tau\lambda_* - H(\varepsilon_i, \kappa_i, -\lambda_*) < \tfrac{1}{2}\lambda_*(\tau - \tau_*). \tag{4.50}$$

From $\kappa_i < \overline{\kappa}_+(\varepsilon_i)$, (4.45) and (4.46), we see that the first three terms on the left-hand side of (4.50) can be estimated from below:

$$\varliminf_{i \uparrow \infty} \{\widehat{\zeta}_0(\varepsilon_i) - \varepsilon_i \kappa_i^2 + \tau\lambda_*\} \geq \tfrac{1}{2}\tau\lambda_*. \tag{4.51}$$

First we consider the case in which $\{\kappa_i\}$ is not bounded. By taking a subsequence, we can assume $\kappa_i \to \infty$ as $i \uparrow \infty$. We let $i \uparrow \infty$ in (4.50). Then from (4.33), we have the inequality: $\frac{1}{2}\tau\lambda_* \le \frac{1}{2}\lambda_*(\tau - \tau_*)$, which contradicts the positiveness of τ_*. We next consider the case in which $\{\kappa_i\}$ is bounded. By taking a subsequence, we can assume that $\kappa_i \to \kappa_0 \ge \underline{\kappa}(0)$ as $i \uparrow \infty$. Hence we have

$$\begin{aligned} &\lim_{i\uparrow\infty} F(\varepsilon_i, \kappa_i, -\lambda_*) \\ &\quad = \widehat{\zeta}_0(0) + \tau\lambda_* - H(0, \kappa_0, -\lambda_*) \\ &\quad = H(0, \underline{\kappa}(0), 0) - H(0, \kappa_0, -\lambda_*) + \tau\lambda_* \quad \text{(from (4.43))}. \end{aligned} \tag{4.52}$$

Using (4.12) and (4.14), we obtain

$$\begin{aligned} &H(0, \underline{\kappa}(0), 0) - H(0, \kappa_0, -\lambda_*) \\ &\quad = \{-\lambda_* + D(\kappa_0^2 - \underline{\kappa}(0)^2)\} \sum_{n=0}^{\infty} \frac{c_1^* c_2^* |\psi_n(x^*)|^2}{\{\gamma_n + D\underline{\kappa}(0)^2\}\{\gamma_n + D\kappa_0^2 - \lambda_*\}} \ge -\lambda_*\tau_*. \end{aligned}$$

Putting together (4.52) and the above inequality, we obtain

$$\lim_{i\uparrow\infty} F(\varepsilon_i, \kappa_i, -\lambda_*) \ge -\lambda_*\tau_* + \tau\lambda_* = \lambda_*(\tau - \tau_*).$$

On the other hand, we have from (4.49),

$$\overline{\lim_{i\uparrow\infty}} F(\varepsilon_i, \kappa_i, -\lambda_*) \le \tfrac{1}{2}\lambda_*(\tau - \tau_*).$$

Combining the above two inequalities, we get $\lambda_*(\tau - \tau_*) \le \frac{1}{2}\lambda_*(\tau - \tau_*)$, which contradicts (4.15). Thus we complete the proof of Lemma 4.7. □

PROPOSITION 4.3. *For any fixed $\varepsilon \in (0, \varepsilon_1)$, the statements* (1) *and* (2) *hold true, where $\underline{\kappa}(\varepsilon)$ and $\overline{\kappa}_+(\varepsilon)$ are given in* (4.44) *and* (4.45), *respectively.*

(1) *If $\lambda \in (-\lambda_*, M_1)$ satisfies* (3.25) *with some $\kappa \notin [\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$, then $\lambda < 0$.*

(2) *A unique $\lambda = \widetilde{\lambda}(\varepsilon, \kappa)$ satisfies* (3.25) *for each $\kappa \in [\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$, where $\widetilde{\lambda}(\varepsilon, \kappa)$ is a real-valued function of C^∞-class for κ and the derivative is given by*

$$\frac{\partial\widetilde{\lambda}}{\partial\kappa}(\varepsilon, \kappa) = -\frac{\partial F}{\partial\kappa}(\varepsilon, \kappa, \widetilde{\lambda}(\varepsilon, \kappa)) \Big/ \frac{\partial F}{\partial\lambda}(\varepsilon, \kappa, \widetilde{\lambda}(\varepsilon, \kappa)). \tag{4.53}$$

Moreover, $\widetilde{\lambda}(\varepsilon, \kappa)$ always remains in $(-\lambda_*, M_1)$.

Proof of Proposition 4.3. As for (1), we see, from the assumption, that κ must lie in $[0, \underline{\kappa}(\varepsilon))$ or $[\overline{\kappa}_+(\varepsilon), \infty)$. In view of (4.40) and (4.44c), we have

$$F(\varepsilon, \kappa, 0) < 0 \quad \text{for any} \quad \kappa \in [0, \underline{\kappa}(\varepsilon)),$$

which, combined with (4.47), leads to

$$F(\varepsilon, \kappa, \lambda) < 0 \quad \text{for any } \lambda \ge 0 \text{ and } \kappa \in [0, \underline{\kappa}(\varepsilon)).$$

We have the desired result when $\kappa \in [0, \underline{\kappa}(\varepsilon))$. Next we consider the case in which $\kappa \in [\overline{\kappa}_+(\varepsilon), \infty)$. In view of the definition of $\overline{\kappa}_+(\varepsilon)$, we have the inequality: $\widehat{\zeta}_0(0) < \sigma^2 < \varepsilon\kappa^2$, which combined with (3.26) yields

$$\begin{aligned} \tau\lambda &= \widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa^2 - H(\varepsilon, \kappa, \lambda) \\ &< \{\widehat{\zeta}_0(\varepsilon) - \sigma^2\} - H(\varepsilon, \kappa, \lambda). \end{aligned} \tag{4.54}$$

We see that $H(\varepsilon,\kappa,\lambda)$ converges to zero as $\varepsilon \downarrow 0$ because of (4.33) and $\kappa \geq \overline{\kappa}_+(\varepsilon)$, while $\{\widehat{\zeta}_0(\varepsilon) - \sigma^2\}$ converges to a negative value $\widehat{\zeta}_0(0) - \sigma^2$. Hence we have $\lambda < 0$ for $\varepsilon \in (0, \varepsilon_1)$ and $\kappa \geq \overline{\kappa}_+(\varepsilon)$. It suffices to prove (2). Making use of Lemmas 4.6 and 4.7, we have a unique $\lambda \in (-\lambda_*, M_1)$ which satisfies (3.25) for each $\kappa \in [\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$. Let this λ be denoted by $\widetilde{\lambda}(\varepsilon,\kappa)$. Lemma 4.6 allows us to apply the standard implicit function theorem to (3.25) around $(\varepsilon,\kappa,\lambda) = (\varepsilon,\kappa,\widetilde{\lambda}(\varepsilon,\kappa))$. Since $F(\varepsilon,\kappa,\lambda)$ depends continuously on ε and real analytically on κ (in particular, it is of $C^\infty-$class for κ), we see that $\widetilde{\lambda}(\varepsilon,\kappa)$ is continuous for ε, of C^∞-class for κ, and the formula (4.53) holds. □

4.4. Asymptotic characterization of the fastest growth wave number and proof of Theorem 4.1. The singular dispersion relation (3.25) has strictly positive solutions somewhere on the interval $[\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$ of Proposition 4.3. In fact, suppose we take $\kappa = \varepsilon^{-\vartheta}$ $(0 < \vartheta < \frac{1}{2})$, then we see from (3.25) and (4.33)

$$\begin{aligned} \tau\widetilde{\lambda}(\varepsilon,\varepsilon^{-\vartheta}) &= \widehat{\zeta}_0(\varepsilon) - \varepsilon(\varepsilon^{-\vartheta})^2 - H(\varepsilon,\varepsilon^{-\vartheta},\widetilde{\lambda}(\varepsilon,\varepsilon^{-\vartheta})) \\ &\to \widehat{\zeta}_0(0) > 0 \qquad (\text{as} \quad \varepsilon \downarrow 0). \end{aligned} \tag{4.55}$$

Since $\widetilde{\lambda}(\varepsilon,\underline{\kappa}(\varepsilon)) = 0$ and $\widetilde{\lambda}(\varepsilon,\overline{\kappa}_+(\varepsilon)) < 0$, it is clear that a continuous function $\widetilde{\lambda}(\varepsilon,\cdot)$ has a positive maximum in $(\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$.

We study in this subsection the asymptotic behaviors of the largest eigenvalue $\lambda(\varepsilon)$ of (3.25) and the associated wave number $\kappa(\varepsilon)$ when $\varepsilon \to 0$.

There is at least one local maximizer of $\widetilde{\lambda}(\varepsilon,\cdot)$ in $(\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$ for any fixed $\varepsilon \in (0, \varepsilon_1)$. Let $\kappa(\varepsilon)$ be an arbitrary one of them and let us put $\lambda(\varepsilon) \equiv \widetilde{\lambda}(\varepsilon,\kappa(\varepsilon))$. From the definition of $\kappa(\varepsilon)$ and that of $\lambda(\varepsilon)$, it follows that

$$F(\varepsilon,\kappa(\varepsilon),\lambda(\varepsilon)) = 0, \tag{4.56a}$$

$$\frac{\partial\widetilde{\lambda}}{\partial\kappa}(\varepsilon,\kappa(\varepsilon)) = 0. \tag{4.56b}$$

In view of Proposition 4.3, we can rewrite (4.56b) as follows:

$$\frac{\partial F}{\partial\kappa}(\varepsilon,\kappa(\varepsilon),\lambda(\varepsilon)) = 0, \tag{4.57}$$

or equivalently (see (4.10b)),

$$\varepsilon - D \cdot I(\varepsilon,\kappa(\varepsilon),\lambda(\varepsilon)) - \varepsilon^2 \cdot J(\varepsilon,\kappa(\varepsilon),\lambda(\varepsilon)) = 0. \tag{4.58}$$

REMARK 4.3. The relation (4.56b) implies that $\kappa(\varepsilon)$ is a stationary point of $\widetilde{\lambda}(\varepsilon,\cdot)$. Although it is not a sufficient condition for $\kappa(\varepsilon)$ to be a local maximizer, we will be able to obtain enough information to control the asymptotic behaviors of $\kappa(\varepsilon)$ and $\lambda(\varepsilon)$ as ε tends to zero.

PROPOSITION 4.4. $\kappa(\varepsilon)$ $(\in (\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon)))$ *satisfies*

$$\begin{aligned} &\overline{\lim_{\varepsilon\downarrow 0}}\, \varepsilon\kappa(\varepsilon)^2 \leq \sigma^2, \\ &\kappa(\varepsilon) \to \infty \quad \textit{as} \quad \varepsilon \downarrow 0, \end{aligned}$$

while $\lambda(\varepsilon)$ *remains in* $(-\lambda_*, M_1)$ *for any* $\varepsilon \in (0, \varepsilon_1)$. M_1 *is the same constant as in Proposition* 4.1.

Proof. The boundedness of $\lambda(\varepsilon)$ is a direct consequence of that of $\widetilde{\lambda}(\varepsilon, \cdot)$ in Proposition 4.3. We can obtain the first inequality from $\kappa(\varepsilon) < \overline{\kappa}_+(\varepsilon)$ and the definition of $\overline{\kappa}_+(\varepsilon)$. It is enough to prove $\kappa(\varepsilon) \to \infty$ as $\varepsilon \downarrow 0$. Let us assume the contrary; then we have a sequence $\{\varepsilon_n\}_{n=1}^{\infty}$ such that $\varepsilon_n \to 0$ as $n \uparrow \infty$, $\{\kappa(\varepsilon_n)\}_{n=1}^{\infty}$ remains bounded and that

$$\varepsilon_n - D \cdot I(\varepsilon_n, \kappa(\varepsilon_n), \lambda(\varepsilon_n)) - \varepsilon_n^2 \cdot J(\varepsilon_n, \kappa(\varepsilon_n), \lambda(\varepsilon_n)) = 0. \tag{4.59}$$

By taking a subsequence, we can assume

$$\kappa(\varepsilon) \to \kappa_0 \in [\underline{\kappa}(0), \infty), \qquad \lambda(\varepsilon_n) \to \lambda_0 \in [-\lambda_*, M_1],$$

as $n \uparrow \infty$. In view of Lemma 4.1, we have

$$D \cdot I(\varepsilon_n, \kappa(\varepsilon_n), \lambda(\varepsilon_n)) \to D c_1^* c_2^* \| K(0, \kappa_0, \lambda_0) \delta(x - x^*) \|_{L^2(I)}^2 > 0. \tag{4.60}$$

Since $J(\varepsilon_n, \kappa(\varepsilon_n), \lambda(\varepsilon_n))$ remains bounded, the first and the second terms of the right-hand side of (4.59) converge to zero as $n \uparrow \infty$. (4.60) is not consistent with (4.59). This is a contradiction and completes the proof. □

We study the asymptotic behaviors of $\kappa(\varepsilon)$ and $\lambda(\varepsilon)$ more accurately. It turns out that the first and the second terms of (4.58) play a dominant role in the limit of $\varepsilon \downarrow 0$ and $\kappa \uparrow \infty$. Without loss of generality, we can assume $\kappa_* < \kappa(\varepsilon)$ for $\varepsilon \in (0, \varepsilon_1)$. Recall that κ_* is a constant appeared in Lemma 4.4. In view of Lemma 4.5 and Proposition 4.4, we have

$$\varepsilon^2 |J(\varepsilon, \kappa, \lambda)| \le \varepsilon^2 M_6 \kappa(\varepsilon)^{-5/2} \le \sigma^4 M_6 \kappa(\varepsilon)^{-13/2}, \tag{4.61}$$

for $\varepsilon \in (0, \varepsilon_1)$. In order to study the second term of the right-hand side of (4.58), we use the next lemma.

LEMMA 4.8. *We assume that* $0 < \varepsilon < \varepsilon_1$, $\kappa_* < \kappa$, *and* $|\lambda| < M_1$. *Then we have*

$$|I(\varepsilon, \kappa, \lambda) - (\widehat{K}(\kappa) h_2(\varepsilon), \widehat{K}(\kappa) h_1(\varepsilon))_{L^2(I)}| < M_7 \cdot \kappa^{-4}, \tag{4.62}$$

where $M_7 = M_7(f, g, D) > 0$, *and* κ_* *is the same constant as in Lemma* 4.4.

Proof. In view of (4.17) and the definition of $I(\varepsilon, \kappa, \lambda)$, we see that $I(\varepsilon, \kappa, \lambda)$ can be rewritten as follows:

$$I(\varepsilon, \kappa, \lambda) \equiv \sum_{j=1}^{4} I_j(\varepsilon, \kappa, \lambda), \tag{4.63}$$

where

$$\begin{aligned}
I_1(\varepsilon, \kappa, \lambda) &\equiv (\widehat{K}(\kappa) \widehat{K}(\kappa) h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}, \\
I_2(\varepsilon, \kappa, \lambda) &\equiv (\widehat{K}(\kappa) \overline{K}(\varepsilon, \kappa, \lambda) h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}, \\
I_3(\varepsilon, \kappa, \lambda) &\equiv (\overline{K}(\varepsilon, \kappa, \lambda) \widehat{K}(\kappa) h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}, \\
I_4(\varepsilon, \kappa, \lambda) &\equiv (\overline{K}(\varepsilon, \kappa, \lambda) \overline{K}(\varepsilon, \kappa, \lambda) h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}.
\end{aligned}$$

Using (4.30) and the fact that $\widehat{K}(\kappa)$ is a selfadjoint operator, we have

$$I_1(\varepsilon,\kappa,\lambda) = (\widehat{K}(\kappa)h_2(\varepsilon), \widehat{K}(\kappa)h_1(\varepsilon))_{L^2(I)},$$
$$I_2(\varepsilon,\kappa,\lambda) = (\overline{K}(\varepsilon,\kappa,\lambda)h_2(\varepsilon), \widehat{K}(\kappa)h_1(\varepsilon))_{L^2(I)},$$
$$I_3(\varepsilon,\kappa,\lambda) = (\overline{R}(\varepsilon,\kappa,\lambda)\widehat{K}(\kappa)\widehat{K}(\kappa)h_2(\varepsilon), \widehat{K}(\kappa)h_1(\varepsilon))_{L^2(I)},$$
$$I_4(\varepsilon,\kappa,\lambda) = (\overline{R}(\varepsilon,\kappa,\lambda)\widehat{K}(\kappa)\overline{K}(\varepsilon,\kappa,\lambda)h_2(\varepsilon), \widehat{K}(\kappa)h_1(\varepsilon))_{L^2(I)}.$$

By virtue of Schwarz's inequality, the estimates in Lemmas 4.3 and 4.4, and Corollary 2.1, we obtain

$$|I_2(\varepsilon,\kappa,\lambda)| < (B_*)^2 M_3 M_5 \cdot \kappa^{-4},$$
$$|I_3(\varepsilon,\kappa,\lambda)| < (B_*)^2 (M_3)^2 M_4 M_5 \cdot \kappa^{-4},$$
$$|I_4(\varepsilon,\kappa,\lambda)| < (B_*)^2 M_3 M_4 (M_5)^2 \cdot \kappa^{-6}.$$

Substituting these into (4.63), we obtain the desired result. □

Using (4.61) and (4.62), we have, from (4.58),

$$|\varepsilon - D(\widehat{K}(\kappa(\varepsilon))h_2(\varepsilon), \widehat{K}(\kappa(\varepsilon))h_1(\varepsilon))_{L^2(I)}| < M_8 \cdot \kappa(\varepsilon)^{-4}, \tag{4.64}$$

where $M_8 = M_8(f,g,D) > 0$.

In order to obtain the asymptotic form of the left-hand side of (4.64), we need preparations. We introduce two positive constants $\theta_1, \theta_2 \in (0,1)$, which will be specified later, and divide $I = (0,1)$ into two subintervals:

$$I = I^1_{\text{in}}(\varepsilon) \cup I^1_{\text{out}}(\varepsilon), \tag{4.65a}$$

for $\varepsilon \in (0, \varepsilon_1)$, where

$$I^1_{\text{in}}(\varepsilon) = (x^* - \varepsilon^{\theta_1}, x^* + \varepsilon^{\theta_1}), \qquad I^1_{\text{out}}(\varepsilon) = I \setminus I^1_{\text{in}}(\varepsilon), \tag{4.65b}$$

and

$$I = I^2_{\text{in}}(\kappa) \cup I^2_{\text{out}}(\kappa), \tag{4.66a}$$

for $\kappa \in (\kappa_*, \infty)$, where

$$I^2_{\text{in}}(\kappa) = (x^* - \kappa^{-\theta_2}, x^* + \kappa^{-\theta_2}), \qquad I^2_{\text{out}}(\kappa) = I \setminus I^2_{in}(\kappa). \tag{4.66b}$$

Let $c_i(\varepsilon)$ be defined by

$$c_i(\varepsilon) \equiv \int_{I^1_{\text{in}}(\varepsilon)} h_i(x)dx \qquad (i = 1,2). \tag{4.67}$$

LEMMA 4.9. *$c_i(\varepsilon) \to c_i^*$ as $\varepsilon \downarrow 0$, $(i = 1,2)$.*

Proof. For $i = 1,2$, we have

$$\int_{I^1_{\text{in}}(\varepsilon)} h_i(x,\varepsilon)dx = \int_I h_i(x,\varepsilon)dx - \int_{I^1_{\text{out}}(\varepsilon)} h_i(x,\varepsilon)dx. \tag{4.68}$$

Since we see that the first term of the right-hand side of (4.68) converges to c_i^* from (2.10), it suffices to show the second term converges to zero. Using (2.12), we have

$$\begin{aligned}\int_{I^1_{\text{out}}(\varepsilon)} |h_i(x,\varepsilon)|dx &= \left(\int_{\varepsilon^{\theta_1-1}}^{(1-x^*)/\varepsilon} + \int_{-x^*/\varepsilon}^{-\varepsilon^{\theta_1-1}}\right) \varepsilon\,|h_i(\varepsilon\eta + x^*, \varepsilon)|\,d\eta \\ &\le \left(\int_{\varepsilon^{\theta_1-1}}^{\infty} + \int_{-\infty}^{-\varepsilon^{\theta_1-1}}\right) B\exp(-\beta|\eta|)d\eta \\ &\le (2B/\beta)\cdot\exp(-\beta\varepsilon^{\theta_1-1}),\end{aligned} \tag{4.69}$$

which completes the proof. □

LEMMA 4.10. *Under the same hypothesis as in Lemma* 4.8, *we have*

$$(\widehat{K}(\kappa)h_i(\varepsilon))(x) = c_i(\varepsilon)\{G(x,x^*,\kappa) + r_i(\varepsilon,x,\kappa)\} + p_i(\varepsilon,x,\kappa) \tag{4.70}$$

for $x \in I$ $(i = 1,2)$, *where* $r_i(\varepsilon,\cdot,\kappa)$, $p_i(\varepsilon,\cdot,\kappa)$ *are functions with*

$$\|r_i(\varepsilon,\cdot,\kappa)\|_{L^\infty(I)} \le M_2 \cdot \varepsilon^{\theta_1} \qquad (i = 1,2), \tag{4.71a}$$

$$\|p_i(\varepsilon,\cdot,\kappa)\|_{L^\infty(I)} \le M_2 \cdot \exp(-\beta\varepsilon^{\theta_1-1}) \qquad (i = 1,2). \tag{4.71b}$$

Proof. Using (4.65a) and Green's function of $\widehat{T}(\kappa)$, we have

$$(\widehat{K}(\kappa)h_i(\varepsilon))(x) = \left(\int_{I^1_{\text{in}}(\varepsilon)} + \int_{I^1_{\text{out}}(\varepsilon)}\right) G(x,\xi,\kappa)h_i(\xi,\varepsilon)d\xi,$$

for $x \in I$ $(i = 1,2)$. From (4.24), we have

$$|G(x,\xi,\kappa) - G(x,x^*,\kappa)| \le M_2|\xi - x^*| < M_2\varepsilon^{\theta_1},$$

for $x \in I$ and $\xi \in I^1_{in}(\varepsilon)$. We rewrite it as follows:

$$G(x,x^*,\kappa) - M_2\varepsilon^{\theta_1} < G(x,\xi,\kappa) < G(x,x^*,\kappa) + M_2\varepsilon^{\theta_1},$$

for $x \in I$ and $\xi \in I^1_{in}(\varepsilon)$. We multiply each side of the above inequality by $h_i(\xi,\varepsilon)$ and integrate it over $I^1_{in}(\varepsilon)$, then using (4.67), we obtain

$$\begin{aligned} c_i(\varepsilon)\{G(x,x^*,\kappa) - M_2\varepsilon^{\theta_1}\} &< \int_{I^1_{\text{in}}(\varepsilon)} G(x,\xi,\kappa)h_i(\xi,\varepsilon)d\xi \\ &< c_i(\varepsilon)\{G(x,x^*,\kappa) + M_2\varepsilon^{\theta_1}\}, \end{aligned}$$

for $x \in I$ and $i = 1,2$. The above inequality implies that, if we write

$$\int_{I^1_{\text{in}}(\varepsilon)} G(x,\xi,\kappa)h_i(\xi,\varepsilon)d\xi = c_i(\varepsilon)\{G(x,x^*,\kappa) + r_i(\varepsilon,x,\kappa)\},$$

then $r_i(\varepsilon,\cdot,\kappa) \in C(\bar{I})$ satisfies (4.71a) for $i = 1,2$. It suffices to show

$$p_i(\varepsilon,x,\kappa) \equiv \int_{I^1_{\text{out}}(\varepsilon)} G(x,\xi,\kappa)h_i(\xi,\varepsilon)d\xi \tag{4.72}$$

satisfies (4.71b). Indeed we have, from (4.69) and (4.23),

$$|\text{the right-hand side of (4.72)}| \le M_2/\kappa^* \cdot (2B/\beta)\exp(-\beta\varepsilon^{\theta_1-1}),$$

for $\kappa > \kappa^*$. Replacing M_2 with a larger one, if necessary, we obtain the desired result. □

LEMMA 4.11. *Under the same hypothesis as in Lemma* 4.8, *we have*

$$\left|(\widehat{K}(\kappa)h_2(\varepsilon), \widehat{K}(\kappa)h_1(\varepsilon))_{L^2(I)} - \frac{c_1(\varepsilon)c_2(\varepsilon)}{4D^2\kappa^3}\right| < M_9\left\{\varepsilon^{\theta_1}\kappa^{-(1+\theta_2)} + \varepsilon^{2\theta_1} + \exp(-d\kappa)\right\}, \tag{4.73}$$

where d *is the same one as in Lemma* 4.2, *and* $M_9 = M_9(f, g, D) > 0$.

Proof. Making use of (4.70), we have

$$\begin{aligned}(\widehat{K}(\kappa)h_2(\varepsilon))(x) \cdot (\widehat{K}(\kappa)h_1(\varepsilon))(x) \\ = [c_1(\varepsilon)\{G(x, x^*, \kappa) + r_1(\varepsilon, x, \kappa)\} + p_1(\varepsilon, x, \kappa)] \\ \times [c_2(\varepsilon)\{G(x, x^*, \kappa) + r_2(\varepsilon, x, \kappa)\} + p_2(\varepsilon, x, \kappa)],\end{aligned} \tag{4.74}$$

for $x \in I$. Using (4.71b), we obtain

$$\text{(the right-hand side of (4.74))} = c_1(\varepsilon)c_2(\varepsilon)\widehat{g}(\varepsilon, x, \kappa) + p(\varepsilon, x, \kappa),$$

for $x \in I$, where

$$\widehat{g}(\varepsilon, x, \kappa) \equiv \{G(x, x^*, \kappa) + r_1(\varepsilon, x, \kappa)\}\{G(x, x^*, \kappa) + r_2(\varepsilon, x, \kappa)\}$$

and $p(\varepsilon, \cdot, \kappa)$ is a function of x which satisfies

$$\|p(\varepsilon, \cdot, \kappa)\|_{L^\infty(I)} < M_2 \exp(-\beta\varepsilon^{\theta_1 - 1}).$$

We replaced M_2 with a larger one, if necessary. It suffices to show that

$$\int_I \widehat{g}(\varepsilon, x, \kappa)dx - \frac{c_1(\varepsilon)c_2(\varepsilon)}{4D^2\kappa^3}$$

is estimated by the right-hand side of (4.73). We have

$$\int_I \widehat{g}(\varepsilon, x, \kappa)dx = \overline{g}_1(\kappa) + \overline{g}_2(\varepsilon, \kappa) + \overline{g}_3(\varepsilon, \kappa) + \overline{g}_4(\varepsilon, \kappa),$$

where

$$\begin{aligned}\overline{g}_1(\varepsilon, \kappa) &\equiv \int_I |G(x, x^*, \kappa)|^2 dx, \\ \overline{g}_2(\varepsilon, \kappa) &\equiv \int_{I^2_{\text{in}}(\kappa)} G(x, x^*, \kappa)\{r_1(\varepsilon, x, \kappa) + r_2(\varepsilon, x, \kappa)\}dx, \\ \overline{g}_3(\varepsilon, \kappa) &\equiv \int_{I^2_{\text{out}}(\kappa)} G(x, x^*, \kappa)\{r_1(\varepsilon, x, \kappa) + r_2(\varepsilon, x, \kappa)\}dx, \\ \overline{g}_4(\varepsilon, \kappa) &\equiv \int_I r_1(\varepsilon, x, \kappa)r_2(\varepsilon, x, \kappa)dx.\end{aligned}$$

We have, from (4.27),

$$\overline{g}_1(\varepsilon, \kappa) = \frac{1}{4D^2\kappa^3} + o(\exp(-d\kappa)).$$

From (4.23), (4.66b), and (4.71a), we obtain

$$|\overline{g}_2(\varepsilon,\kappa)| < 2\kappa^{-\theta_2}\cdot M_2\kappa^{-1}\cdot 2M_2\varepsilon^{\theta_1} = 4(M_2)^2\varepsilon^{\theta_1}\kappa^{-(1+\theta_2)}.$$

By a direct calculation from (4.22), we have

$$\int_{I^2_{\mathrm{out}}(\kappa)} |G(x,x^*,\kappa)|dx = O(\kappa^{-2}\exp(-\kappa^{1-\theta_2})).$$

Using (4.71a) and the above estimate, we obtain

$$\begin{aligned}|\widehat{g}_3(\varepsilon,\kappa)| &< M_2\varepsilon^{\theta_1}\cdot O(\kappa^{-2}\exp(-\kappa^{1-\theta_2})),\\ |\widehat{g}_4(\varepsilon,\kappa)| &< (M_2)^2\varepsilon^{2\theta_1}.\end{aligned}$$

Putting together the above estimates, we conclude (4.73) for some $M_9 = M_9(f,g,D) > 0$. □

Proof of Theorem 4.1. Since we already showed (1) in Proposition 4.2, it is enough to prove (2) and (3) with the aid of Propositions 4.1 and 4.3. Combining (4.64) and (4.73), we have

$$\left|\varepsilon - \frac{c_1(\varepsilon)c_2(\varepsilon)}{4D\kappa(\varepsilon)^3}\right| < \frac{M_8}{\kappa(\varepsilon)^4} + M_9D\left\{\frac{\varepsilon^{\theta_1}}{\kappa(\varepsilon)^{1+\theta_2}} + \varepsilon^{2\theta_1} + \exp(-d\kappa(\varepsilon))\right\}.$$

Multiplying both sides of the above inequality by $\kappa(\varepsilon)^3$, we obtain

(4.75)
$$\begin{aligned}&\left|\varepsilon\kappa(\varepsilon)^3 - \frac{c_1(\varepsilon)c_2(\varepsilon)}{4D}\right|\\ &\quad< M_8\kappa(\varepsilon)^{-1} + M_9D\left\{\varepsilon^{\theta_1}\kappa(\varepsilon)^{2-\theta_2} + \varepsilon^{2\theta_1}\kappa(\varepsilon)^3 + \kappa(\varepsilon)^3\exp(-d\kappa(\varepsilon))\right\}.\end{aligned}$$

We choose θ_1 and θ_2 appropriately in order to make the right-hand side of (4.75) tends to zero as $\varepsilon \downarrow 0$. Let us set θ_1, θ_2 such that

$$2\theta_1 + \theta_2 - 2 > 0, \qquad 4\theta_1 - 3 > 0. \tag{4.76}$$

For instance, we can take $\theta_1 = 7/8, \theta_2 = 7/8$. Since we have $\kappa(\varepsilon) < \overline{\kappa}_+(\varepsilon)$, we have

$$\varepsilon^{\theta_1}\kappa(\varepsilon)^{2-\theta_2} < \varepsilon^{\theta_1}(\sigma\varepsilon^{-1/2})^{2-\theta_2} = \sigma^{2-\theta_2}\varepsilon^{(2\theta_1+\theta_2-2)/2}, \tag{4.77a}$$

$$\varepsilon^{2\theta_1}\kappa(\varepsilon)^3 < \varepsilon^{2\theta_1}(\sigma\varepsilon^{-1/2})^3 \quad = \sigma^3\varepsilon^{(4\theta_1-3)/2}, \tag{4.77b}$$

which yield (4.2). We note that the convergence of (4.2) is uniform regardless of our way of selecting $\kappa(\varepsilon)$. Now we show the unique existence of a zero solution of (3.25) which goes to ∞ as $\varepsilon \downarrow 0$. We see that $\lambda(\varepsilon) \equiv \widetilde{\lambda}(\varepsilon,\kappa(\varepsilon))$ is positive for small ε. From Proposition 4.3, we also see that $\widetilde{\lambda}(\varepsilon,\overline{\kappa}_+(\varepsilon))$ is negative. Since $\widetilde{\lambda}(\varepsilon,\cdot)$ is continuous, we obtain at least one zero point in $(\kappa(\varepsilon),\overline{\kappa}_+(\varepsilon))$. Let $\overline{\kappa}(\varepsilon)$ be an arbitrary one of them, then by the definition of $\overline{\kappa}(\varepsilon)$, it holds that

$$\begin{aligned}&\overline{\kappa}(\varepsilon) \to \infty \qquad (\text{as} \quad \varepsilon \downarrow 0),\\ &F(\varepsilon,\overline{\kappa}(\varepsilon),0) = \widehat{\zeta}_0(\varepsilon) - \varepsilon\overline{\kappa}(\varepsilon)^2 - H(\varepsilon,\overline{\kappa}(\varepsilon)^2,\lambda(\varepsilon)) = 0.\end{aligned}$$

Using (4.33) and the above equality, we obtain $\varepsilon\overline{\kappa}(\varepsilon)^2 \to \widehat{\zeta}_0(0)$ as $\varepsilon \downarrow 0$, which implies, for any given $\delta' > 0$, $\overline{\kappa}(\varepsilon)$ must lie in the interval

$$\left(\sqrt{\widehat{\zeta}_0(0) - \delta'} \cdot \varepsilon^{-1/2}, \sqrt{\widehat{\zeta}_0(0) + \delta'} \cdot \varepsilon^{-1/2}\right)$$

for small $\varepsilon > 0$. Since the asymptotic characterization (4.2) holds for any stationary point of $\widetilde{\lambda}(\varepsilon, \cdot)$, we see that $\widetilde{\lambda}(\varepsilon, \cdot)$ is monotone in the above interval. Accordingly $\overline{\kappa}(\varepsilon)$ is uniquely determined and satisfies (4.1). There are no other zero points of $\widetilde{\lambda}(\varepsilon, \cdot)$ in $[\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$ besides $\underline{\kappa}(\varepsilon)$ and $\overline{\kappa}(\varepsilon)$. We see that $\underline{\kappa}(\varepsilon)$ remains bounded, while $\overline{\kappa}(\varepsilon)$ goes to infinity as $\varepsilon \to 0$. Hence it holds that $0 < 2\underline{\kappa}(\varepsilon) < \overline{\kappa}(\varepsilon) < \infty$ for small ε. Let us take any $\kappa_1(\varepsilon)$ from $(\kappa(\varepsilon) - M, \kappa(\varepsilon) + M)$. Then it follows that $\kappa_1(\varepsilon) \in [\underline{\kappa}(\varepsilon), \overline{\kappa}_+(\varepsilon))$ for small $\varepsilon > 0$, and $\kappa_1(\varepsilon) = O(\varepsilon^{-1/3})$ as $\varepsilon \downarrow 0$. We have

$$\tau\widetilde{\lambda}(\varepsilon, \kappa_1(\varepsilon)) = \widehat{\zeta}_0(\varepsilon) - \varepsilon\kappa_1(\varepsilon)^2 - H(\varepsilon, \kappa_1(\varepsilon), \widetilde{\lambda}(\varepsilon, \kappa_1(\varepsilon))). \tag{4.78}$$

Making use of (4.33), we see that the right-hand side of (4.78) converges to $\widehat{\zeta}_0(0) > 0$ as $\varepsilon \downarrow 0$ uniformly for $\kappa_1(\varepsilon)$. We obtain (4.3), and this completes the proof of Theorem 4.1. □

Appendix 1 (Proof of Lemma 4.3). Let $z \in H^1(I)$ and $h \in H^{-1}(I)$ satisfy the equation: $\widehat{T}(\kappa)z = h$, which is equivalent to

$$D(z_x, \varphi_x)_{L^2(I)} + D\kappa^2(z, \varphi)_{L^2(I)} = {}_{H^{-1}(I)}\langle h, \varphi\rangle_{H^1(I)} \quad \forall\, \varphi \in H^1(I). \tag{1}$$

Substituting z for φ, we obtain

$$\begin{aligned} &D\|z_x\|^2_{L^2(I)} + D\kappa^2\|z\|^2_{L^2(I)} \\ &\quad = {}_{H^{-1}(I)}\langle h, z\rangle_{H^1(I)} \le |{}_{H^{-1}(I)}\langle h, z\rangle_{H^1(I)}|. \end{aligned} \tag{2}$$

We have

$$|{}_{H^{-1}(I)}\langle h, z\rangle_{H^1(I)}| \le \|h\|_{H^{-1}(I)}\|z\|_{H^1(I)} \le \frac{1}{2D}\|h\|^2_{H^{-1}(I)} + \frac{D}{2}\|z\|^2_{H^1(I)}. \tag{3}$$

Putting (2) and (3) together, we get

$$\frac{D}{2}\|z_x\|^2_{L^2(I)} + D\left(\kappa^2 - \frac{1}{2}\right)\|z\|^2_{L^2(I)} \le \frac{1}{2D}\|h\|^2_{H^{-1}(I)}.$$

In particular, we obtain

$$D\left(\kappa^2 - \frac{1}{2}\right)\|z\|^2_{L^2(I)} \le \frac{1}{2D}\|h\|^2_{H^{-1}(I)},$$

which leads to (4.28a). In order to prove (4.28b), we use the expression by Green's function:

$$\begin{aligned} z(x) &= \int_0^1 G(x, \xi, \kappa)h(\xi)d\xi \\ &= {}_{H^1(I)}\langle G(x, \cdot, \kappa), h\rangle_{H^{-1}(I)}, \end{aligned}$$

for $x \in I$. Making use of the Schwarz inequality, we have

$$
\begin{aligned}
|z(x)| &= |{}_{H^1(I)}\langle G(x,\cdot,\kappa), h\rangle_{H^{-1}(I)}| \\
&\leq \|G(x,\cdot,\kappa)\|_{H^1(I)} \|h\|_{H^{-1}(I)},
\end{aligned} \tag{4}
$$

for $x \in I$. We obtain (4.28b) from (4.26) and (4). Next we prove (4.29). Assume that $z \in H^1(I)$ satisfies $\widehat{T}(\kappa) z = \Theta(\kappa)$. Multiplying both side by $\bar{z}$ and integrating over I, we have

$$
\begin{aligned}
D\|z_x\|^2_{L^2(I)} + D\kappa^2 \|z\|^2_{L^2(I)} &= (\Theta(\kappa), z)_{L^2(I)} \\
&\leq \|\Theta(\kappa)\|_{L^2(I)} \|z\|_{L^2(I)}.
\end{aligned}
$$

In particular, we have

$$
D\kappa^2 \|z\|^2_{L^2(I)} \leq \|\Theta(\kappa)\|_{L^2(I)} \|z\|_{L^2(I)},
$$

which, combined with the assumption for $\Theta(\kappa)$, yields (4.29a). Finally we prove (4.29b). Expressing z by Green's function

$$
z(x) = \int_0^1 G(x,\xi,\kappa)\Theta(\xi,\kappa)d\xi \qquad (x \in I),
$$

and using the Schwarz inequality, we obtain

$$
|z(x)| \leq \|G(x,\cdot,\kappa)\|_{L^2(I)} \cdot \|\Theta(\kappa)\|_{L^2(I)}. \tag{5}
$$

Combining (4.25) with (5), we obtain (4.29b). Thus we complete the proof. □

Appendix 2 (Proof of Lemma 4.4). Since the latter half of this lemma, namely (4.32), can be proved by using the former half and Lemma 4.3, it suffices to show (4.30) and (4.31). By the definition of $T(\varepsilon,\kappa,\lambda)$ (see (3.12)), we have

$$
\begin{aligned}
T(\varepsilon,\kappa,\lambda) &= \widehat{T}(\kappa) + R(\varepsilon,\kappa,\lambda) + \lambda \\
&= \widehat{T}(\kappa)\{I + Z(\varepsilon,\kappa,\lambda)\},
\end{aligned} \tag{6}
$$

where I is the identity operator in $L^2(I)$, and

$$
Z(\varepsilon,\kappa,\lambda) \equiv \widehat{K}(\kappa)\{R(\varepsilon,\kappa,\lambda) + \lambda\}. \tag{7}
$$

In view of Lemma 3.1 and Proposition 4.1, we see that $\{R(\varepsilon,\kappa,\lambda) + \lambda\}$ is uniformly bounded in $\mathcal{B}(L^2(I), L^2(I))$. Since the norm of $\widehat{K}(\kappa)$ in $\mathcal{B}(L^2(I), L^2(I))$ is $O(\kappa^{-1})$ from (4.28a), there exists $\kappa_* = \kappa_*(f,g,D)$ such that

$$
\|Z(\varepsilon,\kappa,\lambda)\|_{\mathcal{B}(L^2(I),L^2(I))} < \tfrac{1}{2} \tag{8}
$$

holds for any $\kappa > \kappa_*$. Since $\mathcal{B}(L^2(I), L^2(I))$ is a Banach space, we see that

$$
W(\varepsilon,\kappa,\lambda) \equiv \sum_{j=0}^{\infty} (-1)^j Z(\varepsilon,\kappa,\lambda)^j
$$

exists in $\mathcal{B}(L^2(I), L^2(I))$ and its norm is less than 2. We have

$$\{I + Z(\varepsilon, \kappa, \lambda)\}^{-1} = W(\varepsilon, \kappa, \lambda) = I - Z(\varepsilon, \kappa, \lambda)W(\varepsilon, \kappa, \lambda). \tag{9}$$

Using (9), we rewrite (6) as

$$K(\varepsilon, \kappa, \lambda) = \{I - Z(\varepsilon, \kappa, \lambda)W(\varepsilon, \kappa, \lambda)\}\widehat{K}(\kappa).$$

In view of (4.17) and the above equality, we obtain

$$\overline{K}(\varepsilon, \kappa, \lambda) = -Z(\varepsilon, \kappa, \lambda)W(\varepsilon, \kappa, \lambda)\widehat{K}(\kappa),$$

which, combined with (7), implies that (4.30) holds true, if we set

$$\overline{R}(\varepsilon, \kappa, \lambda) = -\{R(\varepsilon, \kappa, \lambda) + \lambda\}W(\varepsilon, \kappa, \lambda). \tag{10}$$

We see that both $\{R(\varepsilon, \kappa, \lambda) + \lambda\}$ and $W(\varepsilon, \kappa, \lambda)$ are uniformly bounded operators in $L^2(I)$ for $(\varepsilon, \kappa, \lambda)$, and so is $\overline{R}(\varepsilon, \kappa, \lambda)$. We complete the proof of Lemma 4.4. □

Appendix 3 (Proof of Proposition 4.2). We prove Proposition 4.2 by contradiction. Let us assume that there exists $\{(\varepsilon_n, \kappa_n, \lambda_n)\}_{n=1}^{\infty}$ such that each $(\varepsilon_n, \kappa_n, \lambda_n)$ satisfies (3.25) and

$$\varepsilon_n \to 0 \quad \text{as} \quad n \uparrow \infty, \qquad \Im\lambda_n \neq 0 \quad \text{for any} \quad n. \tag{11}$$

By virtue of Proposition 4.1, we may suppose, without loss of generality, that $\lambda_n \to \lambda_0 \in \mathbb{C}_{\lambda_*}$ as $n \uparrow \infty$. We rewrite the right-hand side of (3.26) by

$$F(\varepsilon, \kappa, \lambda) = \widehat{\zeta_0}(\varepsilon) - \varepsilon\kappa^2 - \tau\Re\lambda - X(\varepsilon, \kappa, \lambda) - i\Im\lambda\{\tau - Y(\varepsilon, \kappa, \lambda)\}, \tag{12}$$

where X, Y are real-valued functions defined by

$$H(\varepsilon, \kappa, \lambda) = X(\varepsilon, \kappa, \lambda) - i\Im\lambda Y(\varepsilon, \kappa, \lambda). \tag{13}$$

Since $(\varepsilon_n, \kappa_n, \lambda_n)$ satisfies (3.25) and (11), we obtain

$$\widehat{\zeta_0}(\varepsilon_n) - \varepsilon_n\kappa_n^2 - \tau\Re\lambda_n - X(\varepsilon_n, \kappa_n, \lambda_n) = 0, \tag{14a}$$

$$\tau - Y(\varepsilon_n, \kappa_n, \lambda_n) = 0. \tag{14b}$$

It is apparent from (11) that (14b) is equivalent to

$$\Im H(\varepsilon_n, \kappa_n, \lambda_n) = -\tau\Im\lambda_n. \tag{15}$$

We have two cases $\Im\lambda_0 \neq 0$ and $\Im\lambda_0 = 0$. First we consider the case $\Im\lambda_0 \neq 0$. We can assume that $\{\kappa_n\}$ remains bounded, because if not, we see that the left-hand side of (15) converges to zero by virtue of (4.33), while the right-hand side of (15) converges to $\tau\Im\lambda_0 \neq 0$, which is a contradiction. Hence we can assume $\kappa_n \to \kappa_0$ as $n \uparrow \infty$ without loss of generality. It follows from Lemma 4.1 that

$$\lim_{n\uparrow\infty} Y(\varepsilon_n, \kappa_n, \lambda_n)$$

$$= -\frac{1}{\Im\lambda_0}\Im H(0, \kappa_0, \lambda_0) = \sum_{n=0}^{\infty} \frac{c_1^* c_2^* |\psi_n(x^*)|^2}{(\gamma_n + D\kappa_0^2 + \Re\lambda_0)^2 + (\Im\lambda_0)^2}.$$

By the definition of τ_* and (4.15), we see that the right-hand side of the above equality is less than τ, which contradicts (14b) as $n \uparrow \infty$.

Next we consider the case $\Im\lambda_0 = 0$. We make some preparations under the hypotheses of Lemma 3.1. From now on, all function spaces are considered as those of real functions. Let $P(\varepsilon, \kappa, \lambda)$ and $Q(\varepsilon, \kappa, \lambda)$ be the operators defined by

$$R(\varepsilon, \kappa, \lambda) = P(\varepsilon, \kappa, \lambda) + i\varepsilon\tau\Im\lambda Q(\varepsilon, \kappa, \lambda). \tag{16}$$

We introduce a bilinear form from $H^1(I) \times H^1(I)$ to $\mathbb{R}$ defined by

$$\widetilde{B}(\varepsilon, \kappa, \lambda)(z^1, z^2) \equiv D(z^1_x, z^2_x)_{L^2(I)} + ((P(\varepsilon, \kappa, \lambda) + D\kappa^2 + \Re\lambda)z^1, z^2)_{L^2(I)},$$

for $z^1, z^2 \in H^1(I)$. We also define an operator from $H^1(I)$ to $H^{-1}(I)$ by

$$\widetilde{T}(\varepsilon, \kappa, \lambda)z \equiv -z_{xx} + (P(\varepsilon, \kappa, \lambda) + D\kappa^2 + \Re\lambda)z, \tag{17}$$

for $z \in H^1(I)$. Then we have

LEMMA A.1. *Under the same hypotheses of Lemma* 3.1, *we have* (i)–(v).

(i) *Both* $P(\varepsilon, \kappa, \lambda)$ *and* $Q(\varepsilon, \kappa, \lambda)$ *are uniformly bounded linear operators from* $L^2(I)$ *to* $L^2(I)$ *for* $(\varepsilon, \kappa, \lambda)$.

(ii) $\widetilde{B}(\varepsilon, \kappa, \lambda)$ *is a bounded and coercive sesquilinear form on* $H^1(I)$.

(iii) $\widetilde{T}(\varepsilon, \kappa, \lambda)$ *belongs to* $\mathcal{B}(H^1(I), H^{-1}(I))$, *and has an inverse operator denoted by* $\widetilde{K}(\varepsilon, \kappa, \lambda)$. $\widetilde{K}(\varepsilon, \kappa, \lambda)$ *is a uniformly bounded linear operator from* $H^{-1}(I)$ *to* $H^1(I)$ *for* $(\varepsilon, \kappa, \lambda)$.

(iv) *If we assume* (3.6) *in addition, we have*

$$\widetilde{K}(\varepsilon, \kappa, \lambda) \to K(0, \kappa, \Re\lambda) \quad \textit{in} \quad \mathcal{B}(H^{-1}(I), H^1(I)),$$

as $\varepsilon \downarrow 0$. *The convergence is uniform for* (κ, λ) *in any compact subset of* $[0, \infty) \times \mathbb{C}_{\lambda_*}$.

(v) *We have, for* $\kappa > \kappa_*$,

$$|(\widetilde{K}(\varepsilon, \kappa, \lambda)\widetilde{K}(\varepsilon, \kappa, \lambda)h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}| < M_6 \cdot \kappa^{-5/2}, \tag{18}$$

where M_6, κ_* *are the same constants in Lemma* 4.5.

Proof. By virtue of (3.7), we can obtain (i) by a simple calculation from (3.8a) and (16). The proofs of (ii), (iii), and (iv) can be carried out by the same argument as in [NF, Lemma 3.1]. A similar argument used to obtain (4.34a) is also valid to get (v). □

From the definitions of $T(\varepsilon, \kappa, \lambda)$ and $\widetilde{T}(\varepsilon, \kappa, \lambda)$, we have

$$\begin{aligned} T(\varepsilon, \kappa, \lambda) &= \widetilde{T}(\varepsilon, \kappa, \lambda) + i\Im\lambda\{I + \varepsilon\tau Q(\varepsilon, \kappa, \lambda)\} \\ &= \widetilde{T}(\varepsilon, \kappa, \lambda)\{I + i\Im\lambda\widetilde{Z}(\varepsilon, \kappa, \lambda)\}, \end{aligned} \tag{19}$$

where

$$\widetilde{Z}(\varepsilon, \kappa, \lambda) \equiv \widetilde{K}(\varepsilon, \kappa, \lambda)\{I + \varepsilon\tau Q(\varepsilon, \kappa, \lambda)\}. \tag{20}$$

LEMMA A.2. *Under the same hypotheses of Lemma* 3.1, *there exists* $\nu_* = \nu_*(f, g, D)$ *such that it holds for* $|\Im\lambda| < \nu_*$ *that*

$$K(\varepsilon,\kappa,\lambda) = \{I - i\Im\lambda\widetilde{Z}(\varepsilon,\kappa,\lambda)\}\{I - (\Im\lambda)^2\widetilde{W}(\varepsilon,\kappa,\lambda)\}\widetilde{K}(\varepsilon,\kappa,\lambda), \tag{21}$$

where $\widetilde{W}(\varepsilon,\kappa,\lambda)$ *is a uniformly bounded linear operator in* $L^2(I)$ *for* $(\varepsilon,\kappa,\lambda)$.

Proof. In view of (20) and Lemma A.1, we see that $\widetilde{Z}(\varepsilon,\kappa,\lambda)$ is uniformly bounded in $\mathcal{B}(L^2(I), L^2(I))$ for $(\varepsilon,\kappa,\lambda)$. Since $\mathcal{B}(L^2(I), L^2(I))$ is a Banach space, we have

$$\{I + i\Im\lambda\widetilde{Z}(\varepsilon,\kappa,\lambda)\}^{-1} = \sum_{n=0}^{\infty}\{-i\Im\lambda\widetilde{Z}(\varepsilon,\kappa,\lambda)\}^n, \tag{22}$$

for small $\Im\lambda$. We set

$$\widetilde{W}(\varepsilon,\kappa,\lambda) = \sum_{n=0}^{\infty}(-1)^n(\Im\lambda)^{2n}\{\widetilde{Z}(\varepsilon,\kappa,\lambda)\}^{2n+2}.$$

Then we see that $\widetilde{W}(\varepsilon,\kappa,\lambda)$ is uniformly bounded for $(\varepsilon,\kappa,\lambda)$ in $\mathcal{B}(L^2(I), L^2(I))$ and that from (22),

$$\{I + i\Im\lambda\widetilde{Z}(\varepsilon,\kappa,\lambda)\}^{-1} = \{I - i\Im\lambda\widetilde{Z}(\varepsilon,\kappa,\lambda)\}\{I - (\Im\lambda)^2\widetilde{W}(\varepsilon,\kappa,\lambda)\} \tag{23}$$

in $\mathcal{B}(L^2(I), L^2(I))$. It follows from (19) that

$$K(\varepsilon,\kappa,\lambda) = \{I + i\Im\lambda\widetilde{Z}(\varepsilon,\kappa,\lambda)\}^{-1}\widetilde{K}(\varepsilon,\kappa,\lambda). \tag{24}$$

Combining (23) and (24), we obtain the desired result. □

From the definition of $Y(\varepsilon,\kappa,\lambda)$,

$$\begin{aligned} &Y(\varepsilon_n,\kappa_n,\lambda_n) \\ &\quad = -\Im H(\varepsilon_n,\kappa_n,\lambda_n)/\Im\lambda_n \\ &\quad = -(\Im K(\varepsilon_n,\kappa_n,\lambda_n)h_2(\varepsilon_n), h_1(\varepsilon_n))_{L^2(I)}/\Im\lambda_n, \end{aligned} \tag{25}$$

which, combined with (21) and (20), leads to

$$\begin{aligned} &Y(\varepsilon_n,\kappa_n,\lambda_n) \\ &\quad = (\widetilde{Z}(\varepsilon,\kappa,\lambda)\{I - (\Im\lambda)^2\widetilde{W}(\varepsilon,\kappa,\lambda\}\widetilde{K}(\varepsilon,\kappa,\lambda)h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}\Big|_{\substack{\varepsilon=\varepsilon_n,\kappa=\kappa_n\\ \lambda=\lambda_n}} \\ &\quad = (\widetilde{K}(\varepsilon,\kappa,\lambda)\widetilde{K}(\varepsilon,\kappa,\lambda)h_2(\varepsilon), h_1(\varepsilon))_{L^2(I)}\Big|_{\substack{\varepsilon=\varepsilon_n,\kappa=\kappa_n\\ \lambda=\lambda_n}} + O(\varepsilon_n) + O(|\Im\lambda_n|^2). \end{aligned} \tag{26}$$

If $\{\kappa_n\}$ is unbounded, we may assume that $\kappa_n \to \infty$ as $n \uparrow \infty$ by taking a subsequence. Using Lemma A.1.(v) and the assumption $\Im\lambda_n \to 0$, we see that the right-hand side of (26) converges to zero, which contradicts (14b). Hence $\{\kappa_n\}$ must remain bounded. So we may assume that we have $\kappa_n \to \kappa_0 \in [0,\infty)$ and $\lambda_n \to \lambda_0 \in \mathbb{C}_{\lambda_*}$ as $n \uparrow \infty$. Letting $n \uparrow \infty$ in (26), it follows from Lemma A.1 (iv) and (4.8) that

$$\begin{aligned} &\lim_{n\uparrow\infty} Y(\varepsilon_n,\kappa_n,\lambda_n) \\ &\quad = c_1^*c_2^* \cdot {}_{H^1(I)}\langle K(0,\kappa_0,\Re\lambda_0)K(0,\kappa_0,\Re\lambda_0)\delta(x-x^*), \delta(x-x^*)\rangle_{H^{-1}(I)} \\ &\quad = \sum_{n=0}^{\infty}\frac{c_1^*c_2^*|\psi_n(x^*)|^2}{(\gamma_n + D\kappa_0^2 + \Re\lambda_0)^2} \quad < \tau, \end{aligned}$$

which contradicts (14b) when $n \uparrow \infty$. We complete the proof of Proposition 4.2. □

Acknowledgment. The first author expresses his sincere gratitude to Professor Hiroshi Matano for many stimulating discussions and continuous encouragement. Special thanks go to Professor Masashi Katsurada for useful suggestions on numerical computations.

REFERENCES

[Ca] G. CAGINALP, *An analysis of a phase field model of a free boundary*, Arch. Rational Mech. Anal., 92 (1986), pp. 205–245.

[Ch] X.-Y. CHEN, *Dynamics of interfaces in reaction diffusion systems*, Hiroshima Math. J., 21 (1991), pp. 47–83.

[CH] R.COURANT AND D.HILBERT, *Methods of Mathematical Physics Vol.* I, Wiley-Interscience, New York, 1953.

[F] P. C. FIFE, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497–521.

[GGI] Y. GIGA, S. GOTO, AND H. ISHII, *Global existence of weak solutions for interface equations coupled with diffusion equations*, preprint.

[GJ] R. GARDNER AND C. K. R. T. JONES, *Stability of travelling wave solutions of diffusive predator-prey systems*, Trans. Amer. Math. Soc., 327 (1991), pp. 465–524.

[HNM] D. HILHORST, Y. NISHIURA, AND M. MIMURA, *A free boundary problem arising in some reacting-diffusing system*, Proc. Roy. Soc. Edinburgh, 118A (1991), pp. 355–378.

[Ito] M. ITO, *A remark on singular perturbation methods*, Hiroshima Math. J., 14 (1985), pp. 619–629.

[MS] W. W. MULLINS AND R. F. SEKERKA, *Stability of a planar interface during solidification of a dilute binary alloy*, J. Appl. Phys., 3 (1964), pp. 444–451.

[MTH] M. MIMURA, M. TABATA, AND Y. HOSONO, *Multiple solutions of two-point boundary value problems of Neumann type with a small parameter*, SIAM J. Math. Anal., 11 (1980), pp. 613–631.

[N1] Y. NISHIURA, *Singular limit approach to stability and bifurcation for bistable reaction diffusion systems*, Rocky Mountain J. Math., 21 (1991), pp. 727–767.

[N2] ———, *Coexistence of Infinitely Many Stable Solutions to Reaction Diffusion Systems in the Singular Limit*, Dynamics Reported, C. K. R. T. Jones, L. Kirchgraber, and H. O. Walther, eds., Springer-Verlag, in press.

[NF] Y. NISHIURA AND H. FUJII, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 1726–1770.

[NM] Y. NISHIURA AND M. MIMURA, *Layer oscillations in reaction-diffusion systems*, SIAM J. Appl. Math., 49 (1989), pp. 481–514.

[OMK] T. OHTA, M. MIMURA, AND R. KOBAYASHI, *Higher-dimensional localized patterns in excitable media*, Physica D, 34 (1989), pp. 115–144.

[Sa] K. SAKAMOTO, *Construction and stability analysis of transition layer solutions in reaction-diffusion systems*, Tôhoku Math. J., 42 (1990), pp. 17–44.

[St] J. STRAIN, *Linear stability of planar solidification fronts*, Physica D, 30 (1988), pp. 297–320.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 135–147, January 1994

004

TIKHONOV REGULARIZATION FOR FINITELY AND INFINITELY SMOOTHING OPERATORS*

B. A. MAIR†

Abstract. The main goal of this paper is to obtain a unified theory of Tikhonov regularization, incorporating explicit asymptotic rates of convergence based on a priori assumptions, which cover both the finitely and infinitely smoothing forward operators, and to extend a classic result of Natterer to this more general framework. More specifically, it is shown that, for a large class of operators, as in the finitely smoothing case obtained by Natterer, the stabilizing functional involved in the minimization process can be determined by larger norms over much smaller classes than those determined by the a priori assumption for the true solution.

Key words. Tikhonov regularization, Hilbert scales, spectral measure

AMS subject classifications. 45L05, 35R25

1. Introduction. The method of Tikhonov regularization is one of the most widely applied methods for solving ill-posed inverse problems that arise in a wide variety of problems in science and engineering. As is well known, the main difficulty in applying this method occurs in the choice of a regularizing parameter, usually denoted by α, depending on the error in the data. Both a priori and a posteriori methods of choosing α have been developed, mainly for the case of finitely smoothing forward operators (cf. [12] and [23]). However, many problems in partial differential equations (cf. [3], [6], [7], [10], [11], [13]–[15], [25]–[27]), linear systems theory (cf. [10] and [18]), statistics (cf. [8]), optics (cf. [4]), and astronomy (cf. [9]), just to mention a few areas, give rise to linear ill-posed problems determined by infinitely smoothing forward operators. The main goal of this paper is to obtain a unified theory, incorporating explicit asymptotic rates of convergence based on a priori assumptions, which covers both the finitely and infinitely smoothing forward operators. We also show that, as in the finitely smoothing case (cf. [22]), the stabilizing functional involved in the minimization process can be determined by larger norms over much smaller classes than those determined by the a priori assumption for the true solution. In other words, it certainly does no harm to "over-regularize." In fact, the application of over-regularization to severely ill-posed inverse heat conduction problems (cf. [3], [15], [26], and [27]) have greatly improved the accuracy of the resulting numerical algorithms.

More specifically, we consider the problem of finding an approximate solution to an operator equation of the form

$$T(x) = y,$$

based on inaccurate data y_δ. Although the results in §2 cover a more general case, for this discussion assume that T is injective with dense range, and that the data satisfies $\|y_\delta - y\| \leq \delta$. As in [19]–[23] and [27], consider approximations obtained by minimizing

$$\|Tx - y_\delta\|^2 + \alpha^2 \|Bx\|^2$$

*Received by the editors October 8, 1992; accepted for publication March 15, 1993. This work was partially supported by National Science Foundation grant DMS 9006308.

†Department of Mathematics, University of Florida, Gainesville, Florida 32611.

over the domain of a suitable constraint operator B, where it is assumed that the true solution x_0 satisfies, $\|Bx_0\| \leq E$.

By using techniques in [20], [21], and [26] and a condition describing the degree of ill-posedness of quite general T in terms of the spectral measure of B^*B, it is shown in §2 that the Tikhonov regularized solution $x_{\alpha,\delta}$ satisfies

$$\|x_{\alpha,\delta} - x_0\| = O\left(\sqrt{\varphi^{-1}(\delta^2)}\right) \quad \text{if } \alpha = \frac{\delta}{E},$$

where φ is a convex function which quantifies the ill-posedness of the operator T.

This generalizes the classical result (cf. [2], [22], and [23]) for finitely smoothing operators, in which $\varphi^{-1}(t)$ is simply a power of t. For the general case considered here, there is no closed form for $\varphi^{-1}(t)$. Explicit convergence rates can only be obtained from estimates of the behavior of $\varphi^{-1}(t)$ as $t \to 0$. The surprising fact is that this basic result enables us to generalize a well-known result of Natterer [22], thus providing a new result for an important class of inverse problems (see §5).

In §3, we specialize this result to the case when the approximation is taking place in a Hilbert scale, $\{X_s : s \in \mathbb{R}\}$, to obtain the usual result (cf. [2], [22], and [23]) for finitely smoothing T and a logarithmic rate of convergence for a general class of infinitely smoothing operators (cf. [2] and [26]).

To demonstrate the significance of the general result, §4 applies it to problems of deconvolution and inversion of dilationally invariant transforms (cf. [1], [2], [4], [5], [7]–[9], [13]–[15], and [24]). There we obtain a convergence rate for a very general class of deconvolution problems that is faster than the classical one obtained in [2] for the special case considered there.

In the final section we show that for a very general class of operators, if approximation is taking place in a Hilbert scale and it is assumed that the true solution $x_0 \in X_q$, then the unique minimizer of $\|Tx - y_\delta\|^2 + \alpha^2\|x\|_p^2$ over $x \in X_p$, for any $p \geq q$, and an appropriate choice of α, produces an approximation of x_0 with the same rate of convergence as when $p = q$. This generalizes the result of Natterer for finitely smoothing operators obtained in [22]. It is interesting to note that the proof here is very different from that in [22]. It simply uses the basic result in §2 and denseness properties of the spaces in a Hilbert scale, avoiding the very involved use of interpolation theorems present in the classical proof.

2. A general error estimate for Tikhonov regularization. Let X, Y be Hilbert spaces and $T : X \longrightarrow Y$ be a bounded, linear operator with nonclosed range $\mathcal{R}(T)$. Then the Moore–Penrose inverse $T^\dagger$ is not continuous and the problem of solving the equation

$$Tx = y$$

is ill-posed, even if $y \in \mathcal{R}(T)$.

In this section we use basic properties of Moore-Penrose inverses contained in [2], [17], [20]. Here, we consider the problem of obtaining an approximate "solution" to the "equation"

$$Tx = y_0, \tag{2.1}$$

based on inaccurate data y_δ, where y_0 is not even assumed to be in $\mathcal{R}(T)$.

We assume y_0 is in the domain, $\mathcal{D}(T^\dagger)$, of the Moore–Penrose inverse $T^\dagger$, which is $\mathcal{R}(T) + \mathcal{R}(T)^\perp$.

For any such y_0, let

$$x_0 = T^\dagger(y_0). \tag{2.2}$$

This is the so-called best approximate solution to (2.1) and is the unique classical solution if T is injective and $y_0 \in \mathcal{R}(T)$.

Since y_0 is usually not known, we need to obtain an approximation to x_0 based on the approximate data y_δ.

Let Q denote the orthogonal projection of Y onto $\overline{\mathcal{R}(T)}$ and assume that the appproximate data y_δ satisfies

$$\|Q(y_\delta - y_0)\| \leq \delta. \tag{2.3}$$

To construct a regularized solution of (2.1), we need a "constraint operator" B that quantifies the smoothness constraints imposed on our solution x_0 (cf. [2], [19]–[21], and [27]).

DEFINITION 2.1. Let B be a closed, densely defined operator on X, mapping its domain $V \subset X$ onto a Hilbert space Z. For each α and $\delta > 0$, define the quadratic functional $L_{\alpha,\delta}$ on V by

$$L_{\alpha,\delta}(x) = \|Tx - y_\delta\|^2 + \alpha^2\|Bx\|^2.$$

Then we seek an approximation to x_0 obtained by minimizing $L_{\alpha,\delta}$ over V.

To guarantee the existence of a unique minimizer, we assume the following.

Assumption 2.2. There exists $\beta > 0$ such that $\|Tx\| \geq \beta\|x\|$ for all $x \in \mathcal{N}(B)$.

Then, by reformulating the minimization as a least squares problem, the following is shown in [20].

THEOREM 2.3. *$L_{\alpha,\delta}$ has a unique minimizer $x_{\alpha,\delta} \in V$. Furthermore, $x_{\alpha,\delta} \in \mathcal{D}(B^*B)$ and is the unique solution of $(T^*T + \alpha^2 B^*B)x_{\alpha,\delta} = T^*y_\delta$.*

Now, to obtain explicit convergence rates for the error in approximating x_0 by $x_{\alpha,\delta}$, assume the following.

Assumption 2.4. T is injective.

The general case can be reduced to this, by considering the restriction of T to $\mathcal{N}(T)^\perp$.

DEFINITION 2.5. *For each $\varepsilon > 0$, $\nu(\varepsilon) = \sup\{\|x\| : x \in V,\ \|Tx\| \leq \varepsilon,\ \|Bx\| \leq 1\}$.*

THEOREM 2.6. *Under Assumption 2.4, for any $\alpha > 0$, $\|x_{\alpha,\delta} - x_0\| \leq \sqrt{(\delta^2/\alpha^2) + \|Bx_0\|^2}\ \nu(\alpha)$.*

Proof. We need to go back to the characterization of $x_{\alpha,\delta}$ as the result of a least squares method. As in [20], for $\alpha > 0$ define $K_\alpha : V \longrightarrow Y \oplus Z$ by $K_\alpha(x) = [Tx, \alpha Bx]$, where $[y, z]$ represents the generic element in $Y \oplus Z$, and the inner product on this space is: $\langle [y, z], [y', z'] \rangle_{Y\oplus Z} = \langle y, y' \rangle_Y + \langle z, z' \rangle_Z$. Then $x_{\alpha,\delta} = K_\alpha^\dagger([y_\delta, 0])$, where $K_\alpha^\dagger$ is the Moore–Penrose inverse of K_α. Hence $x_{\alpha,\delta}$ is characterized as the unique solution in V of the equation

$$K_\alpha x_{\alpha,\delta} = Q_\alpha([y_\delta, 0]), \tag{2.4}$$

where Q_α is the projection of Y onto $\overline{\mathcal{R}(K_\alpha)} = \mathcal{R}(K_\alpha)$.

Now, for any $x \in V$,

$$\langle [y_\delta - Qy_\delta, 0], K_\alpha x \rangle_{Y\oplus Z} = \langle y_\delta - Qy_\delta, Tx \rangle_Y = 0,$$

since $y_\delta - Qy_\delta \in \mathcal{R}(T)^\perp$. Hence,

$$Q_\alpha[y_\delta, 0] - [Qy_\delta, 0] \in \mathcal{R}(K_\alpha)^\perp. \tag{2.5}$$

Now, by (2.4),

$$\begin{aligned} K_\alpha x_0 - [Qy_\delta, 0] &= K_\alpha(x_0 - x_{\alpha,\delta}) + K_\alpha x_{\alpha,\delta} - [Qy_\delta, 0] \\ &= K_\alpha(x_0 - x_{\alpha,\delta}) + Q_\alpha[y_\delta, 0] - [Qy_\delta, 0]. \end{aligned}$$

Hence by (2.5) and (2.3),

$$\begin{aligned} \|K_\alpha(x_0 - x_{\alpha,\delta})\|^2 &\le \|K_\alpha x_0 - [Qy_\delta, 0]\|^2 \\ &= \|Tx_0 - Qy_\delta\|^2 + \alpha^2\|Bx_0\|^2 \\ &= \|Qy_0 - Qy_\delta\|^2 + \alpha^2\|Bx_0\|^2 \\ &\le \delta^2 + \alpha^2\|Bx_0\|^2. \end{aligned}$$

Thus,

$$\|T(x_0 - x_{\alpha,\delta})\| \le \sqrt{\delta^2 + \alpha^2\|Bx_0\|^2},$$

$$\|B(x_0 - x_{\alpha,\delta})\| \le \sqrt{\frac{\delta^2}{\alpha^2} + \|Bx_0\|^2}.$$

The result follows from Definition 2.5.

Now, to obtain an a priori error estimate, assume that the true solution x_0 satisfies the following.

Assumption 2.7. $\|Bx_0\| \le E$, for some fixed constant E.

Then, from Theorem 2.6 we obtain the following.

THEOREM 2.8. *Let $\alpha(\delta) = \delta/E$, and define $x_\delta = x_{\alpha(\delta),\delta}$. Then*

$$\|x_\delta - x_0\| \le \sqrt{2}\, E\, \nu\left(\frac{\delta}{E}\right).$$

This is a slight improvement and generalization of Theorem 3.4 in [2]. To make this result more applicable we need to obtain an estimate of $\nu(\varepsilon)$. To do this, assume the following (cf. [26]).

Assumption 2.9. There exists a continuous function $\varphi : [0,\infty) \to [0,\infty)$ such that: (i) the map $s \mapsto \varphi(s)/s$ is increasing on $(0,\infty)$;

(ii) $\varphi(s) = 0$ if and only if $s = 0$;

(iii) φ is convex on an interval containing $\{1/\lambda : \lambda$ is in the spectrum of $B^*B\}$;

(iv) there exists a constant $m > 0$ such that

$$m^2 \int \varphi\left(\frac{1}{\lambda}\right) \lambda \, d\mathcal{M}_{x,x}(\lambda) \le \|Tx\|^2 \quad \text{for all } x \in \mathcal{D}(B^*B),$$

where $\mathcal{M}$ is the spectral measure of B^*B.

THEOREM 2.10. *Assuming* 2.9, $\nu(\varepsilon) \le \sqrt{\varphi^{-1}(\varepsilon^2/m^2)}$ *for all* $\varepsilon > 0$.

Proof. Let $x \in \mathcal{D}(B^*B)$. Then, by Jensen's inequality and Assumption 2.9,

$$\begin{aligned} \varphi\left(\frac{\|x\|^2}{\|Bx\|^2}\right) = \varphi\left(\frac{\int (1/\lambda)\lambda \, d\mathcal{M}_{x,x}(\lambda)}{\int \lambda \, d\mathcal{M}_{x,x}(\lambda)}\right) &\le \frac{\int \varphi(1/\lambda)\lambda \, d\mathcal{M}_{x,x}(\lambda)}{\|Bx\|^2} \\ &\le \frac{\|Tx\|^2}{m^2\|Bx\|^2}. \end{aligned}$$

Now, the graph of the restriction of B to $\mathcal{D}(B^*B)$ is dense in the graph of B. Hence, the above inequality is valid for $x \in V$.

Now, if $\|Bx\| \leq 1$, then $\|x\| \leq \|x\|/\|Bx\|$; so

$$\frac{\varphi(\|x\|^2)}{\|x\|^2} \leq \frac{\varphi(\|x\|^2/\|Bx\|^2)}{\|x\|^2/\|Bx\|^2} \leq \frac{\|Tx\|^2}{m^2\|x\|^2}.$$

Hence $\varphi(\|x\|^2) \leq \|Tx\|^2/m^2$. The proof is completed by noting that φ^{-1} exists and is increasing.

THEOREM 2.11. *Under Assumptions* 2.4, 2.7, *and* 2.9,

$$\|x_\delta - x_0\| \leq \sqrt{2}E\sqrt{\varphi^{-1}\left(\frac{\delta^2}{m^2E^2}\right)} = O\left(\sqrt{\varphi^{-1}(\delta^2)}\right).$$

3. Approximation in Hilbert scales. In this section we consider the frequently used constraint of assuming that the best approximate solution is in some fixed ball in a Hilbert scale. The case when T is a finitely smoothing operator has been well studied (cf. [2], [22], and [23], and the references therein). Here we obtain the classical result as a special case of Theorem 2.11 and obtain a new result applicable to a general class of infinitely smoothing operators when the Hilbert scale is viewed as Sobolev spaces.

Assumption 3.1. Assume that $\{X_s : s \in \mathbb{R}\}$ is a Hilbert scale generated by a selfadjoint, densely defined, unbounded operator L, on X, with $\|Lx\| \geq \|x\|$, for all $x \in \mathcal{D}(L)$. As usual, $X_0 = X$, and $X_s = \mathcal{D}(L^s)$.

The usual method of regularization by smoothing in a Sobolev space is obtained from the general case by setting $V = Z = X_p$, for some $p > 0$, and B to be the identity on X_p.

Then, the approximation $x_{\alpha,\delta}$ obtained in Theorem 2.3 is the unique minimizer of the quadratic functional

$$x \mapsto \|Tx - y_\delta\|^2 + \alpha^2\|x\|_p^2,$$

where $\|\cdot\|_p$ denotes the norm in X_p. Now, under these assumptions, it is easy to see that $\mathcal{D}(B^*B) = X_{2p}$ and $B^*B = L^{2p}$. Hence, by Theorem 2.3, we obtain

$$(T^*T + \alpha^2L^{2p})x_{\alpha,\delta} = T^*y_\delta.$$

Now, consider the classical choice of $\alpha(\delta) = \delta/E$ (so $x_{\alpha,\delta} = x_\delta$), where it is assumed that the solution x_0 satisfies

$$\|x_0\|_p \leq E.$$

To obtain an estimate on the rate of convergence of x_δ to x_0, it is often assumed that there exist constants $a, m, M > 0$, such that

$$m\|x\|_{-a} \leq \|Tx\| \leq M\|x\|_{-a} \quad \text{for all } x \in X. \tag{3.1}$$

To fit this into the framework of Assumption 2.9, let $\mathcal{L}$ be the spectral measure of L. Then the spectral measure of $B^*B = L^{2p}$ is given by

$$d\mathcal{M}(\lambda) = d\mathcal{L}(\lambda^{1/2p}). \tag{3.2}$$

We now deduce the classical convergence rate (cf. [2] and [22]) from our results in §2.

THEOREM 3.2. *If* $m\|x\|_{-a} \leq \|Tx\|$ *for all* $x \in X_{2p}$, *then*

$$\|x_\delta - x_0\| \leq \sqrt{2}\left(\frac{E^a}{m^p}\right)^{1/(a+p)} \delta^{p/(a+p)}.$$

Proof.

$$\begin{aligned}\|x\|_{-a}^2 &= \int_1^\infty \lambda^{-2a} d\mathcal{L}_{x,x}(\lambda) \\ &= \int_1^\infty \lambda^{-a/p}\, d\mathcal{M}_{x,x}(\lambda) \\ &= \int_1^\infty \left(\frac{1}{\lambda}\right)^{1+a/p} \lambda\, d\mathcal{M}_{x,x}(\lambda).\end{aligned}$$

Hence, setting $\varphi(s) = s^{1+a/p}$, we see that Assumption 2.9 is satisfied.

Clearly, $\varphi^{-1}(s) = s^{p/(a+p)}$ and the result follows from Theorem 2.11.

To deal with a general class of infinitely smoothing operators, we need to investigate properties of functions of the following form.

DEFINITION 3.3. For each $\beta, \gamma > 0$, define the function

$$\psi_{\beta,\gamma}(s) = s \exp\left(-\frac{\beta}{s^\gamma}\right) \quad \text{if } s > 0 \quad \text{and} \quad \psi_{\beta,\gamma}(0) = 0.$$

Then it is easy to see what follows in Lemma 3.4.

LEMMA 3.4.

(i) *The map* $s \mapsto (1/s)\psi_{\beta,\gamma}(s)$ *is increasing on* $(0, \infty)$.

(ii) $\psi_{\beta,\gamma}(s) = 0 \iff s = 0$.

(iii) *If* $\beta \geq 1$, *then* $\psi_{\beta,\gamma}$ *is convex on* $[0, 1]$.

LEMMA 3.5. $\psi_{\beta,\gamma}^{-1}(s) = (\beta/(\log 1/s))^{1/\gamma}\,(1 + o(1)) \quad$ *as* $s \to 0+$.

Proof. Let $t = \psi_{\beta,\gamma}^{-1}(s)$. Then

$$\log s = \log t - \frac{\beta}{t^\gamma}. \tag{3.3}$$

Hence

$$t = \left(\frac{\beta}{\log t - \log s}\right)^{1/\gamma} = \left(\frac{\beta}{|\log s|}\right)^{1/\gamma} G(s),$$

where

$$\begin{aligned}G(s) &= \left|\frac{\log t}{\log s} - 1\right|^{-1/\gamma} \\ &= \left|\frac{\log t}{\log t - \beta t^{-\gamma}} - 1\right|^{-1/\gamma}, \quad \text{by (3.3)} \\ &= \left|\left(1 - \frac{\beta}{t^\gamma \log t}\right)^{-1} - 1\right|^{-1/\gamma}.\end{aligned}$$

Hence $G(s) \to 1$ as $s \to 0+$.

This result (not the proof) is similar to one in [26].

THEOREM 3.6. *If there exist positive constants a, m, and b such that*

$$m^2 \int \exp(-b\lambda^a)\, d\mathcal{L}_{x,x}(\lambda) \le \|Tx\|^2 \quad \textit{for all } x \in X_{2p},$$

then

$$\|x_\delta - x_0\| \le \sqrt{2}E \left(\frac{\max\{b,1\}}{2\log(mE/\delta)} \right)^{p/a} (1 + o(1)) = O\left(\left[\log \frac{1}{\delta} \right]^{-p/a} \right).$$

Proof. From (3.2) it is clear that

$$m^2 \int \varphi\left(\frac{1}{\lambda}\right) \lambda \, d\mathcal{M}_{x,x}(\lambda) \le \|Tx\|^2 \quad \text{for all } x \in X_{2p},$$

where $\varphi(s) = s \exp\left(-\max\{b,1\} s^{-a/2p}\right)$.

Hence, by using Lemma 3.4, it is clear that Assumption 2.9 is satisfied.

It is interesting to note that this convergence rate is obtained for such a wide class of problems by using the classical choice of $\alpha(\delta) = \delta/E$, contrary to results in [2, §10.2], which indicate a different choice for the particular example considered there.

4. Applications. It is well known that inverse problems for the heat equation have important applications in science and engineering. The basic, so-called Inverse Heat Conduction Problem (cf. [3]) gives rise to the problem of solving convolution equations with kernels whose Fourier transforms decrease exponentially fast (cf. [6], [7], and [15]). Our result applies easily to a general class of such deconvolution problems.

For a given kernel $K \in L^2(\mathbb{R})$, consider the problem of solving the convolution equation

$$K * f = g \tag{4.1}$$

for $f, g \in L^2(\mathbb{R})$, given approximate data $g_\delta \in L^2(\mathbb{R})$, satisfying

$$\|g_\delta - g\|_{L^2} \le \delta. \tag{4.2}$$

Assumption 4.1.

(i) There exist constants $a, b, c > 0$ such that the Fourier transform of K satisfies $|\hat{K}(s)| \ge ce^{-b|s|^a}$ for all $s \in \mathbb{R}$.

(ii) For some $p > 0$, the true solution lies in a fixed ball in the usual Sobolev space H^p, i.e., $\|f\|_{H^p} \le E$.

THEOREM 4.2. *Assuming* (4.2) *and Assumption* 4.1, *let f_δ satisfy*

$$\hat{f}_\delta(s) = \frac{\overline{\hat{K}}(s)}{|\hat{K}(s)|^2 + (\delta/E)^2(1+s^2)^p} \, \hat{g}_\delta(s).$$

Then $\|f_\delta - f\|_{L^2} = O([\log(1/\delta)]^{-p/a})$.

Proof. Consider $X = Y = L^2(\mathbb{R})$, $Tf = K * f$, $Sf = f - f''$, and $L = S^{1/2}$.

Then $H^p = \mathcal{D}(L^p) = X_p$, and we are in the framework of §3.

Now, $\|Tf\|_{L^2}^2 = \int_1^\infty |\hat{K}(\sqrt{\lambda^2-1})|^2 d\mathcal{L}_{f,f}(\lambda)$, where $\mathcal{L}$ is the spectral measure of L.

Hence, by Assumption 4.1,

$$\|Tf\|_{L^2}^2 \geq c^2 \int_1^\infty \exp(-2b\lambda^a)\, d\mathcal{L}_{f,f}(\lambda).$$

The result follows from Theorem 3.6, and the representation of $f_\delta = (T^*T + (\delta^2/E^2)L^{2p})^{-1}T^*g_\delta$.

The classical inverse problem of determining initial temperature from later temperature readings (cf. [10]–[14] and [25]) also gives rise to an ill-posed problem with an exponential rate of decrease of its spectral information. Hence, the analysis in [14], which provided only partially explicit convergence rates, can now be modified using Theorem 3.6 in order to obtain a slightly different numerical scheme with completely explicit error bounds.

Many inverse problems in optics and astronomy (cf. [4] and [9]) can be modeled, at least approximately, by the problem of solving an integral equation of the type

$$\int_0^\infty k(st)f(t)dt = g(s), \tag{4.3}$$

given approximate data g_δ.

As in [1] and [4], (4.3) is equivalent to solving the convolution equation

$$G = K * F, \tag{4.4}$$

where

$$K(t) = e^{-t}k(e^{-t}), \quad F(t) = f(e^t), \quad G(t) = e^{-t}g(e^{-t}). \tag{4.5}$$

Thus an approximation to f can be obtained from the regularized solution of (4.4) if $K \in L^2(\mathbb{R})$. Error estimates can be obtained from the classical Theorem 3.2 or from Theorem 3.6.

To state our results here, we will find it useful to introduce the following (cf. [24]).

DEFINITION 4.3.

(i) For any $f : [0,\infty) \to \mathbb{R}$, define $Jf(t) = tf(t)$.

(ii) Define the measure μ on $[0,\infty)$ by $d\mu(t) = dt/t$.

(iii) $H^1(\mu)$ denotes the set of all absolutely continuous functions f on $[0,\infty)$ such that $f(0) = 0$ and $f' \in L^2(\mu)$, with the usual norm.

As noted in [24], μ is the Haar measure of the group $[0,\infty)$ under multiplication (cf. [8]).

Now, consider the problem of Laplace transform inversion. An application of Theorem 3.6 in the special case of $p = 1$ gives the following (cf. [24]).

THEOREM 4.4. *Suppose the data g_δ satisfies $\|J(g_\delta - g)\|_{L^2(\mu)} \leq \delta$, and the true solution, f, to the equation,*

$$\int_0^\infty e^{-st}f(t)dt = g(s),$$

satisfies $\|Jf\|_{H^1(\mu)} \leq E$. Let $f_\delta(t) = F_\delta(\log t)$, where

$$\hat{F}_\delta(s) = \frac{\Gamma(1-is)}{\left(\left|\frac{\pi s}{\sin h\pi s}\right| + \frac{\delta^2}{E^2}(1+s^2)\right)}\hat{G}_\delta(s)$$

and $G_\delta(t) = e^{-t}g_\delta(e^{-t})$. *Then* $\|f_\delta - f\|_{L^2(\mu)} = O([\log(1/\delta)]^{-1})$.

Proof. As in [1], $|\hat{k}(s)|^2 = \pi s/|\sin h\pi s| \geq ce^{-\pi s}$. The result follows by applying Theorem 3.6 with $p = a = 1$.

This same error estimate was obtained in [1].

Now, consider problem (4.3) with the kernel $k(t) = J_1^2(t)/t^2$, where J_1 is the usual Bessel function (cf. [4]). Using formulas in [15], it is easy to see that the Fourier transform of the corresponding kernel K given by (4.5) satisfies

$$|\hat{K}(s)| = \frac{8}{\pi^{1/4}(1+s^2)(9+s^2)^{1/2}}.$$

Hence this problem is not as ill-posed as the Laplace-inversion problem and can be solved by the classical result for finitely smoothing operators. In fact, under the same conditions as in Theorem 4.4, the convergence rate for this problem is $O\left(\delta^{2/5}\right)$.

5. Higher-order regularization. This section combines the general framework in §2 with the basic assumptions for approximation in Hilbert scales (§3) to extend the result in [22] to a more general framework that includes both the finitely and infinitely smoothing operators. Also, the method of proof differs from that in [22] even for the special case considered there.

More specifically, assume (2.3) and Assumptions 2.4 and 3.1. The a priori assumption on the solution x_0 is the following.

Assumption 5.1. There exists $q > 0$ such that $x_0 \in X_q$ and $\|x_0\|_q \leq E$ for some constant E.

The degree of ill-posedness of T is characterized by the following.

Assumption 5.2. There exists a decreasing, continuous function $w : (0, \infty] \longrightarrow [0, \infty)$ such that

(i) $w(s) = 0 \Leftrightarrow s = \infty$;

(ii) the function $s \mapsto s\,w(s^{-\gamma})$ is convex on $[0, 1]$ for each $\gamma > 0$;

(iii) there exist constants m, $M > 0$, and $0 < \rho \leq 1$ such that

$$m^2 \int w \, d\mathcal{L}_{x,x} \leq \|Tx\|^2 \leq M^2 \int w^\rho \, d\mathcal{L}_{x,x} \quad \text{for all } x \in X_{2q},$$

where $\mathcal{L}$ is the spectral measure of L.

DEFINITION 5.3. For each $r > 0$, define the function $\varphi_r : [0, \infty) \longrightarrow [0, \infty)$ by $\varphi_r(\lambda) = \lambda\, w(\lambda^{-1/2r})$.

Since the spectral measure, $\mathcal{M}^{(r)}$, of L^{2r} is given by $d\mathcal{M}^{(r)}(\lambda) = d\mathcal{L}(\lambda^{1/2r})$, then from Assumption 5.2, we see that the following holds.

Remark 5.4. For each $r > 0$, the function φ_r satisfies Assumption 2.9 with $B^*B = L^{2r}$.

The following technical result will be used in the subsequent theorem to compare the convergence rates of two components of the total error.

LEMMA 5.5. *For any* $p \geq q > 0$ *and constant* $C > 0$,

(i) $\varphi_q^{-1}(C\lambda) = O(\varphi_q^{-1}(\lambda))$;

(ii) $\varphi_p^{-1}(C\lambda[\varphi_q^{-1}(\lambda)]^{(p-q)/q}) = O([\varphi_q^{-1}(\lambda)]^{p/q})$, *as* $\lambda \to 0+$.

Proof. Let $C_1 = \max(1, C)$. Then for any $t > 0$, since w is decreasing, $C\varphi_p(t) = Ctw(t^{-1/2p}) \leq C_1 tw\left((C_1 t)^{-1/2p}\right)$. Hence, $\varphi_p^{-1}\left(C\varphi_p(t)\right) \leq C_1 t$.

Part (i) follows easily. To prove (ii), let $t = \varphi_q^{-1}(\lambda)$. Then,

$$\begin{aligned}\varphi_p^{-1}\left(C\varphi_q(t)t^{(p-q)/q}\right) &= \varphi_p^{-1}\left(Cw(t^{-1/2q})t^{p/q}\right) \\ &= \varphi_p^{-1}\left(C\varphi_p(t^{p/q})\right) \leq C_1 t^{p/q}.\end{aligned}$$

THEOREM 5.6. *Assume Assumptions* 5.1, 5.2, $p \geq q$, *and let* $x_{\alpha,\delta} = (T^*T + \alpha^2 L^{2p})^{-1}T^*y_\delta$. *Then*

$$\|x_{\alpha(\delta),\delta} - x_0\| = O\left(\left[\frac{\varphi_q^{-1}(\delta^{2\rho})}{\varphi_q^{-1}(\delta^2)}\right]^{p/2q}\sqrt{\varphi_q^{-1}(\delta^2)}\right), \quad \textit{where}$$

$$\alpha(\delta) = \frac{1+EM}{E}\delta^\rho\left[\varphi_q^{-1}(\delta^2)\right]^{(p-q)/2q}.$$

Proof. The basic idea is quite natural: to use the denseness of X_p to obtain an approximation, x_1 to x_0. Then, use the basic Tikhonov regularization to estimate x_1 in X_p. However, these approximations have to be carefully done to maintain suitable orders of convergence.

It will be essential to observe that for each $\delta > 0$ there exists $\tau(\delta) \uparrow \infty$ as $\delta \downarrow 0$ such that

$$\tau(\delta) = \left[\varphi_q^{-1}(\delta^2)\right]^{-1/2q} \quad \text{and} \quad \tau(\delta)^{-2q}w(\tau(\delta))^\rho \leq \delta^{2\rho}. \tag{5.1}$$

To obtain asymptotic error estimates it suffices to assume δ is so small that $\delta \leq 1$ and $\tau(\delta) \geq 1$.

Let $x_1 = \int_{[1,\tau(\delta)]} d\mathcal{L}_{x_0}$. Then, $\|x_1 - x_0\|^2 = \int_{(\tau(\delta),\infty)} d\mathcal{L}_{x_0,x_0} \leq 1/\tau(\delta)^{2q}\int_{(\tau(\delta),\infty)} \lambda^{2q}\, d\mathcal{L}_{x_0,x_0}(\lambda)$.

Hence, from Assumption 5.1,

$$\|x_1 - x_0\| \leq \frac{E}{\tau(\delta)^q}. \tag{5.2}$$

From Assumption 5.2,

$$\begin{aligned}\|T(x_1 - x_0)\|^2 &\leq M^2\int_{(\tau(\delta),\infty)} w(\lambda)^\rho\, d\mathcal{L}_{x_0,x_0}(\lambda)\\ &\leq M^2 w(\tau(\delta))^\rho\int_{(\tau(\delta),\infty)} d\mathcal{L}_{x_0,x_0}\\ &\leq E^2M^2w(\tau(\delta))^\rho\tau(\delta)^{-2q}\end{aligned}$$

by Assumption 5.1. Hence, by (5.1),

$$\|T(x_1 - x_0)\| \leq EM\delta^\rho. \tag{5.3}$$

By using the triangle inequality, (5.3), and the data error in (2.3), we obtain

$$\|Qy_\delta - Tx_1\| \leq (1+EM)\delta^\rho. \tag{5.4}$$

To use the data y_δ to recover x_1 by minimizing $\|Tx - y_\delta\|^2 + \alpha^2\|x\|_p^2$ over $x \in X_p$, we need an estimate of the size of $\|x_1\|_p$.

$$\|x_1\|_p^2 = \int_{[1,\tau(\delta)]} \lambda^{2p}\, d\mathcal{L}_{x_0,x_0}(\lambda) \leq \tau(\delta)^{2(p-q)}\int_{[1,\tau(\delta)]} \lambda^{2q}\, d\mathcal{L}_{x_0,x_0}\,(\lambda).$$

Thus,

$$\|x_1\|_p \leq E\tau(\delta)^{p-q}. \tag{5.5}$$

By applying Remark 5.4, the error estimate (5.4) and the a priori norm bound (5.5), to Theorem 2.11, it follows that if

$$\alpha(\delta) = \frac{(1+EM)\delta^\rho}{E\tau(\delta)^{p-q}},$$

then, by (5.1) and Lemma 5.5,

$$\|x_{\alpha(\delta),\delta} - x_1\| \le \sqrt{2}E\tau(\delta)^{p-q}\sqrt{\varphi_p^{-1}\left(\left(\frac{(1+EM)\delta^\rho}{E\tau(\delta)^{p-q}}\right)^2\right)}.$$

The result follows from (5.1), Lemma 5.5, and (5.2).

Now, to obtain the result in [22], set $\rho = 1$, and $w(\lambda) = \lambda^{-2a}$, $a > 0$. Then $\varphi_r(\lambda) = \lambda^{1+a/r}$ and Assumption 5.2 reduces to the one in [22]:

$$m\|x\|_{-a} \le \|Tx\| \le M\|x\|_{-a}. \tag{5.6}$$

Since $\varphi_r^{-1}(\lambda) = \lambda^{r/(a+r)}$, the result in Theorem 5.6 says that, under (5.6) and Assumption 5.1,

$$\|x_{\alpha(\delta),\delta} - x_0\| = O(\delta^{a/(a+q)}) \quad \text{where } \alpha(\delta) = O(\delta^{(a+p)/(a+q)}).$$

We now obtain an analogous result for the general class of "infinitely smoothing" operators discussed in §§3 and 4.

COROLLARY 5.7. *Suppose that there exist positive constants a, b, c, m, and M such that*

$$m^2 \int \exp(-b\lambda^a)\, d\mathcal{L}_{x,x}(\lambda) \le \|Tx\|^2 \le M^2 \int \exp(-c\lambda^a)\, d\mathcal{L}_{x,x}(\lambda)$$

for all $x \in X_{2q}$, and the best approximate solution x_0 satisfies

$$\|x_0\|_q \le E.$$

Given data y_δ satisfying $\|Q(y_\delta - y_0)\| \le \delta$, let $x_{\alpha,\delta} = (T^*T + \alpha^2 L^{2p})^{-1}T^*y_\delta$, where $p \ge q$.

Then

$$\|x_{\alpha(\delta),\delta} - x_0\| = O\left(\left[\log\frac{1}{\delta}\right]^{-q/a}\right), \quad \textit{where}$$

$$\alpha(\delta) = O\left(\delta^\rho\left[\log\frac{1}{\delta}\right]^{-(p-q)/a}\right) \quad \textit{and} \quad \rho = \min\left[\frac{c}{\max\{b,1\}}, 1\right].$$

More specifically,

$$\alpha(\delta) = \frac{1+EM}{E}\delta^\rho\,\eta(\delta)^{p-q}, \quad \textit{where}$$

$$\eta(\delta)^{2q}\exp\left(-\max\{b,1\}\eta(\delta)^{-a}\right) = \delta^2.$$

Proof. Let $\beta = \max\{b,1\}$, so $\rho = \min\{c/\beta, 1\}$.

Set $w(\lambda) = \exp(-\beta\lambda^a)$. Then, for all $\lambda > 0$, $w(\lambda) \le \exp(-b\lambda^a)$ and $\exp(-c\lambda^a) \le w(\lambda)^\rho$.

Now, from Definition 3.3, $s\,w(s^{-\gamma}) = \psi_{\beta,a\gamma}(s)$. Hence, by using Lemma 3.4, Assumption 5.2 is satisfied. The result follows from Theorem 5.6 and Lemma 3.5, by noting that $\varphi_q = \psi_{\beta,a/2q}$ and $\varphi_q^{-1}(\delta^{2\rho}) = O(\varphi_q^{-1}(\delta^2))$. □

Preliminary testing indicates that the numerical accuracy of inversion schemes are improved by using this result.

Acknowledgment. The author is indebted to Scott McCullough and Murali Rao for many stimulating discussions.

REFERENCES

[1] D. ANG, J. LUND, AND F. STENGER, *Complex variable and regularization methods of inversion of the Laplace transform*, Math. Comp., 53 (1989), pp. 589–608.

[2] J. BAUMEISTER, *Stable Solution of Inverse Problems*, Friedr. Vieweg & Sohn, Braunsweig, Wiesbaden, Germany, 1987.

[3] J. V. BECK, B. BLACKWELL, AND C. R. ST. CLAIR, JR., *Inverse Heat Conduction: Ill-posed Problems*, John Wiley, New York, 1985.

[4] M. BERTERO AND E. PIKE, *Exponential-sampling method for Laplace and other dilationally invariant transforms: I. Singular-system analysis*, Inverse Problems, 7 (1991), pp. 1–20.

[5] P. BRIANZI AND M. FRONTINI, *On the regularized inversion of the Laplace transform*, Inverse Problems, 7 (1991), pp. 355–368.

[6] J. R. CANNON, *The One-Dimensional Heat Equation*, in Encyclopedia of Mathematics and its Applications, Vol. 23, Addison-Wesley, Reading, MA, 1984.

[7] A. CARASSO, *Determining surface temperatures from interior observations*, SIAM J. Appl. Math., 42 (1982), pp. 558–574.

[8] R. J. CARROLL, A. C. M. VAN ROOIJ, AND F. H. RUYMGAART, *Theoretical aspects of ill-posed problems in statistics*, Acta Appl. Math., 24 (1991), pp. 113–140.

[9] I. J. D. CRAIG AND J. C. BROWN, *Inverse Problems in Astronomy*, Adam Hilger Ltd., Bristol, UK, and Boston, MA, 1986.

[10] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Systems, 18, Springer-Verlag, New York, 1978.

[11] A. EL JAI AND A. J. PRITCHARD, *Sensors and Controls in the Analysis of Distributed Systems*, John Wiley, New York, 1988.

[12] H. W. ENGL AND H. GFRERER, *A posteriori parameter choice for general regularization methods for solving linear ill-posed problems*, Appl. Numer. Math., 4 (1988), pp. 395–417.

[13] D. S. GILLIAM, Z. LI, AND C. F. MARTIN, *Discrete observability of the heat equation on bounded domains*, Internat. J. Control, 48 (1988), pp. 755–780.

[14] D. S. GILLIAM, B. A. MAIR, AND C. F. MARTIN, *Determination of initial states of parabolic systems from discrete data*, Inverse Problems, 6 (1990), pp. 737–747.

[15] ———, *An inverse convolution method for regular parabolic equations*, SIAM J. Control Optim., 29 (1991), pp. 71–88.

[16] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[17] C. W. GROETSCH, *Generalized Inverses of Linear Operators*, Marcel Dekker, New York, 1977.

[18] M. HAZEWINKEL, *On families of linear systems*, in Algebraic and Geometric Methods in Linear Systems Theory, Lectures in Applied Math., 18, American Mathematical Society, Providence, RI, 1980, pp. 157–189.

[19] M. M. LAVRENTE'V, V. G. ROMANOV, AND S. P. SHISHATSKII, *Ill-Posed Problems of Mathematical Physics and Analysis*, Transl. Math. Monographs, 64 (1986).

[20] J. LOCKER AND P. PRENTER, *Regularization with differential operators, I: general theory*, Math. Anal. Appl., 74 (1980), pp. 504–529.

[21] K. MILLER, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.

[22] F. NATTERER, *Error bounds for Tikhonov regularization in Hilbert scales*, Appl. Anal., 18 (1984), pp. 29–37.

[23] A. NEUBAUER, *An a posteriori parameter choice for Tikhonov regularization in Hilbert scales leading to optimal convergence rates*, SIAM J. Numer. Anal., 25 (1988), pp. 1313–1326.

[24] A. C. M. VAN ROOIJ AND F. H. RUYMGAART, *Regularized inversion of noisy Laplace transforms*, Report 9002, Dept. of Math., Kath. Univ. Nijmegen, the Netherlands, 1990.

[25] Y. SAKAWA, *Observability and related problems for partial differential equations of parabolic type*, SIAM J. Control, 13 (1975), pp. 15–27.

[26] G. TALENTI AND S. VESSELLA, *A note on an ill-posed problem for the heat equation*, J. Austral. Math. Soc. Ser. A, 32 (1982), pp. 358–368.

[27] A. N. TIKHONOV AND V. Y. ARSENIN, *Solutions of Ill-Posed Problems*, John Wiley, New York, 1977.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 148–173, January 1994

005

REGULARIZING MICROSCOPES AND RIVERS*

MARC DIENER†

This paper is dedicated to Professor Jean-Louis Callot, in memoriam.

Abstract. This paper proposes a generalization of the existing geometric studies of resonance. The Riccati equation associated with any second-order linear equation is extended to any C^∞ first-order equation. The Morse-critical point is generalized to any "generic" critical point. The resonant solution becomes a general canard solution. The paper explains how to find the regularizing blowup, and shows how classical special functions become enlarged in *rivers*, i.e., some resurgent solutions of polynomial differential equations. The paper shows a matching principle that connects the slow solutions with these rivers. The method to show the existence of canards is applied for some *Union-Jack equations*, i.e., equations with a critical point where three smooth curves intersect.

Key words. resonance, turning points, singular perturbations, matching, canard, river, Union-Jack, macroscope, microscope, nonstandard analysis, Newton polygon

AMS subject classifications. 34E20, 34E05, 03H10

Introduction. For the problem of turning points of singularly perturbed linear second-order differential equations [21], one is indebted to Kopell [14] for a geometric study of *resonances*, remarkable solutions, discovered by Ackerberg and O'Malley [1], of some boundary value problems in the neighbourhood of the equation's turning points.

At the heart of the proof of the main theorem, Kopell introduces a crucial blowup that turns the singular perturbation problem into a regular one. The first approximation of that new equation is a Hermite equation $\ddot{U} - X\dot{U} + kU = 0$ depending on the real parameter k. It is well known that this equation admits, for noninteger k, a basis of solutions, generally denoted by H_k^+ and H_k^-, with polynomial growth at $+\infty$ and $-\infty$, respectively. When k becomes an integer, these two solutions become equal to each other and turn into a polynomial: the Hermite polynomial. As observed independently by Kopell and Callot [5], it is the "crossing" of these two special functions that makes possible (and necessary) the existence of the resonant solutions. We want to show here that the special functions (or more precisely the inverse of their logarithmic derivatives) admit a generalization, the "rivers," that make it possible to solve this kind of turning point problem for a much more general class of equations.

Let's briefly sketch Callot's approach. One first factors out the invariance of the set of solutions of the second-order differential equation $\varepsilon\ddot{u} - f(x)\dot{u} + g(x,a)u = 0$ by considering the new unknown $y = u/\dot{u}$. So, one considers the slow-fast Riccati equation $\varepsilon dy/dx = -f(x)y + g(x,a)y^2 + \varepsilon$ whose slow curve $\{g(x,a)y^2 - f(x)y = 0\}$ is the union of the two smooth curves $C_1 = \{g(x,a)y = f(x)\}$ and $C_2 = \{u = 0\}$ (see Fig. 1(a)). In those variables, a resonant solution is a slow solution (i.e., infinitely close to the slow curve) staying infinitely close to C_1 on both sides of some "critical point" x_0, i.e., here, of a point such that $f(x_0) = 0$, and such that, near x_0, C_1 is attracting for $x < x_0$ and repelling for $x > x_0$. This is precisely what is usually called a *canard* [3], [10].

Now comes the trick we want to consider here in a general set-up. In order to determine the behaviour of the slow solutions when they come across the halo*[1] of

* Received by the editors September 16, 1991; accepted for publication (in revised form) January 14, 1993.

† Laboratoire Centre National de la Receherche Scientifique, J. A. Dieudonné, Université de Nice, 06108 Nice Cedex 2, France (diener@math.unice.fr).

[1] We use here the methods of nonstandard asymptotics. We recall in the appendix (§4) most nonstandard definitions and results involved here: we indicate by a star * the words defined there.

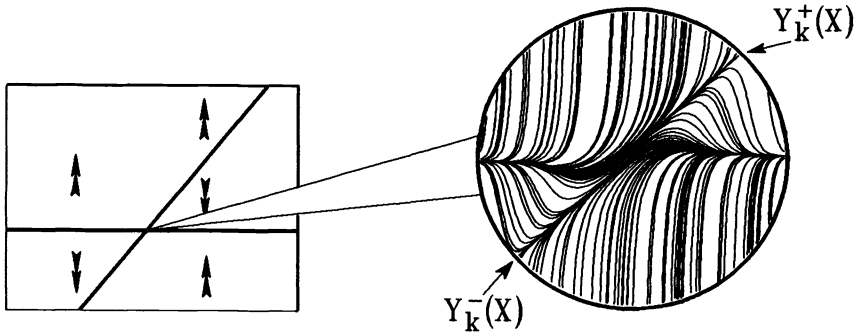


FIG. 1. *Sketch of Callot's proof of the existence of resonant solutions.* (a) *The slow curve of the Riccati equation* $\varepsilon dy/dx = -f(x)y + g(x,a)y^2 + \varepsilon$; (b) *The rivers* $Y^{\pm}(X)$ *of the Riccati–Hermite equation* $dY/dX = -XY + kY^2 + 1$ *for* k *near to* 1. *For larger* k, *the river would have an analogous behaviour at infinity, but would exhibit* $\operatorname{Int} k - 1$ *simple poles.*

the critical point $(x_0, 0)$ (and see how to choose the value of the parameter a in order that such a solution is a canard), one introduces a microscope

$$x - x_0 = \sqrt{\varepsilon}X, \qquad y = \sqrt{\varepsilon}Y$$

that is *regularizing*, in the sense that it turns the singular perturbation of the Riccati equation into a *regular* perturbation of the so-called Riccati–Hermite equation

$$dY/dX = -XY + kY^2 + 1,$$

where $k = k(a) := g(x_0, a)/f'(x_0)$.

Let $Y_k^{\pm} := H_k^{\pm}/H_k^{\pm}$. The function Y_k^+ and Y_k^- can clearly be perceived in Fig. 1(b): they are the only trajectories asymptotic to a (neither horizontal nor vertical) straight line, here $X = kY$. The striking fact about these solutions is that the other nearby solutions depart from it in an "exponential way." This kind of behaviour has been studied since that time for general polynomial differential equations, and these solutions are now called *rivers* (see below).

Let $U \subseteq \mathbb{R}^2$ be a nonempty standard open set, $\varepsilon > 0$ an infinitesimal, and f a function defined on U, with regular ε-shadow expansion*.

In this paper, we consider the general nonlinear singular perturbation problem associated with equation

$$\varepsilon dy/dx = f(x,y) \quad (\varepsilon > 0 \text{ infinitesimal}), \tag{1}$$

where $f(x,y) = f_0(x,y) + \varepsilon f_1(x,y) + \cdots$ with (f_n) a standard sequence of $\mathcal{C}^\infty$ functions of $(x,y) \in U$. One may think of $f(x,y) = F(x,y,\varepsilon)$ for some standard smooth function F, but it could as well be $f(x,y) := G(x,y,\varepsilon,\overline{a})$ with $\overline{a}$ the sum to the smallest term of some diverging series $\sum a_n \varepsilon^n$.

The purpose of this paper is first to show how to associate with any equation (1) one, or if necessary several, *regularizing microscopes* in the neighbourhood of a *critical* point of the slow curve $f_0(x,y) = 0$, one for each growth type of the branches of this curve. This makes it possible to convert the singular perturbation into the regular perturbation of one or more *polynomial* differential equation: the *local models.*

We shall then see that the special functions of the above example are not a "miracle": the local model indeed has rivers; the rivers are solutions that are transcendent, but, as in the case of the special functions of physics such as the Airy function, they have remarkable asymptotic behaviour that lead to efficient numerical approximations. We shall recall the definition and main properties of the rivers at §1.2.

The final problem is to relate the behaviour of the slow solutions of (1) when they are infinitely close to the critical point with the rivers of the local model. This study will lead to Theorem 3.1, a matching principle, which is the central result of this paper. It shows how these rivers generalize the special functions $Y_k^{\pm}$ introduced in the recalled study, and make it possible to specify the behaviour of the slow solutions of (1) near the critical points.

In the introduction to §2, we shall give an example in which the critical point is the intersection of three smooth curves. Using a transversality argument (the "crossing" of two rivers) on the local model, we will show (Corollary 3.2) the existence of canards.

1. Slow-curve branches, Newton polygons, and rivers. The local models near critical points of the slow curve depend strongly on the geometry of this slow curve at that point. We give here some elementary tools that yield crucial information about that geometry from a finite number of terms of the Taylor expansion that give dominant balance near the critical point.

This first section will recall some definitions and define notation that we shall need in the sequel concerning the branches of curves $f(x,y)=0$. Then it will introduce our main tool, the rivers.

1.1. Branches of generic $\mathcal{C}^\infty$ curves.

DEFINITIONS. Let f be a function defined in the neighbourhood of (x_0,y_0). Let φ be a continuous function, whose domain $\mathcal{D}(\varphi)$ is a closed interval with nonempty interior, and having x_0 as one of its ends.We call φ a *branch* at (x_0,y_0) of the curve $\mathcal{C}$ of equation $f(x,y)=f(x_0,y_0)$ if $\varphi(x_0)=y_0$ and $f(x,\varphi(x))=f(x_0,y_0)$ for all $x\in\mathcal{D}(\varphi)$. One specifies that φ is a *positive* or *negative* branch, according to the sign of $x-x_0$ for $x\in\mathcal{D}(\varphi)$, $x\neq x_0$. Let k and r be real numbers, $k\neq 0$ if $r\neq 0$. We say that φ is a branch of type (k,r) at(x_0,y_0) (or, for short, a (k,r)-branch) if $\varphi(x)-y_0\sim k|x-x_0|^r$ when $x\to x_0$ in $\mathcal{D}(\varphi)$ in the case $k\neq 0$, or $\varphi(x)\equiv y_0(=0)$ in the case $k=r=0$.

We denote by $\sim$ the classical relation "is asymptotic to" in the neighbourhood of x_0 or $\pm\infty$, according to the context. In nonstandard words this is equivalent, for standard φ, x_0, $k\neq 0$, and r, to $\varphi(x)=y_0+k|x-x_0|^r(1+\delta)$ with $\delta\simeq 0$ as soon as $x\simeq x_0$ if $x_0\in\mathbb{R}$. If $x_0=\pm\infty$, it means $\varphi(x)=kx^r(1+\delta)$ with $\delta\simeq 0$ as soon as x is unlimited*, of the sign of $\pm\in\{-1,+1\}$.

Puiseux's theory [12], [15] shows that if f is analytic in the neighbourhood of (x_0,y_0), the curve $\mathcal{C}$ is a union of (k,r)-branches with r rational (and possibly the straight line $x=x_0$). The assumption that f is $\mathcal{C}^\infty$ is much too flabby to force an analogous result, as $\mathcal{C}$ could be any closed subset of $\mathbb{R}^2$. Nevertheless one recovers an analogous result if f is not too degenerate; for that purpose, we introduce now what we shall call the *lower* Newton polygon $\mathcal{N}(f\,;x_0,y_0)$. Proposition 1.2 will show how to determine the growth types of the various *branches* of $f=0$ using this lower Newton polygon and the polynomial ${}_rf_0$ that we shall now define (see also Fig. 2).

DEFINITION. Let f be a $\mathcal{C}^\infty$ function on a neighbourhood of $(x_0,y_0)\in\mathbb{R}^2$. We call the *Taylor set* of f at the point (x_0,y_0) the set of couples of integers $\mathcal{T}(f;x_0,y_0):=\{(m,n)\neq(0,0)|f^{(m+n)}_{x^my^n}(x_0,y_0)\neq 0\}$; if $(x_0,y_0)=(0,0)$ we'll just write $\mathcal{T}(f)$ for

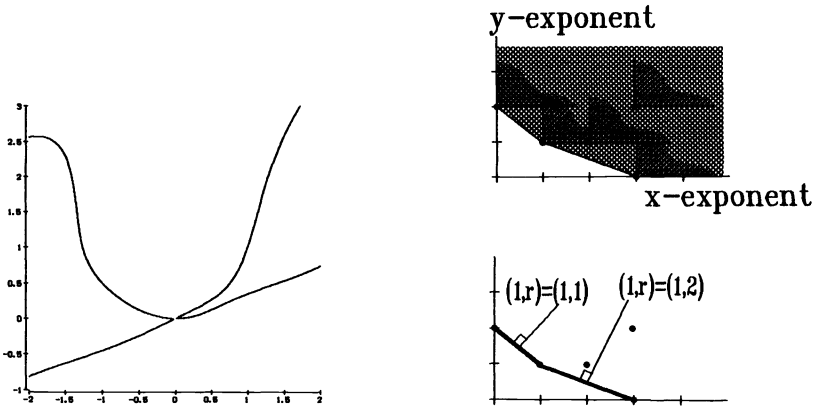


FIG. 2. *The curve $f = 0$ and the lower Newton polygon at $(0,0)$ associated with $f(x,y) = (x-2y)(2y-x^2)+x^3y^2\cos(x-y)$; here ${}_1f(x,y) = 2y(x-2y)$ and ${}_2f(x,y) = x(2y-x^2)$. The curve shows two negative and two positive branches at $(0,0)$, of growth-type $x/2$ and $x^2/2$ at $(0,0)$.*

$\mathcal{T}(f\,;0,0)$. Let $\mathcal{E} = \mathcal{E}(f;x_0,y_0)$ be the convex hull in $\mathbb{R}^2_{mn}$ of the union of all quadrants $m \geq \overline{m}$ and $n \geq \overline{n}$ for all $(\overline{m},\overline{n}) \in \mathcal{T}(f;x_0,y_0)$.

The (*lower*) *Newton polygon* of f at the point (x_0,y_0) is the union $\mathcal{N}(f;x_0,y_0)$ of all oblique segments that build the border of $\mathcal{E}(f;x_0,y_0)$. If $(x_0,y_0) = (0,0)$, we'll just write $\mathcal{N}(f)$ for $\mathcal{N}(f\,;0,0)$.

We call *coslope* of a nonhorizontal segment σ the real number r such that $(1,r)$ is orthogonal to σ. If r is the coslope of some segment σ contained in $\mathcal{N}(f;x_0,y_0)$ we say that r is a coslope of $\mathcal{N}(f;x_0,y_0)$; r is then necessarily a positive rational (see Fig. 2). If there is no point $(m,n) \in \mathcal{N}(f;x_0,y_0)$ such that $n = 0$ (i.e., if $(y-y_0)$ is a factor of $f(x,y)-f(x_0,y_0)$, we shall say that $r = 0$ is (also) a coslope of $\mathcal{N}(f;x_0,y_0)$.

In §1.2 we shall define an *upper* Newton polygon for f any *polynomial.*

DEFINITION. We assume that $\mathcal{T}(f;x_0,y_0) \neq \emptyset$. We define the *r-valuation* of f at the point (x_0,y_0) as the minimum $\mu_r(f;x_0,y_0)$ of all $m+rn$ for $(m,n) \in \mathcal{T}(f;x_0,y_0)$. We'll just write $\mu_r(f)$ when $(x_0,y_0) = (0,0)$. We denote by ${}_rf(X,Y)$ the polynomial, sum of all monomials of the Taylor expansion of f at the point (x_0,y_0) with r-valuation equal to $\mu := \mu_r(f;x_0,y_0)$:

$$ {}_rf(X,Y) := \sum_{m+rn=\mu} \frac{1}{m!n!} f^{(m+n)}_{x^m y^n}(x_0,y_0) X^m Y^n. $$

Remark. The polynomial ${}_rf$ is *r-homogeneous* of *r-degree*$=\mu := \mu_r(f;x_0,y_0)$, that is, ${}_rf(\lambda X,\lambda^r Y) = \lambda^\mu {}_rf(X,Y)$ for any $\lambda > 0$. For $r \neq 0$, its Taylor set is contained in the segment σ of $\mathcal{N}(f;x_0,y_0)$ with coslope r if any and then $\mathcal{N}({}_rf;x_0,y_0) = \sigma$, or is just equal to one point of $\mathcal{N}(f;x_0,y_0)$.

LEMMA 1.1. *Let $f \in \mathcal{C}^\infty$ be a standard function, such that $f(0,0) = 0$ and $\mathcal{T}(f) \neq \emptyset$. Let $\alpha > 0$, $\alpha \simeq 0$, and $r \in \mathbb{R}^+$ standard. Set $\mu = \mu_r(f)$, $x = \alpha X$, and $y = \alpha^r Y$. For all limited (X,Y), one has $f(x,y) = \alpha^\mu({}_rf(X,Y) + \phi)$.*

Proof. As f and r are standard, so* is μ. Let N be a standard integer, $N \geq \mathrm{Max}\{\mu, \mu/r\}$. We write f as $f = T_N f + R_N f$, where $T_N f$ is the Taylor polynomial

of f at $(0,0)$ of degree N, and $R_N f$ is the rest of the Taylor expansion. For all infinitesimal (x,y), one has $R_N f(x,y) = (\mathrm{Max}\{|x|,|y|\})^{N+1}\mathcal{L}$ ($\mathcal{L}$ denoting a generic limited real number), and thus $R_N f(\alpha X, \alpha^r Y) = \alpha^\mu \phi$ as $N+1 \geq \mathrm{Max}\{\mu, \mu/r\}$.

Thus, it suffices to show the lemma in the case where f is a standard polynomial $T_N f(x,y)$; this case is trivial. □

PROPOSITION 1.2. *Let $f \in \mathcal{C}^\infty$ be any function defined on some neighbourhood $(0,0)$, such that $f(0,0) = 0$ and $\mathcal{T}(f) \neq \emptyset$. Let k and r be nonzero numbers. If the curve $f = 0$ admits a negative (k,r)-branch at $(0,0)$, then r is a coslope of $\mathcal{N}(f)$ and ${}_r f(-1,k) = 0$.*

Proof. By transfer*, we may assume that f and the branch φ are standard. Since φ is standard, and as $\varphi(x) \sim k|x|^r$, the real numbers k and r are standard. Let $\alpha > 0$ be any infinitesimal. The hypothesis $\varphi(x) \sim k|x|^r$ for $x < 0$ implies that $\varphi(-\alpha) = k\alpha^r(1+\phi)$. Let $\mu := \mu_r(f\,; 0,0)$. Lemma 1.1 implies that

$$f(-\alpha, \varphi(-\alpha)) = f(-\alpha, k\alpha^r(1+\phi)) = \alpha^\mu({}_r f(-1, k(1+\phi)) + \delta)$$

with $\delta \simeq 0$. Dividing by α^μ the relation $f(\alpha, \varphi(\alpha)) = 0$, one gets ${}_r f(-1, k(1+\phi)) \simeq 0$. Since ${}_r f$ is a standard continuous function, one has

$${}_r f(-1,k) = {}_r f({}^o(-1, k(1+\phi))) = {}^o({}_r f(-1, k(1+\phi))) = 0.$$

Thus ${}_r f(-1,k) = 0$; since $k \neq 0$, the polynomial ${}_r f(X,Y)$ cannot be just equal to one monomial; r is thus a coslope of $\mathcal{N}(f)$. □

Here is a somewhat more general result that we will *not* use here.

PROPOSITION 1.3. *Let $f \in \mathcal{C}^\infty$ be standard such that $f(0,0) = 0$ and $\mathcal{T}(f) \neq \emptyset$. Let $(\alpha, \phi) \simeq (0,0)$, $\alpha > 0$, such that $f(\alpha,\phi) = 0$. Assume that $\phi = \alpha^r$ with r appreciable. Then some standard $k_0 \neq 0$ and some $r_0 = {}^o r$ exist such that $\phi = k_0\alpha^{r_0}(1+\phi)$; r_0 is a coslope of $\mathcal{N}(f; x_0, y_0)$ and ${}_{r_0} f(1, k_0) = 0$.*

Proof. Let $r_0 := {}^o r \neq 0$, and $\mu := \mu_{r_0}(f)$, which is standard, as are f and r_0. Let $N \geq \mathrm{Max}\{\mu, \mu/r\}$. Using the factorization $\alpha^r = \alpha^{r-r_0}\alpha^{r_0}$ and reasoning as in proof of Lemma 1.1, one checks that $f(\alpha,\phi) = \alpha^\mu({}_{r_0} f(1, \alpha^{r-r_0}) + \phi)$. Since $f(\alpha,\phi) = 0$, dividing by α^μ, one sees that ${}_{r_0} f(1, \alpha^{r-r_0}) \simeq 0$. Since $p(K) := {}_{r_0} f(1,K)$ is a nonconstant standard polynomial, thus unlimited for any unlimited K, $k := \alpha^{r-r_0}$ has to be limited, thus near-standard in $\mathbb{R}$. Let $k_0 := {}^o k$. One has $0 = {}^o({}_{r_0} f(1,k)) = {}_{r_0} f(1, {}^o k) = {}_{r_0} f(1, k_0)$. We have to show that $k_0 \neq 0$. If we set $s := 1/r$ and $g(x,y) := f(y,x)$, as $\alpha = \phi^s$, the previous reasoning shows that $l := \alpha^{s-{}^o s}$ is near-standard in $\mathbb{R}$, and $\alpha = l\phi^{s_0}$ with $s_0 := {}^o s$. Thus $\phi = k\alpha^{r_0} = kl^{r_0}\phi^{r_0 s_0} = kl^{r_0}\phi$, with k and l limited. Finally, $1 = kl^{r_0}$ and ${}^o k \neq 0$. □

1.2. Rivers of polynomial differential equations. The polynomial differential equations that occur as local models for the behaviour of the slow solutions near the critical points of the slow curve have a few remarkable solutions called "rivers" that "organize" the qualitative behaviour of the other solutions. These solutions are of polynomial growth and attract or repel exponentially the nearby solutions. These rivers generalize the (logarithmic derivative) of the special functions, the distinguished solutions of (the Riccati equations associated with) the second-order linear differential equations occuring in mathematical physics. We recall here briefly the definition of rivers of a polynomial differential equation and the effective methods to determine the rivers using an *upper* Newton polygon and the polynomials ${}_r P$ associated with P that we also define now. For more details see [11], [7], and [20].

Let us consider the following differential equation:

$$\frac{dY}{dX} = P(X,Y) \tag{2}$$

with $P(X,Y)$ any polynomial with real coefficients.

DEFINITION (see [4]). Let $\overline{Y}$ be any solution of (2), and k and r two real numbers, with $k \neq 0$ if $r \neq 0$. We say that $\overline{Y}$ is a solution of type (k,r) of (2) at $X = +\infty$ (respectively, $X = -\infty$) if there exists a real number X_0 such that $\overline{Y}$ is defined on $[X_0, +\infty)$ (respectively, $(-\infty, X_0]$), and if

$$\overline{Y}(X) \sim k|X|^r \quad \text{at } X = +\infty \quad (\text{respectively, } X = -\infty).$$

We say that $\overline{Y}$ is a *river* of type (k,r) of (2) at $X = +\infty$ (respectively, $X = -\infty$) if $\overline{Y}$ is a solution of type (k,r) of (2) at $X = +\infty$ (respectively, $X = -\infty$) and if

$$\lim_{X \to \pm\infty} X \cdot P'_Y(X, \overline{Y}(X)) = \pm\infty.$$

We shall use the expression "of growth type kX^r at $X = \pm\infty$" as a synonym "of type (k,r) at $X = \pm\infty$."

This last hypothesis suffices to ensure the exponential attracting or repelling of nearby solutions (see [4]). Here come two definitions introducing objects useful to determine the rivers of a polynomial differential equation.

DEFINITION. Let $P := \sum a_{mn} X^m Y^n$, and let $\mathcal{D} = \mathcal{D}(P)$ be the convex hull in $\mathbb{R}^2_+$ of the half-lines $m \leq \overline{m}$ and $n = \overline{n}$, for all $(\overline{m}, \overline{n})$ such that $a_{\overline{mn}} \neq 0$. We call the *upper* Newton polygon of the polynomial P the union $\mathcal{M}(P)$ of all oblique segments of the border of $\mathcal{D}$.

If σ is a segment contained in $\mathcal{M}$ of coslope r, we say that r is a coslope of $\mathcal{M}(P)$. If $(Y - k)$ is a factor of P for some $k \in \mathbb{R}$, we shall say that $r = 0$ is (also) a coslope of $\mathcal{M}(P)$.

DEFINITION. Let $r \in \mathbb{R}$ and $P(X,Y) := \sum a_{mn} X^m Y^n$ for any polynomial. We define the r-degree of P to be the number $\partial_r P := \operatorname{Max}\{m + rn \mid a_{mn} \neq 0\}$. We set

$${}^r P(X,Y) = \sum_{m+rn=\partial_r P} a_{mn} X^m Y^n.$$

Remarks. As we already pointed out, for any smooth function f at (x_0, y_0), the polynomial ${}_r f$ is *r-homogeneous of r-degree=μ* $:= \mu_r(f; x_0, y_0)$, and, if $r \neq 0$, its Taylor set is contained in the segment σ of $\mathcal{N}(f; x_0, y_0)$ with coslope r if any and then $\mathcal{N}({}_r f; x_0, y_0) = \sigma$, or is just equal to one point. Thus $\mathcal{N}({}_r f; x_0, y_0) = \mathcal{M}({}_r f)$, ${}^r({}_r f) = {}_r f = {}_r({}_r f)$ and $\partial_r({}_r f) = \mu_r(f\,; x_0, y_0)$.

By construction of $\mathcal{N}({}_r f; x_0, y_0)$, any of its coslopes r are nonnegative; this was done intentionally, for the sake of simplicity, choosing to define $\mathcal{E}$ as the convex hull of a union of quadrants. Indeed, a curve of type (k,r) at (x_0, y_0) with $r < 0$ would tend to infinity when x goes to x_0; the problem would no longer be local so we would need some rigidity on f, such as "f is a polynomial in y^α" (this is beyond the purpose of this paper); nevertheless, in §3 we give an example on singular deformation to show how curves of type (k,r) with $r < 0$ arise.

As we consider here a river solution of *polynomial* differential equations, it is no longer necessary to exclude the case $r < 0$: this is why the upper Newton polygon was introduced as the convex hull of horizontal half-lines instead of quadrants as in the previous case. So $\mathcal{M}(P)$ may have some negative r as coslope.

There is a result [11], [4] in which (a) and (b) are the analogs, for rivers, of Proposition 1.2 for the branches of the slow curve at a critical point; condition (c) is related to the behaviour of the other solutions with respect to the river: for X unlimited (that is, for $x := \varepsilon X \not\simeq 0$, where $\varepsilon \simeq 0$), the river behaves as a slow solution of a (usually other) slow-fast differential equation. As rivers are concerned with the behaviour of the equation at infinity, it is the *upper* Newton polygon that is of interest here. Proposition 2.1, Theorem 2.2, and finally the matching principle 3.1 will make precise the relation between growth type of a branch at a critical point and rivers of the related local model under the regularizing blowup.

PROPOSITION 1.4. *If* (2) *has a solution of type* (k, r) *at* $X = \pm\infty$, *then* (a) r *is a coslope of the upper Newton polygon* $\mathcal{M}(P)$; (b) ${}^rP(\pm 1, k) = 0$. *Moreover, this solution is a river at* $X = \pm\infty$ *if and only if* (c) $c(r) := 1 - r + \partial_r P > 0$.

Conversly, if beyond properties (a), (b), and (c) above, one has

$$ {}^rP'_Y(1, k) \neq 0, $$

then the *rivers-existence theorem* [11] implies that (2) has indeed a river of type (k, r).

One of the interests of the river solutions lies in the fact that they admit, as most special functions, a (diverging) asymptotic expansion of Gevrey type [7], and that they are resurgent [13], [6] and, in particular, summable in the sense of the summation of diverging series [17]. In practice, in the "good cases," i.e., if ${}^rP'_Y(\pm 1, k) \neq 0$ and for $r = p/q$ with p and $q > 0$ integers, any river has an asymptotic expansion $kx^r \sum_{n\geq 0} a_n x^{-n/q}$ for which it is easy to compute as many terms as desirable, with help, for example, from programs such as *Maple* or *Mathematica*. The fact that the expansion is of Gevrey type, that is, (a_n) grows no faster as $n\,!$, implies among other things [19] that the error committed when truncating the sum is of order equal to the first "neglected" term. So one may compute approximations of the rivers of order of the smallest term $a_n x^{-n/q}$ ("summation up to the smallest term"), or excellent approximation with few terms, as for the diverging expansions of special functions. The fact that the expansion is resurgent is related to its "transasymptotic" expansion (exponential corrections). So rivers lend themselves nicely to numerical computations.

2. Regularizing microscopes and local models.

Examples and definition. Let us consider the following example of slow-fast differential equation (Fig. 3):

$$ \varepsilon dy/dx = (y - ax)(y - bx)(y - cx) + p(x, y), \quad \text{with } abc \neq 0 \tag{3} $$

with $a < b < c$ three standard real numbers, and p a standard $\mathcal{C}^\infty$ function with $\mathcal{T}_3(p) \equiv 0$, i.e., zero Taylor polynomial of degree 3. The slow curve is the union of three smooth curves passing through the origin, and tangent to the three straight lines $y = ax$, $y = bx$, and $y = cx$, respectively, that motivate the name, *Union-Jack* equation, that we give to it; $(0, 0)$ is a more degenerate critical point than in the case of a Riccati equation. Let us perform the following change of variable:

$$ \alpha X = x, \quad \alpha Y = y, \quad \text{with } \alpha = \varepsilon^{1/3}. \tag{4} $$

Since $\varepsilon \simeq 0$, this operates as a microscope on the phase-space, and the image is a blowup of this phase space. The behaviour of the solutions for (X, Y) limited corresponds to the behaviour of solutions of (3) in the α-galaxy of $(0, 0)$. Equation (3) becomes

$$ dY/dX = (Y - aX)(Y - bX)(Y - cX) + \varepsilon^{-1} p(\alpha X, \alpha Y). \tag{5} $$

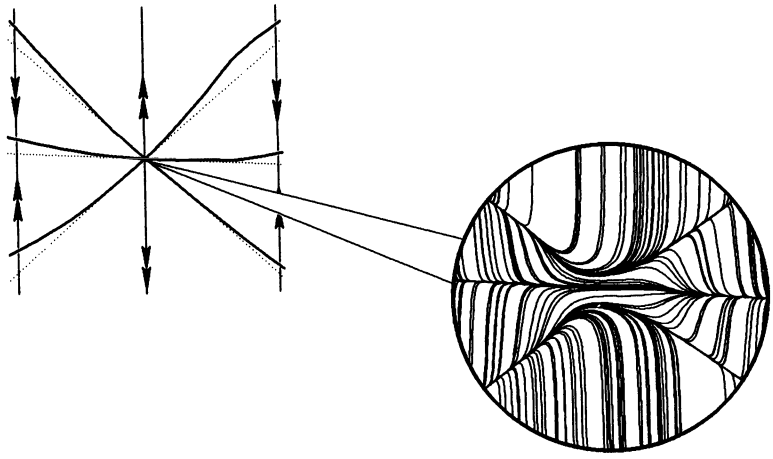


FIG. 3. *Slow-curve of a Union-Jack equation* 3 *and its image under a regularizing blowup.*

Since, by assumption, the Taylor polynomial of degree 3 of p is zero, we see that the term $\varepsilon^{-1}p(\alpha X, \alpha Y)$ is infinitesimal for (X, Y) limited. In other words, the blow up (4) made it possible, in the vicinity (i.e., infinitesimal) of the critical point, to convert the singular perturbed equation (3) into a regular one. The shadow of the solutions of (5) are solutions of the standard differential equation

$$dY/dX = (Y - aX)(Y - bX)(Y - cX). \tag{6}$$

And this is now a polynomial differential equation. An easy study of its rivers shows that this equation has three rivers at $X = +\infty$, and as many at $X = -\infty$, all being asymptotic to kX^r for some $k = a, b, c$, and $r = 1$. Theorem 3.1 will show that any slow solution following $\{y = bx\}$ for $x < 0$ (attracting solution), or $\{y = ax\}$ or $\{y = cx\}$ for $x > 0$ (repelling solution), is infinitely close to the corresponding river at the scale of (X, Y).

In this study, we want to elucidate a double "miracle": how to associate with any critical point of a large class such a desingularizing blowup, and how to approximate the slow solutions in the very vicinity of the critical point by a river of some standard, polynomial, differential equation. As an example of possible application of that method, we shall show how to deduce the existence of *canards* in a one-parameter family of Union-Jack equations.

DEFINITION. Let $\varphi(x) \sim y_0 + k(x - x_0)^r$ be a positive (respectively, negative) branch of the slow curve $\{f_0 = 0\}$ of (1) in the neighbourhood of (x_0, y_0). We call this the *regularizing microscope* of (1) in the neighbourhood of (x_0, y_0) with respect to the branch φ any change of variable $x = x_0 + \alpha X$, $y = y_0 + \beta Y$, with $\alpha > 0$, $\alpha \simeq 0$ and any $\beta > 0$, that converts that equation into $dY/dX = F(X, Y)$, with F of class S^0, ${}^oF \neq 0$, such that the standard equation $dY/dX = {}^oF(X, Y)$ admits a solution asymptotic to kX^r at $X = +\infty$ (respectively, $X = -\infty$). Equation $dY/dX = F(X, Y)$ will be called a *regularizing blowup* of (1) (by the considered regularizing microscope).

Remark. For small enough $\alpha = \beta$ it is always possible to get an infinitesimal F, that is, a regular perturbation of the trivial equation $dY/dX = 0$. This way to "regularize" the singular perturbation is of course of low interest: it is just a way to express the classical theorem of local straightening of locally Lipschitz differential equations' solutions. This is why we ask for ${}^oF \neq 0$; as we shall see, this needs a more

subtle choice of α and β, neither too large nor too small.

2.1. Case $f = f_0$: the simple equation. We first consider the case of the *simple* equation

$$\varepsilon dy/dx = f_0(x,y) \quad (\varepsilon > 0 \text{ infinitesimal}), \tag{7}$$

i.e., the case where f is a standard function. It is easy to see that the microscope

$$x = x_0 + \alpha X, \qquad y = y_0 + \alpha^r Y, \tag{8}$$

centered at a standard point (x_0, y_0) of the slow curve, yields the following blowup of (7):

$$dY/dX = (\alpha^{1-r}/\varepsilon) f(x_0 + \alpha X, y_0 + \alpha^r Y) = (\alpha^{1-r+\mu}/\varepsilon)\,({}_r f_0(X,Y) + \delta)\,,$$

with $\mu = \mu_r(f_0\,; x_0, y_0)$ and $\delta \simeq 0$ for any limited (X,Y). In order to get a regularizing microscope, one has to consider those choices of α such that $\alpha^{1-r+\mu}/\varepsilon$ is appreciable, for example, equal to 1, which means that $\alpha = \varepsilon^{1/(1-r+\mu)}$. The crucial point of this study consists in observing that for some convenient choices of r the shadow of the near-standard equation associated in that way is a *polynomial equation exhibiting rivers.*

PROPOSITION 2.1 (case of the simple equation). *Assume f is standard (i.e., $f = f_0$) and that $(0,0)$ belongs to the slow curve $\{f_0 = 0\}$. Let $r \geq 0$ be any coslope of the lower Newton polygon $\mathcal{N}(f_0)$, $\mu = \mu_r(f_0)$, and k be any root of (the algebraic) equation ${}_r(f_0)(\pm 1, k) = 0$. Let $s = 1/(1-r+\mu)$. In that case $s > 0$, $\alpha := \varepsilon^s$ is infinitesimal, the microscope*

$$x = \varepsilon^s X, \qquad y = \varepsilon^{sr} Y \tag{9}$$

is regularizing for (7), and the resulting blowup is infinitely close to the polynomial equation

$$dY/dX = {}_r f_0(X,Y). \tag{10}$$

If $({}_r f_0)'_Y(\pm 1, k) \neq 0$, this equation has a river of type kX^r at $X = \pm\infty$.

Proof. One has $1/(1-r+\mu) > 0$: let $(m,n) \in \mathcal{N}(f_0)$ be the point of the segment σ of the lower Newton polygon of f_0 of coslope r that has the largest ordinate n; one has $n \geq 1$ and thus

$$1 - r + \mu = 1 - r + \mu_r(f_0) = 1 - r + m + rn = (1+m) + r(n-1) \geq 1.$$

Thus $1 - r + \mu$ is positive and is standard, whence $\alpha := \varepsilon^{1/(1-r+\mu)} \simeq 0$.

The microscope is regularizing: Lemma 1.1 implies that

$$f(x,y) = \alpha^{\mu}({}_r f(X,Y) + \delta),$$

with $\delta \simeq 0$ for all limited (X,Y), whence

$$dY/dX = \alpha^{1-r}\alpha^{\mu}\varepsilon^{-1}\,({}_r f_0(X,Y) + \delta) = {}_r f_0(X,Y) + \delta =: F(X,Y),$$

and ${}^{\circ}F = {}_r f_0$. The *short-shadow lemma** [9] implies that the shadows of the solutions at this scale really are solutions of (10).

Equation $dY/dX = {}^oF(X,Y)$ does have rivers: the polynomial ${}_rf_0$ is r-homogeneous (of r-degree $\mu := \mu_r(f_0)$), and its *upper* Newton polygon $\mathcal{M}({}_rf_0)$ is just equal to the segment σ. Since $({}_rf_0)'_Y(\pm 1, k) \neq 0$, the *rivers-existence theorem* [11] implies the existence of one river of type kX^r at $X = \pm\infty$. □

Examples. Case of a regular point of the slow curve. Assume $f = f_0$, i.e., f is standard, and that $(0,0)$ is a regular point of the slow curve of (1), that is, $f_0(0,0) = 0$ and $f_{0x}'f_{0y}'(0,0) \neq 0$. The implicit-function theorem implies that the slow curve $\{f_0 = 0\}$ is, in the neighbourhood of $(0,0)$, a standard smooth curve tangent to the straight line $\{y = kx\}$, with $k = -A/B$, where $A := f_{0x}'(0,0)$ and $B := f_{0y}'(0,0)$. So, it has two branches (one for $x \geq 0$ and one for $x \leq 0$), of type $(k,1)$.

The microscope $x = \varepsilon X$, $y = \varepsilon Y$ is regularizing and turns (1) into a near-standard equation, the shadows of the trajectories of which being solutions of

$$\frac{dY}{dX} = AX + BY.$$

This equation has the explicit solution $\overline{Y}(X) := -XA/B - A/B^2$, which is a river, both at $X = +\infty$ and at $X = -\infty$, of the same type as the two branches of the slow curve.

Case of a fold point. In [16], Mishchenko and Rosov study the behaviour of the solutions of (1) near a fold point at $(0,0)$, that is, in the case $f_{0x}'(0,0) \neq 0$ and $f_{0y}'(0,0) = 0$, but $f_{0y^2}''(0,0) \neq 0$: the slow curve looks like a horizontal parabola: it has two branches (assumed to be negative) of type $(k, \frac{1}{2})$.

Let $s_0 = \frac{2}{3}$; the microscope $x = \varepsilon^{s_0}X$, $y = \varepsilon^{s_0/2}Y$ is regularizing at such a fold-point. It turns (1) into a near-standard equation, the shadows of the trajectories being solutions of

$$\frac{dY}{dX} = AX + BY^2$$

with $A = f_{0x}'(0,0)$ and $B = \frac{1}{2}f_{0y^2}''(0,0)$. It is a Liouville equation that exhibits [11] two families of rivers (Fig. 4), one containing just one isolated river (asymptotic to $y = +\sqrt{-x}$ on the figure, and that the authors call the "dividing solution," which they are able, in that case, to express in terms of Bessel functions), and the other containing an infinity of rivers, all asymptotic to each other (and to $-\sqrt{-x}$ on the figure). The types of the two families of rivers are equal to those of the two branches of slow curve.

2.2. Regular deformations. We now come back to the general case, where $f = f_0 + \varepsilon f_1 + \cdots$ that we shall consider in different ways, according to whether the microscope (9) used to regularize the simple (or *simplified*) equation (7), obtained by replacing f by f_0, is still regularizing for the *complete* equation (1) or not.

Let us first consider the example of the following equation:

$$\varepsilon dy/dx = \pm(y-x)(y-x^2) + \varepsilon(a+bx), \quad \text{with } a,b \in \mathbb{R}. \tag{11}$$

The slow curve, with equation $(y-x)(y-x^2) = 0$, exhibits some branches at $(0,0)$ of type x^r with $r = 1$ and $r = 2$. Let's consider successively the cases $r = 1$ and $r = 2$.

For $r = 1$, considering the simplified equation leads to selecting $s = \frac{1}{2}$, the microscope $x = \varepsilon^{1/2}X$, $y = \varepsilon^{1/2}Y$ yields the blowup

$$dY/dX = \pm(Y-X)(Y-\varepsilon^{1/2}X^2) + a + \varepsilon^{1/2}bX, \tag{12}$$

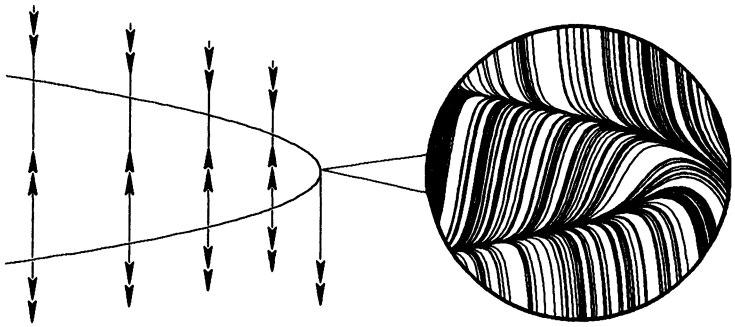


FIG. 4. *Slow-fast equation with slow curve exhibiting a fold point, and its image under a regularizing microscope. This last equation has a unique river asymptotic to* $\sqrt{-x}$, *and an infinity of rivers all asymptotic to* $-\sqrt{-x}$, *and actually with the same asymptotic expansion: they are exponentially close to each other.*

which, for limited a and b, is infinitely close to a standard equation (i.e., is a regular perturbation), namely, $dY/dX = \pm Y(Y - X) + {}^{\circ}a$.

For $r = 2$, considering the simplified equation suggests once more letting $s = \frac{1}{2}$; the microscope $x = \varepsilon^{1/2}X$, $y = \varepsilon Y$ transforms (11) into

$$dY/dX = \mp X(Y - X^2) \pm \varepsilon^{1/2}Y(Y - X^2) + a\varepsilon^{-1/2} + bX. \tag{13}$$

We notice here that if a is standard nonzero, the term $a\varepsilon^{-1/2}$ is unlimited: the microscope that is regularizing for the simplified equation associated with (11) is no longer regularizing for the complete equation (11).

In terms of the following definition, the "deformation term" $\varepsilon(a + bx)$ is a *regular deformation* for the growth-type x^1, and a *singular deformation* for the growth-type x^2 (if ${}^{\circ}a \neq 0$).

DEFINITION. We say that (1) is a *regular deformation at the point* (x_0, y_0) *of the simple equation* (7) *for the growth type* x^r, if, for $s_0 = 1/(1 - r + \mu_r(f_0; x_0, y_0))$, the microscope

$$x = x_0 + \varepsilon^{s_0}X, \qquad y = y_0 + \varepsilon^{s_0 r}Y \tag{14}$$

is regularizing for (1). If not, we call it a *singular deformation at the point* (x_0, y_0) *of* (7) *for the growth type* x^r.

The purpose of the following theorem is first to show that only an explicit finite number of terms of the expansion of $f = \varepsilon f_1 + \varepsilon^2 f_2 + \cdots$ may introduce a singular deformation of the simple equation for a given branch-growth type. Then it establishes a relation existing between the local model of the simple equation and that of the complete equation, in case of regular deformation. This will make it possible to show (Corollary 2.3) that these two models have essentially the "same kind" of rivers.

THEOREM 2.2. *Let* $s_0 = 1/(1 - r + \mu_r(f_0))$. *A necessary and sufficient condition for* (1) *to be a regular deformation at the point* $(0, 0)$ *for the simplified equation for the growth type* x^r *is that, for all* p *such that* $1 \leq p < 1 + s_0(r - 1)$, *the following inequality holds:*

$$p - 1 + s_0\big(1 - r + \mu_r(f_p)\big) \geq 0. \tag{15}$$

If so, the blowup $dY/dX = F(X, Y)$ *of equation* (1) *by the microscope* (14) *is infinitely close to some standard polynomial differential equation*

$$dY/dX = P(X, Y), \quad \text{with } {}^{r}P(X, Y) = {}_{r}f_0,$$

called the local model *of* (1) *at the point* $(x_0, y_0) = (0,0)$ *for branches of growth type* $|x - x_0|^r$.

Proof. Let $p^* > 1 + s_0(r-1)$ be some fixed standard integer and denote $\mu(f_p)$ just by μ_p. One has

$$f(x,y) = f_0(x,y) + \varepsilon f_1(x,y) + \cdots + \varepsilon^{p^*}\left(f_{p^*}(x,y) + \phi\right),$$

for all $(x,y) \simeq (0,0)$; thus, for all limited (X,Y)

$$\begin{aligned} F(X,Y) &= \varepsilon^{-1+s_0(1-r)} f(\varepsilon^{s_0}X, \varepsilon^{rs_0}X) \\ &= \sum_{p<p^*} \varepsilon^{-1+s_0(1-r)+p+s_0\mu_p}\left({}_r f_p(X,Y) + \phi\right) \\ &\quad + \varepsilon^{p^*-1+s_0(1-r)}\left(f_{p^*}(0,0) + \phi\right) \\ &\simeq \sum_{p<p^*} \varepsilon^{-1+s_0(1-r)+p+s_0\mu_p}\left({}_r f_p(X,Y) + \phi\right). \end{aligned}$$

So, the deformation is regular if and only if all the exponents of this later sum are positive or zero, that is, if (15) holds for all $p < 1 + s_0(r-1)$. If so, let $\Pi_* := \{p < p^* \mid p + s_0\mu_p = 1 + s_0(r-1)\}$; so, for any limited (X,Y),

$$F(X,Y) \simeq \sum_{p\in\Pi_*} {}_r f_p(X,Y) =: P(X,Y).$$

Since this defines a standard polynomial P, one finally has ${}^oF = P$.

Still in that case, let $p \geq 1$, $p \leq 1 + s_0(r-1)$, and assume that f_p brings a noninfinitesimal contribution to $F(X,Y)$, that is, assume $-1 + s_0(1-r) + p + s_0\mu_r(f_p) = 0$. By definition of s_0, we thus have $p + s_0\mu_r(f_p) = 0 + s_0\mu_r(f_0)$, whence $\mu_r(f_0) - \mu_r(f_p) = p/s_0$, which is strictly positive, since $p \geq 1$ and $s_0 > 0$ as we saw in Proposition 2.1. Thus $\mu_r(f_0) > \mu_r(f_p)$, and ${}^rP = {}^r({}^oF) = {}_r f_0$. □

Example. If $r \leq 1$, the condition $1 \leq p < 1 + s_0(r-1)$ is never satisfied, and thus any deformation is automatically satisfied for the growth type x^r: at a regular point or at a fold point, or at a Morse point ($f(0,0) = 0$, $\mathrm{Jac}\,(f)(0,0) = (0,0)$, but $\mathrm{hess}\,(f)(0,0) \neq 0$) any deformation is regular for all growth types of the branches that reach such a point. This explains why all the existing studies of canards never came across the problem of singular deformations.

COROLLARY 2.3 (regular deformations). *Assume* $(x_0, y_0) = (0,0)$. *Let* $r \geq 0$ *be a coslope of the lower Newton polygon* $\mathcal{N}(f_0)$, *and* k_0 *any root of the algebraic equation* ${}_r f_0(\pm 1, k)$. *Let* $s_0 = 1/(1 - r + \mu_r(f_0))$. *If* (1) *is, at the point* $(0,0)$, *a regular deformation of the simplified equation* (7) *for the growth type* x^r, *the microscope* (14) $x = \varepsilon^{s_0}X$, $y = \varepsilon^{rs_0}Y$ *is regularizing for* (1), *and the image of equation* (1) *by this microscope* (14) *is infinitely close to the standard polynomial equation*

$$dY/dX = P(X,Y), \tag{16}$$

with ${}^rP(X,Y) = {}_r f_0$. *If* $({}_r f_0)'_Y(\pm 1, k_0) \neq 0$, *this equation has a river of type* k_0X^r *at* $X = \pm\infty$.

Proof. The previous theorem implies that ${}^rP = {}_r f_0$, and the existence of rivers of type (k,r) for $dY/dX = P(X,Y)$ depends only on that polynomial ${}^rP$: so the corollary follows immediately from Theorem 2.2. □

We shall come back to the problem of singular deformations at the end of the study of preresonant solutions.

3. Preresonant trajectories.

3.1. Entrance in the halo of a critical point. The two previous sections were dedicated to the geometric study of (1), that is, in some sense, to the formal solutions of that equation. We can now come to the study of the behaviour of the (slow) *solutions* of that equation, when x becomes infinitely close to a critical point. Actually, to know that the solution is slow for x not infinitely close to the critical point does not always give strong information about its behaviour when x becomes infinitely close to that point. Indeed, consider, for example, the equation $\varepsilon dy/dx = -2xy$, which has a critical point at $(0,0)$, and whose solutions are given, as a function of the initial condition $y_- = \overline{y}(x_-)$, by

$$\overline{y}(x) = y_- e^{x_-^2/\varepsilon} e^{-x^2/\varepsilon}.$$

The hypothesis that the solution is equal to an infinitesimal $y_- = \overline{y}(x_-)$ at some "initial condition" $x_- \not\simeq 0$ does not suffice to give some control on $\overline{y}$ for $x \simeq 0$: for example, for $x_- = -1$ and $x = 0$, y_- may be infinitesimal and $\overline{y}(0) = y_- e^{1/\varepsilon}$ may take, according to the value of y_-, any infinitesimal value, and $\overline{y}(0)$ may even be appreciable or illimited, for some convenient choice of $y_- \simeq 0$. This comes, essentially, from the fact that the slow curve $y = 0$ is repelling for $x < 0$ (or attracting for $x > 0$). In the general case, we can get a good precision on the behaviour of slow solutions in the halo of the critical point (x_0, y_0) and more precisely under a regularizing microscope, only for solutions that follow an attracting curve for $x < x_0$, or a repelling one for $x > x_0$. Such a solution, defined and satisfying that condition on *both* sides of x_0 is a *canard* [10], also called [18] "resonant in the sense of N. Kopell." This is why we shall call any solution satisfying that condition on *one* side (at least) of the critical point *pre*resonant.

DEFINITION. Let $\varphi : \mathcal{D}(\varphi) \longrightarrow \mathbb{R}$ be a (k,r)-branch at the point (x_0, y_0) of the slow-curve $f_0(x,y) = 0$ of (1). We say that φ is a *preresonant branch* at (x_0, y_0) if and only if φ is standard, $(f_0)'_y(x_0, y_0) = 0$, and $(f_0)'_y(x, \varphi(x))$ is nonzero and of the same sign as $(x - x_0)$ for all $x \in \mathcal{D}(\varphi)$ not equal to x_0 (i.e., if and only if (x_0, y_0) is critical, and for $x \neq x_0$, $(x, \varphi(x))$ is attracting if $x < x_0$, or repelling if $x > x_0$).

Let φ be a (k,r)-branch at (x_0, y_0) of the slow curve of (1) and $\overline{y}$ a solution of (1). We say that $\overline{y}$ is a *preresonant solution at* (x_0, y_0) *attached to* φ if and only if φ is preresonant, and for all $x \not\simeq x_0$ in $\mathcal{D}(\varphi)$, $\overline{y}(x)$ is defined and $\overline{y}(x) \simeq \varphi(x)$.

Examples. For a Union-Jack equation (3) (with $a < b < c$) (Fig. 3), the slow-curve at $(0,0)$ has one preresonant branch defined for $x < 0$ which is tangent to $y = bx$, and two preresonant branches defined for $x > 0$, tangent to $y = ax$ and $y = cx$, respectively.

At a fold point (Fig. 4), one of the branches of the slow curve is preresonant and the other is not.

To each preresonant branch there corresponds, on the regularizing blowup, an isolated river that is a shadow of the image by the microscope of any preresonant solution infinitely close to that preresonant branch. Theorem 3.1 shows that this is a general fact.

We shall consider in this main theorem the case of a preresonant branch defined for $x < x_0$, with $x_0 = 0$; one has of course an analogous result for any standard x_0, and also for any preresonant branch defined for $x > x_0$.

THEOREM 3.1 (matching principle). *Let* $k \neq 0$, $r > 0$, $x_- < 0$ *be standard, and* $\varphi : [x_-, 0] \longrightarrow \mathbb{R}$ *be a preresonant* (k,r)*-branch at* $(0,0)$ *of the slow curve of* (1); *Let*

$s_0 := 1/(1 - r + \mu_r(f_0))$. *Assume that* $({}_r f_0)'_y(-1, k) \neq 0$, *and that* (1) *is a regular deformation for the growth type of* φ *of the simplified equation*

$$\varepsilon \frac{dy}{dx} = f_0(x, y).$$

Let $P(X, Y)$ *be the polynomial such that the shadow of the blowup of* (1) *by the regularizing microscope*

$$x = \varepsilon^{s_0} X, \qquad y = \varepsilon^{r s_0} Y \tag{17}$$

is

$$\frac{dY}{dX} = P(X, Y). \tag{18}$$

Let $\overline{y}$ *be any maximal solution of* (1) *such that* $\overline{y}(x_-) \simeq \varphi(x_-)$. *Then*

(1) *There exists some limited* $X_+ \leq 0$ *such that* $\overline{y}$ *is defined and preresonant, attached to* φ *on* $[x_-, x_+]$, *with* $x_+ := \varepsilon^{s_0} X_+ \simeq 0$.

(2) *For all* $x \in [x_-, x_+]$ *such that* $x \simeq 0$, *if* x/ε^{s_0} *is unlimited, then* $\overline{y}(x) = k|x|^r(1 + \phi)$.

(3) *Equation* (18) *has a unique river* $\hat{Y} : (-\infty, a) \longrightarrow \mathbb{R}$ *of type* (k, r) *at* $X = -\infty$; *for any* X *near-standard in* $(-\infty, a)$ *the image* $\overline{Y}(X)$ *of* $\overline{y}(x)$ *by the microscope* (17) *is infinitely close to* $\hat{Y}(X)$. *So for* $x := \varepsilon^{s_0} X$, $\overline{y}(x)$ *is defined and satisfies*

$$\varepsilon^{-r s_0} \overline{y}(\varepsilon^{s_0} X) \simeq \hat{Y}(X).$$

Proof. The proof uses a zoom technique introduced by Callot for the Riccati–Hermite equation [5], and presented in a more general case by Benoit in [2]. It consists of a three-step study, the middle step connecting the scales of the first and the last step using a typically nonstandard technique.

Behaviour of $\overline{y}$ *for* $x \ll 0$ *and for* $x \leq x_0 \simeq 0$. As the branch $y = \varphi(x)$ is negative by assumption and thus attracting, as it is preresonant, the solution $\overline{y}$ is defined and satisfies $\overline{y}(x) \simeq \varphi(x)$ for all x such that $x_- \leq x \ll 0$ and thus, by permanence (Fehrele's principle*), there exists $x_0 \simeq 0$, $x_0 \leq 0$, such that this stays true for all $x \in [x_-, x_0]$.

As φ is asymptotic (tangent) to $k|x|^r$ at $x = 0$ with $k \neq 0$, there exists some standard $x'_- \in [x_-, 0)$ such that $\varphi(x) \neq 0$ for all $x'_- \in [x_-, 0)$. As the result is local, without loss of generality, we may assume that $x_- = x'_-$.

As φ is standard, $\varphi(x)$ is appreciable on $[x_-, 0)$ for all $x \not\simeq 0$, and thus $\overline{y}(x) = \varphi(x) + \phi = \varphi(x)(1 + \phi)$ if $x \not\simeq 0$. By Fehrele's principle, there exists some infinitesimal $x'_0 \leq x_0$ such that this stays true for all $x \leq x'_0$. Possibly choosing a smaller infinitesimal x'_0, we may also assume that x'_0/ε^{s_0} is unlimited (negative). Without loss of generality we now change x_0 into $x_0 := x'_0$.

Behaviour for $x \simeq 0$ *outside the* ε^{s_0}*-galaxy of* 0. We now assume that $x_0 \leq x < 0$, $|x|$ large enough for x/ε^{s_0} to be unlimited. We both have to show that $\overline{y}(x)$ is defined and that $\overline{y}(x) = k|x|^r(1 + \phi)$.

Let $\kappa(x) := \overline{y}(x)/|x|^r$, and choose $k_- < k_+$ standard with the same sign such that $k_- < k < k_+$, and such that $[k_-, k_+]$ contains no other root than k of the algebraic equation ${}_r f_0(-1, K) = 0$. We shall show that it is absurd to assume that $\kappa(x)$ leaves (k_-, k_+) on the external domain under consideration here. This will imply, on one hand, that $\overline{y}(x)$ stays defined, as the compact $\{(x, y) | x_0 \leq x \leq 0, k_-|x|^r \leq \overline{y}(x) \leq$

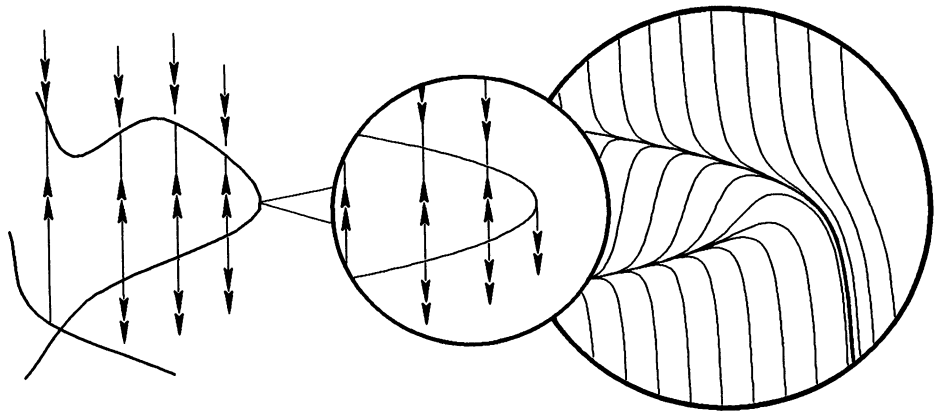


FIG. 5. *The equation* (1), *its image by the microscope* (19), *and its image by the regularizing microscope* (21).

$k_+|x|^r\}(\subseteq \text{hal}\,(0,0))$ is contained in the domain of (1). On the other hand, this implies also that $\kappa(x) \simeq k$ as k_- and k_+ are standard and can be chosen arbitrarily close to k, and so assertion (3.1) will be shown.

So assume that $\kappa(x)$ leaves (k_-, k_+) at $x_1 \geq x_0$, with x_1/ε^{s_0} unlimited, that is, $k_-|x|^r < \overline{y}(x) < k_+|x|^r$ for all $x \in [x_0, x_1)$ and $\overline{y}(x_1) = k_\pm|x|^r$. Let $s_1(< s_0)$ be such that $x_1/\varepsilon^{s_1} = -1$, and consider the microscope

$$x = \varepsilon^{s_1}\xi, \qquad y = \varepsilon^{rs_1}\eta. \tag{19}$$

Let $\overline{\eta}$ be the image of $\overline{y}$ by this microscope; $\overline{\eta}(\xi) = \varepsilon^{rs_1}\overline{y}(\xi/\varepsilon^{s_1})$ and thus $k_-|\xi|^r < \overline{\eta}(\xi) < k_+|\xi|^r$ for all $\xi \leq -1$ limited, and $\overline{\eta}(-1) = k_\pm$.

This is absurd; indeed, the image of (1) by the microscope (19) is a slow-fast differential equation

$$\varepsilon^d \frac{d\eta}{d\xi} = g(\xi, \eta) \tag{20}$$

with $d := 1 - s_1(1 - r + \mu_r(f_0)) > 0$ and $g(\xi, \eta) := \varepsilon^{-s_1\mu_r(f_0)} f(\varepsilon^{s_1}\xi, \varepsilon^{rs_1}\eta) \simeq {}_r f_0(\xi, \eta)$, the slow curve of which, for $\xi < 0$, is the union of the branches $y = K|x|^r$ for the various roots K of ${}_r f_0(-1, K) = 0$. The differential equation is thus fast at the point $(-1, \overline{\eta}(-1)) = (-1, k_\pm)$, and oriented towards the branch $k|x|^r$, as by asumption $({}_r f_0)'_y(-1, k) < 0$, which contradicts that $(\xi, \overline{\eta}(\xi))$ is contained in the crescent $\eta \in [k_-\xi^r, k_+\xi^r]$ for $\xi \leq -1$.

Behaviour for x in the ε^{s_0}-galaxy of 0. We just showed that $\overline{y}(x)$ is defined and satifies the inequalities $k_-|x|^r < \overline{y}(x) < k_+|x|^r$ for all $x \in [x_0, 0]$ such that x/ε^{s_0} is unlimited. Thus by Cauchy's permanence principle* there exists some $x_+ < 0$ such that $X_+ := x_+/\varepsilon^{s_0}$ is limited, and such that this internal property stays true for all $x \in [x_0, x_+]$, which implies, in particular, assertion (3.1).

In other words, the image $\overline{Y}$ of the solution $\overline{y}$ by the microscope

$$x = \varepsilon^{s_0}X, \qquad y = \varepsilon^{rs_0}Y \tag{21}$$

keeps contained, for $X \leq X_+$ limited, in the region $k_-|X|^r \leq \overline{Y}(X) \leq k_+|X|^r$ (Fig. 6) as, for $X \leq X_+$ limited, $x := \varepsilon^{s_0}X \geq x_0$; $\overline{Y}(X)$ is thus limited for all limited

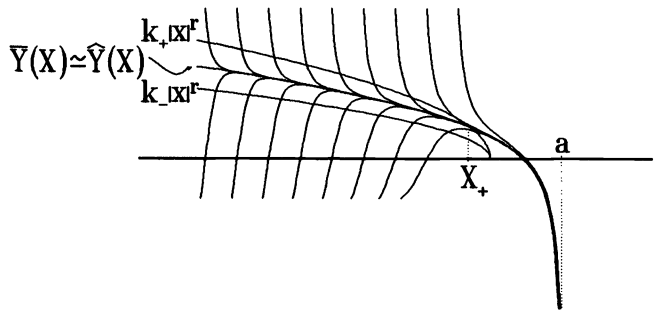


FIG. 6. *The unique river of* (22) *which is contained in the crescent-shaped region* $X < X_+$, $k_-|X|^r \le Y \le k_+|X|^r$ *for all (limited)* X *sufficently negative.*

$X \le X_+$. By the short-shadow lemma, $\overline{Y}$ has thus a shadow $\hat{Y}$, also contained in the close, standard region $k_-|X|^r \le \hat{Y}(x) \le k_+|X|^r$, which is necessarily a solution of the shadow (22) of (1) by the regularizing microscope (21):

$$\frac{dY}{dX} = P(X, Y). \tag{22}$$

But, by Corollary 2.3, one has ${}^rP(X, Y) = {}_rf_0$. As equation ${}^rP(-1, K)$ $(= {}_rf_0(-1, K)) = 0$ has a unique root, k, between k_- and k_+, by Proposition 1.4, $\hat{Y}$ is asymptotic to $k|X|^r$ at $X = -\infty$. As, moreover, $({}^rP)'_Y(-1, k) < 0$, by the rivers-existence theorem [11], $\hat{Y}$ is necessarily the unique river of (22), that is, of type (k, r). As $\hat{Y}$ is the shadow of $\overline{Y}$, one has $\overline{Y}(X) \simeq \hat{Y}(X)$ for all limited X. As k_- and k_+ are standard and of the same sign, $\hat{Y}(X)$ is appreciable for all limited $X \le X_+$, and thus $\overline{Y}(X) = \hat{Y}(X)(1 + \phi)$ for all limited X; hence assertion (3.1). □

Here is, as an example of application, a corollary giving the existence of canards for certain one-parameter famillies of Union-Jack equations. Such canards are not of class S^1 (i.e., the shadow of their image exhibits angles).

COROLLARY 3.2. *Consider a continuous, one-parameter* $d \in \mathbb{R}$ *family of Union-Jack equations*

$$\varepsilon \frac{dy}{dx} = (y - ax)(y - bx)(y - cx) + p(x, y, d) \tag{23}$$

with $a < b < c$ *fixed standard numbers,* p *an internal function with regular* ε*-shadow expansion,* $p(x, y) =: p_0(x, y) + \varepsilon(p_1(x, y, d) + \varepsilon\phi)$*, on some standard neighbourhood of* $\{(0, 0)\} \times [A, C]$*, with* $\mathcal{T}_3(p_0)(x, y) \equiv 0$*,* p_1 *standard continuous such that* $a = p_1(0, 0, A)$*, and* $c = p_1(0, 0, C)$.

Then there exist values $d_a \in (A, B)$ *(respectively,* $d_c \in (B, C)$*) of the parameter* d *for which* (23) *has a canard, that is, more precisely, a slow solution, defined on the halo of* 0 *(and further), following, for* $x < 0$*, the attracting branch of a slow-curve tangent to* $y = bx$*, and following, for* $x > 0$*, the repelling branch tangent to* $y = ax$ *(respectively,* $y = cx$*).*

Proof. As the Taylor polynomial of p_0 of degree 3, $\mathcal{T}_3(p_0)$ is zero, the slow curve of (23) has, at $(0, 0)$, three negative branches, φ_a^-, φ_b^-, φ_c^-, and three positive branches φ_a^+, φ_b^+, φ_c^+, tangent, respectively, to the three straight lines $y = ax$, $y = bx$, and $y = cx$ (Fig. 7(a)); the three preresonant branches are φ_b^-, φ_a^+, and φ_c^+. We shall show the existence of a value $d_c \in (B, C)$ and of a canard, for $d = d_c$, following φ_b^-

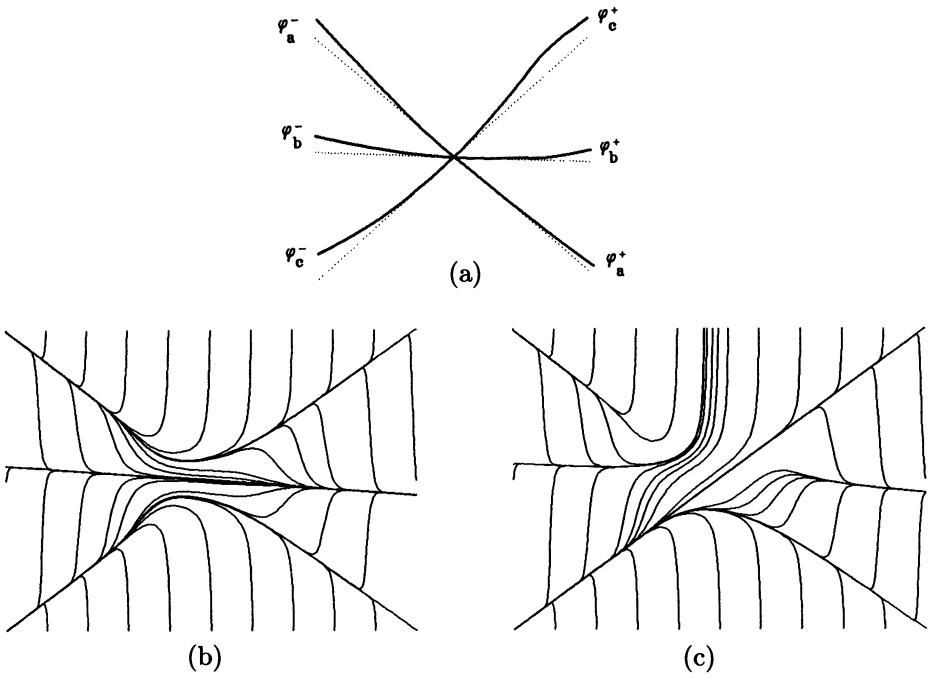


FIG. 7. (a) *the six branches, at* $(0,0)$ *of the (fixed) slow curve of the family* (23); (b) *and* (c) *are some trajectories of* (24), *image by the regularizing microscope of* (23) *for* $d = B$ *and* $d = C$.

and φ_c^+. One would proceed analogously for the existence of a canard following φ_a^+ for some $d = d_a \in (A, B)$.

Let $x_- \ll 0 \ll x_+$ be such that $x_- \in \mathcal{D}(\varphi_b^-)$ and $x_+ \in \mathcal{D}(\varphi_c^+)$. Choose $y_- \simeq \varphi_b^-(x_-)$ and $y_+ \simeq \varphi_c^+(x_+)$; so, (x_-, y_-) and (x_+, y_+) are two "initial conditions" belonging, respectively, to the halo of the two considered branches of slow curve: let $\overline{y}_-$ and $\overline{y}_+$ be the maximal solutions of (23) passing through these two initial conditions; as in (23), $\overline{y}_-$ and $\overline{y}_+$ are dependent continuously on the parameter d.

LEMMA 3.3. *Let* $dY/dX = F(X, Y, d)$ *be a continuous one-parameter* $d \in [D_-, D_+]$ *family of locally Lipschitz differential equations defined for all* $(X, Y) \in [X_-, X_+] \times \mathbb{R}$. *For each* $d \in [D_-, D_+]$, *let* $(X, d) \mapsto \hat{Y}_-$ *and* $(X, d) \mapsto \hat{Y}_+$ *be the maximal solution through* Y_- *and* Y_+ *for* $X = X_-$ *and* $X = X_+$, *respectively* ($Y_\pm$ *may be constant or may depend continuously on the parameter* d). *If there exist* X_0^- *and* X_0^+ *in* $[X_-, X_+]$ *such that* $Y_-(X_0^-, D_-) < Y_+(X_0^-, D_-)$ *and* $Y_-(X_0^+, D_+) > Y_+(X_0^+, D_+)$, *then there exists* $d_c \in [D_-, D_+]$ *such that* $\hat{Y}_-(X, d_c) = Y_+(X, d_c)$ *for all* $X \in [X_-, X_+]$.

Proof. Consider the map $\varphi : [D_-, D_+] \longrightarrow \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ that to each value of the parameter d associates the value of $\hat{Y}_-(X_+, d)$ if this value is defined, or $\pm\infty$ if $\hat{Y}_-$ stops to be defined because it tends to $\pm\infty$ for some $X_0 < X_+$. By continuity of F and Y_- with respect to d, φ is continuous. By uniqueness $\varphi(D_-) < Y_+$ and $\varphi(D_+) > Y_+$. Thus, by continuity with respect to d of φ and Y_+, there exists d_c such that $\varphi(d_c) = Y_+$. By uniqueness, for $d = d_c$, $\hat{Y}_- = \hat{Y}_+$. □

Here is the principle of the end of the proof.

Looking at the solutions $\overline{y}_-$ and $\overline{y}_+$ on the regularizing blowup, where, by Theorem 3.1, these solutions are infinitely close to the isolated rivers $\overline{Y}_-$ and $\overline{Y}_+$, shows

that for $d = B$ and $d = C$, the solutions $\overline{y}_-$ and $\overline{y}_+$ are defined and slow until $x = 0$, that is, on $[x_-, 0]$ and $[0, x_+]$, respectively, which for $d = B$, one has $\overline{y}_-(0) < \overline{y}_+(0)$, and that for $d = C$, these values at 0 of the solutions are in the inverse order. Thus by Lemma 3.3 there exists some $d_c \in (B, C)$ such that for this value of d, $\overline{y}_- \equiv \overline{y}_+$: this solution is the desired canard. Let us check the various facts stated in this reasoning.

Consider the following regularizing microscope

$$\varepsilon^{1/3} X = x, \qquad \varepsilon^{1/3} Y = y$$

that we already used in example (11). The shadow of the corresponding blowup of (23) is

$$\frac{dY}{dX} = (Y - aX)(Y - bX)(Y - cX) + D, \quad \text{with } D = p_1(0, 0, d). \tag{24}$$

Let $\overline{Y}_\pm(X) := \varepsilon^{-1/3} \overline{y}_\pm(\varepsilon^{1/3} X)$ be the images of the solutions $\overline{y}_\pm(x)$. Theorem 3.1 implies that $\overline{Y}_-(X)$ is defined and infinitely close to the unique river $\hat{Y}_-$ asymptotic to $Y = bX$, if X is near-standard in the domain of $\hat{Y}_-$. Analogously $\overline{Y}_+(X)$ is defined and infinitely close to the unique river $\hat{Y}_+$ asymptotic to $Y = cX$ if X is near-standard in the domain of $\hat{Y}_+$.

An elementary study of the two solutions $\hat{Y}_\pm$ of (24) shows that for $D = b$ and $D = c$, these solutions are well defined on an open interval containing $(-\infty, 0]$ and $[0, +\infty)$, respectively.

For $D = b$, $\hat{Y}_-(X) \equiv bX$ is an obvious solution, and $\hat{Y}_+(X) > bX \equiv \hat{Y}_-(X)$, for X positive and large enough, thus for all $X \in \mathcal{D}\hat{Y}_+$, by uniqueness of solutions (Fig. 7(b)). Moreover, as for large values of $Y > 0$, $dY/dX > 0$, solutions of (24) "above" the obvious solution Y_- have to be defined down to $X = -\infty$. So, in particular, $\hat{Y}_+(0)$ is defined and $\hat{Y}_+(0) > \hat{Y}_-(0)$. Now, still for $D = b$, that is, $d = B$, using that $\hat{Y}_+(0)$ and $\hat{Y}_+(0)$ are standard, one has

$$\varepsilon^{-1/3} \overline{y}_-(0) = \overline{Y}_-(0) \simeq \hat{Y}_-(0) \ll \hat{Y}_+(0) \simeq \overline{Y}_+(0) = \varepsilon^{-1/3} \overline{y}_+(0),$$

and thus $\overline{y}_-(0) < \overline{y}_+(0)$.

One uses an analog way of reasoning for $d = C$ (Fig. 7(c)), that is, $D = c$, observing here that $\hat{Y}_+(X) \equiv cX$ is now an obvious solution of (24), and that $\hat{Y}_+ < \hat{Y}_-$. So, for $d = C$, $\overline{y}_-(0) > \overline{y}_+(0)$, which shows the corollary, as explained above. □

Remark. A consequence of Corollary 3.2 is that there exists, for (24) a value $D(= {}^o d)$ for which the unique river asymptotic to bX at $X = -\infty$, and the unique river asymptotic, for example, to aX at $X = +\infty$, are equal. Figure 8 shows a numerical computation of that value. We would be highly interested in an analytical method to determine this value.

3.2. Exiting a critical point's halo. In Theorem 3.1 we considered the question of the image, on a convenient regularizing blowup of any preresonant slow solution. The example considered in Corollary 3.2 can also be used to illustrate the inverse problem, which will be solved below with Theorem 3.4: to the contrary of the canard behaviour of a solution of (23), the "ordinary" behaviour of a preresonant solution that enters, say, for $x < 0$ in the scope of the regularizing microscope, is to stay infinitely close, on the blowup, to a repelling river (at $X = -\infty$). This river solution may also be (see, for example, the left-hand side of Fig. 8) a river at $X = +\infty$ but *attracting* now, so belonging to a one-parameter family of rivers, all with same asymptotic expansion at $X = +\infty$. Such a slow solution crosses the halo of the critical

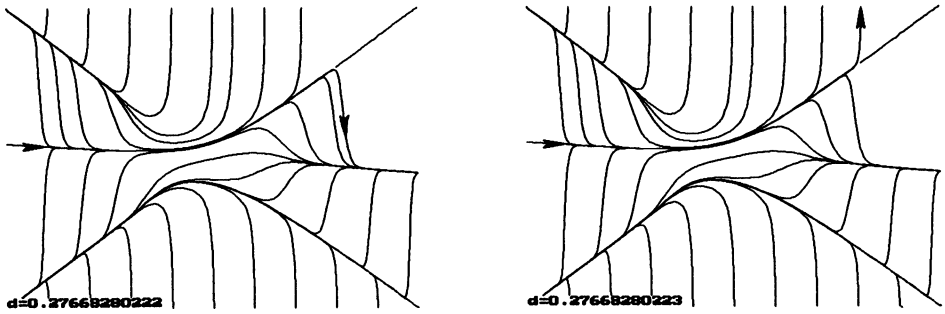


FIG. 8. *Numerical computation of the standard part of the canard value d_c for equation* (23), *in the case* $(a,b,c) = (-1,-0.1,1)$. *The behaviour at* $X = +\infty$ (*"descending" or "ascending" arrow*) *of the river at* $X = -\infty$ *of type* $(b,1)$ (*"left-entering arrow"*) *changes very quickly with the value of d, displayed in the lower left corner.*

point and then follows, for $x > 0$, an attracting branch $\varphi_+ : [0, x_+] \longrightarrow \mathbb{R}$, of type (k_+, r_+), where (k_+, r_+) is precisely the type of the considered river at $X = +\infty$.

Once more, we assume that $(x_0, y_0) = (0,0)$, but that φ is defined on $[0, x_+]$ this time, and is attracting (thus *nonpreresonant*). There is of course an analogous result for any standard (x_0, y_0), and also any repelling branch defined on some "left-interval" $[x_-, x_0]$.

THEOREM 3.4. *Let* $k \neq 0$, $r > 0$, $x_+ > 0$ *be standard numbers, and* $\varphi : [0, x_+] \longrightarrow \mathbb{R}$ *be any attracting* (k,r)*-branch at* $(0,0)$ *of the slow curve of equation* (1); *put* $s_0 := 1/(1-r+\mu_r(f_0))$. *Assume that* $({}_r f_0)'_y(-1,k) \neq 0$, *and that* (1) *is a regular deformation for the growth-type of the branch* φ *of the simplified equation*

$$\varepsilon \frac{dy}{dx} = f_0(x,y).$$

Denote by $P(X,Y)$ *the polynomial such that the shadow of the blowup of* (1) *by the regularizing microscope*

$$x = \varepsilon^{s_0} X, \qquad y = \varepsilon^{r s_0} Y \tag{25}$$

is equal to

$$\frac{dY}{dX} = P(X,Y). \tag{26}$$

Under these assumptions, (26) *has an attracting river* $\hat{Y} : (a, +\infty)$ *of type* (k,r) *at* $X = +\infty$.

Let $\overline{y}$ *be any maximal solution of* (1), *and let* $\overline{Y}(X) := \varepsilon^{-r s_0} \overline{y}(\varepsilon^{s_0} X)$ *be its image by the regularizing microscope* (25). *Let* $X_0 \gg a$ *be limited, and define* $x_0 = \varepsilon^{s_0} X_0$.

(1) *If* $\overline{Y}(X_0) \simeq \hat{Y}(X_0)$, *then for any* X *near-standard in* $]a, +\infty[$ *and for* $x := \varepsilon^{s_0} X$, $\overline{y}(x)$ *is defined and satisfies*

$$\varepsilon^{-r s_0} \overline{y}(\varepsilon^{s_0} X) \ \ (= \overline{Y}(X)) \simeq \hat{Y}(X).$$

(2) *For all* $x \in [0, x_+]$ *such that* $x \simeq 0$, *if* x/ε^{s_0} *is unlimited, then* $\overline{y}(x) = kx^r(1+\not{o})$.

(3) *For any appreciable* $x \in]0, x_+]$, $\overline{y}(x) \simeq \varphi(x)$.

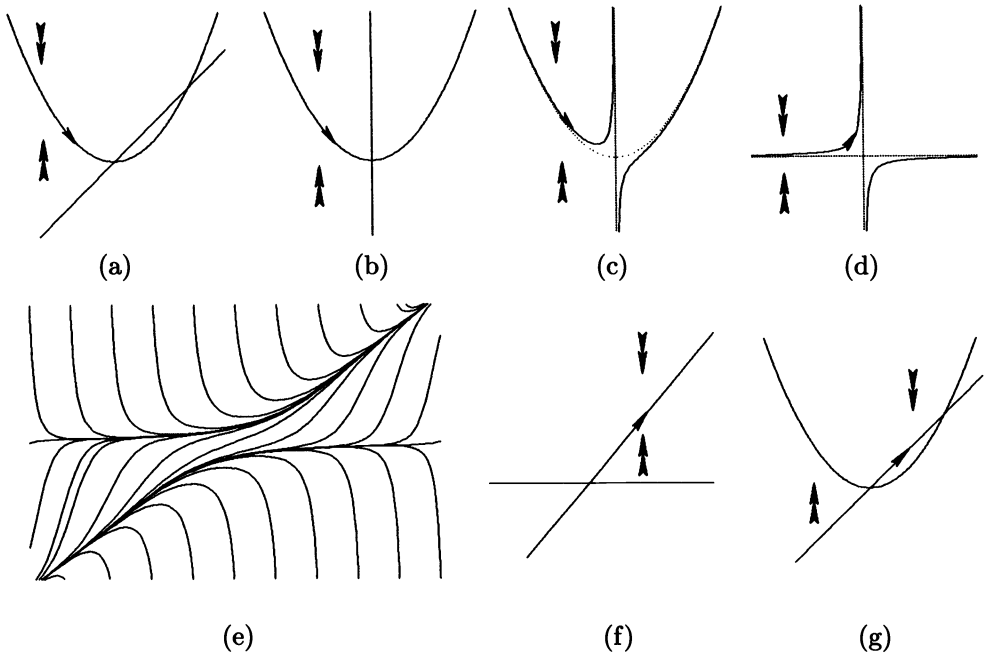


FIG. 9. *The passage through the halo of the critical point of* (27). (a) *and* (g): $\varepsilon^s = 1$, *the initial scale*; (b): $\varepsilon^s \simeq 0$ *and* $\varepsilon/\varepsilon^{3s} \simeq 0$; (c): $\varepsilon^{3s} = \varepsilon$, *one puts* $\eta = \varepsilon^{1/3}$; (c)–(e): *the microscopes of Theorem* 3.1 *in the case* $0 > r \ (= -1)$; (e)–(g): *the microscopes of Theorem* 3.4 *applied to the positive branch* $y = x$.

Proof. By Corollary 2.3, we have ${}^rP(X,Y) = {}_rf_0$. By Proposition 1.2, we have ${}^rP(1,k) = {}_rf_0(1,k) = 0$, and, as the branch φ is attracting, $({}_rf_0)'_y(1,k) < 0$. Thus by the rivers existence theorem, there exists an attracting river $\hat{Y}$: $(a,+\infty) \longrightarrow \mathbb{R}$, of type (k,r). The short-shadow lemma implies that the solution $\overline{Y}$ of blowup of (1) by the regularizing microscope (25) is defined for all X near-standard in $]a,+\infty[$, and that $\overline{Y}(X) \simeq \hat{Y}(X)$; hence assertion (3.4).

Assertions (2) and (3) are obtained by reasoning with the same microscopes as in the proof of Theorem 3.1; the existence of the slow solution up to x_+ follows from the attractivity of the slow curve at each of these scales. □

3.3. Singular deformations. We shall raise the question of equations with singular deformations for some branches of the slow curve on an example, and more precisely on the example of (11) which we already used to introduce that notion. This will also give us the oportunity to sketch how to deal with slow curves at some x_0 with branches of type (k,r), with $r < 0$. For the sake of simplicity, we do not consider here this question in the general case (that is, only relevant $f_0(x,y)$ that are polynomials in y^α).

Recall that (11) is equation

$$\varepsilon\frac{dy}{dx} = (y - x^2)(x - y) + \varepsilon(a + bx) \tag{27}$$

for which the term $\varepsilon(a + bx)$ is a singular deformation for the branch $y = x^2$ for any standard $a \neq 0$. This branch being preresonant for $x < 0$, we shall, more precisely, be interested in the behaviour of any preresonant solution $\overline{y}$ with initial condition

(x_-, y_-) such that x_- is appreciable and negative and $y_- \simeq x_-^2$. The branch $y = x^2$ being attracting, $\overline{y}(x)$ is defined and $\overline{y}(x) \simeq x^2$ for all x such that $x_- \leq x \ll 0$, and thus, by Fehrele's principle, for some $x \simeq 0$. As soon as $x \simeq 0$, we study $\overline{y}$ using the "zoom" (microscope with strengh varying with s)

$$\varepsilon^s X = x, \qquad \varepsilon^{2s} Y = y. \tag{28}$$

Let $\overline{Y}_s$ be the image of $\overline{y}$ by this microscope; thus it is a solution of equation

$$\frac{dY}{dX} = \varepsilon^{-1-s}[\varepsilon^{3s}(Y - X^2)(X - \varepsilon^s Y) + \varepsilon(a + bX)]. \tag{29}$$

As long as $s \ll \frac{1}{3}$, and more precisely as long as $\varepsilon/\varepsilon^{3s} \simeq 0$, (29) is slow-fast, with slow curve $X(Y - X^2)$, which is attracting for $X < 0$ (Fig. 9). Using the same reasoning as in the proof of Theorem 3.1, we see that for all these values of s and all appreciable $X < 0$, $\overline{Y}_s(X) \simeq X^2$, and thus $\varepsilon^{-2s}\overline{y}(\varepsilon^s X) \simeq X^2$.

Increasing the strengh of the zoom, that is, increasing s such that the ratio $\varepsilon^{3s}/\varepsilon$ becomes appreciable, say equal to 1, that is $s = \frac{1}{3}$, then $\overline{Y}_{1/3}$ is a solution of

$$\eta \frac{dY}{dX} = X(Y - X^2) + a + \eta(bX - Y(Y - X^2)), \tag{30}$$

with $\eta := \varepsilon^{1/3}$; this is still a slow-fast equation, but a change occured in the slow curve (Fig. 9(c)): it's no longer a parabola, but the curve $X(Y - X^2) + a = 0$, which, as $a \neq 0$, is now only asymptotic to the parabola $\{Y = X^2\}$ at $X = -\infty$; but at $X = 0$, it is asymptotic to $Y = -a/X$. *This part of the slow curve has to be seen as a (negative) branch φ of type $(a, -1)$ at $(0, 0)$, which means asymptotic to $a|X|^r$, with* $r = -1$.

Notice that now the term ηbX is a *regular* deformation term with respect to the preresonant branch φ, of type $(a, -1)$, for the simplified equation $\eta(dY/dX) = X(Y - X^2) + a$, for which the "microscope" (it is indeed a microscope with respect to X, but it is a *macroscope* with respect to Y)

$$\eta^{\frac{1}{2}} \mathbf{X} = X, \qquad \eta^{-\frac{1}{2}} \mathbf{Y} = Y \tag{31}$$

changes (30) into a regular perturbation for the polynomial differential equation with rivers

$$\frac{d\mathbf{Y}}{d\mathbf{X}} = -\mathbf{Y}^2 + \mathbf{X}\mathbf{Y} + a, \tag{32}$$

which exhibits in particular a river $\hat{\mathbf{Y}}_-$ of type $-a/\mathbf{X}$ at $\mathbf{X} = -\infty$. Observe now that in Theorem 3.1, we assumed that r, the growth type of the branch, is positive; this was done for the sake of simplicity, and also for the sake a generality with respect to the equation (which was assumed to be C^∞ and not only a polynomial), a negative growth type at x_0 making no sense in the C^∞ case. One checks easily that we only used that $r \neq 0$ (and thus $kr \neq 0$). So we come to the same conclusion here and see that for limited $\mathbf{X}$ sufficiently negative,

$$\hat{\mathbf{Y}}_-(\mathbf{X}) \simeq \overline{\mathbf{Y}}_{\frac{1}{3}}(\mathbf{X}),$$

where $\overline{\mathbf{Y}}_{1/3}$ denotes the image of the solution $\overline{Y}_{1/3}$ on the blowup using the microscope (31).

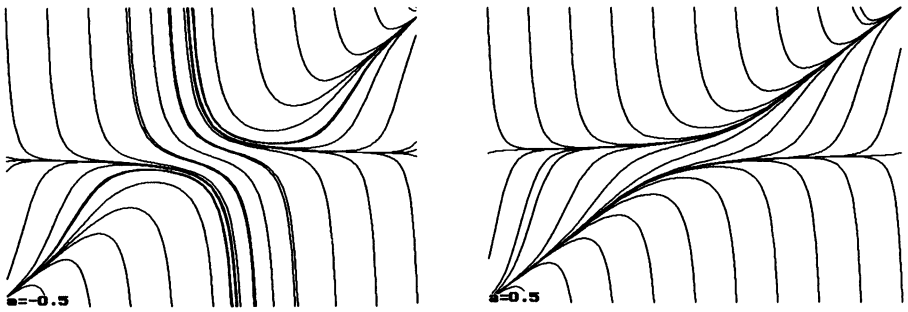


FIG. 10. *The blowup of equation* (27) *by the regularizing microscope* (31) *for some negative and some positive value of the parameter* a.

It is easy to study the behaviour of the rivers of (32) (see Fig. 10). One shows in particular that for $a > 0$, the river $\hat{\mathbf{Y}}_-$ at $\mathbf{X} = -\infty$ is also a river at $\mathbf{X} = +\infty$, but attracting and of type $(1,1)$ this time.

The time has come to express the microscope we obtained by the two successive microscopes we applied, namely, $x = \varepsilon^{1/3}X$, $y = \varepsilon^{2/3}Y$, and $X = \varepsilon^{2/3}\mathbf{X}$, $Y = \varepsilon^{-1/3}\mathbf{Y}$, as $\eta = \varepsilon^{1/3}$. Finally, we have

$$x = \varepsilon^{\frac{1}{2}}\mathbf{X}, \qquad y = \varepsilon^{\frac{1}{2}}\mathbf{Y} :$$

we merely resulted in the regularizing microscope relative to the branch of type $(1,1)$ (with respect to which the complete equation (27) is a regular deformation of the simplified equation). So we can apply Theorem 3.4 for $\mathbf{X} > 0$, and we get (Fig. 9(e)–9(g)) that for $a > 0$ (standard), the solution $\overline{y}$ stays slow for $x > 0$, and follows the attracting curve $y = x$.

Let us observe that for $a < 0$ (standard), Fig. 10(b) suggests what the behaviour of $\overline{y}$ will be for $x \gg 0$: when $\mathbf{X}$ increases, $\hat{\mathbf{Y}}_-$ decreases faster than any linear function and thus, coming back to the initial scale, $\overline{y}$ jumps towards y unlimited negative.So we can deduce that $\overline{y}$ will assume, for some $a \simeq 0$, intermediate behaviours, and especially that for any limited b, there exists some $a(b)$ such that $\overline{y}_{a(b),b}$ is a canard of (27).

This example shows how to reduce a singular deformation problem: r being chosen with respect to the preresonant branch φ followed by the considered solution, one increases gradually the value of s in the zoom (28): for small values of $s > 0$, at that scale the solution follows the slow curve equal to the growth type of φ. For some smallest standard value $s_1 (< s_0$, unless the deformation would just be regular), one or several terms in the deformation get the same size order as those of f_0: Proposition 1.2 shows that such terms can only rise from those f_p such that $1 \leq p < 1 + s_0(r-1)$. One gets a new slow-fast equation ("ε" becomes "η"), the standard part of which being henceforth a polynomial, and the slow curve being an *algebraic curve*, which may exhibit new critical points and which may once more be analysed in branches at these critical points, but also branches at $X = \pm\infty$, or $Y = \pm\infty$ $(r < 0)$. Once the critical points with finite abscissa have been found (this is the only difficult operation), one can study theses branches with convenient Newton polygons. If the slow curve at the initial scale exhibits a preresonant branch of type (k,r) for $\pm(x - x_0) > 0$, the slow curve of the blowup will exhibit a branch of type (k,r) at $X = \pm\infty$ such that the solution will follow up to some new critical point of finite abscissa and finite or infinite ordinate: it suffices to reapply the same process; as already mentioned, the

Theorems 3.1 and 3.4 apply in both cases, and this makes it possible to "follow" the considered solution, at least when no canard situations occur, which would imply an additional canard study. We do not doubt that this process will finally result (if no nonzero occuring function is infinitely flat at some of the critical points, which will not happen if the functions are assumed to be analytic) in a regularizing blowup, but it would of course be interesting to solve this question in an effective manner (symbolic machine computations). We will not consider here this purely algebraic question that involves only the equation, our goal being to show the role played by rivers in the behaviour of the slow *solutions*.

4. Appendix. Let us briefly sketch the nonstandard tools involved in this paper.

The vocabulary of infinitesimals had gradually been forsaken by most mathematicians because of the contradictions that this terminology seemed to introduce. To overcome these contradictions, logicians like Löwenheim and Skolem introduced so-called "nonstandard models," mathematical constructions that would, for example, add "infinitely large elements" to the integers. Abraham Robinson extended these techniques in order to build efficient nonstandard models that contain all of what is needed by mathematical analysis (Banach spaces, measure spaces, etc.) He also pointed out that his construction should indicate how to use safely the infinitesimal terminology within the usual mathematical sets, like $\mathbb{R}$, the main idea being not to apply blindly classical results to "external sets" like the set of standard integers. It is Nelson that achieved this program, in his October 1977 paper in the Bulletin of the American Mathematics Society that indicates what to do (and what *not* to do) using that terminology. Actually, Nelson's Internal Set Theory introduces just one new word, *standard*, and gives the rules for correct use of it.

We use Nelson's approach (and thus the usual real line, not some elaborate model); here is how it works. Every classical object is standard or not: we just need tools to determine which is which. Any uniquely defined object (possibly using other objects that have already been shown to be standard) is standard. So $\emptyset$, 0, 1, π, sin, $\mathbb{R}$, $\mathcal{C}_0(\mathbb{R})$ are standard. An equivalent statement, called *transfer*, asserts that classical-type theorems are true if and only if they are true for standard elements. Things become more interesting with the theorem asserting that any infinite set contains nonstandard elements (infinite means, as usual, in one–one correspondence with one of its proper subsets). So $\mathbb{N}$ has nonstandard elements, and any nonstandard integer, say ω, is larger than any standard one. So we have an infinitely large number ω. Taking its inverse (in $\mathbb{Q}$ or $\mathbb{R}$), we get a nonzero *infinitesimal* $\varepsilon = 1/\omega$, i.e., $|\varepsilon|$ is smaller than $1/n$ for all standard $n > 0$. A number is *limited* if its absolute value is smaller than some standard number (one should keep the word "finite" for questions of cardinality: the set $\{1, \ldots, \omega\} \subset \mathbb{N}$ is finite, but its cardinal is infinitely large, or better: *unlimited*). It is *appreciable* if it is neither infinitesimal nor unlimited. Here are two notations: ϕ (pronounce *zerobar*) will always denote an infinitesimal and $\pounds$ a limited number, but two occurences of ϕ are *not* necessarily equal, and analogously for $\pounds$. Please notice the difference between the empty-set symbol $\emptyset$ and this symbol ϕ. We write $x \simeq y$ if and only if $x - y$ is infinitesimal, and, to the contrary, $x \ll y$ if and only if $x - y$ is not infinitesimal.

As $\mathbb{R}$ is complete, any limited real number l is infinitely close to a standard one, called its standard part, and denoted by ${}^{\circ}l$. This extends easily to any standard finite-dimensional Banach space B. For any standard $C \subseteq B$, if ${}^{\circ}l \in C$, we say that l is near-standard in C. If $\gamma \subseteq B$ has all its limited points near-standard in C, it is *infinitely close to* C. If γ is a curve (think of a trajectory), we also say that γ

follows C. If, moreover, C is standard and any of its standard points is infinitely close to some point of γ, then C is called the *shadow* of γ, and we write ${}^o\gamma = C$ (actually, on any standard compact subset, it *is* a standard part for the Haussdorf metric). A standard function f_0 is the shadow of the function f if their graphs are also. We use this terminology for differential equations, just identifying the equation $y' = f(x,y)$ with the function f. Given a differential equation $y' = f(x,y)$ and its shadow $y' = {}^of(x,y)$, both assumed to be locally Lipschitz, the *short-shadow lemma* [9, Thm. 8.2.2] indicates the relationship existing between the shadow of some segments of solution of the initial equation and a standard solution of the shadow equation. The segment should be short in the sense that the difference between two elements in its domain has to be limited. The lemma also gives existence results for solutions of the initial equation from existence assumptions on the shadow equation.

To make it possible to assume that $\varepsilon > 0$ is a *fixed* infinitesimal is a key point in the use of nonstandard analysis to singular perturbation theory. For example, dealing with the question of *existence*, the fact that solutions have their values in a (finite-dimensional and thus) locally compact space, one can refer to the prolongation theorem, which ensures that the maximal solution leaves any closed subset of the domain of the equation; it might be more cumbersome to deal with an (ad hoc) infinite-dimensional space (of $\mathbb{R}^2$-valued functions, defined for $\varepsilon \in (0, \varepsilon_0)$). A still more important reason is that it makes it easier to let other parameters vary. Expansions involving ε may enter "by themselves" for a *number* a, as, for example, for the necessary value $a = 1 - \frac{1}{8}\varepsilon - \frac{3}{32}\varepsilon^2 + \varepsilon^2\phi$ of some parameter (for the existence of canards [3]).

A number a has an ε-shadow expansion, that we shall write $a = \sum a_n\varepsilon^n$, if $(a_n)_{n\geq 0}$ is a standard sequence of numbers and for any *standard* n, $a = a_0 + a_1\varepsilon + \cdots + a_n\varepsilon^n + \varepsilon^n\phi$ (this formal expansion usually does not converge for any nonzero value of ε). A function f has an ε-shadow expansion, write $f = \sum f_n\varepsilon^n$, if $(f_n)_{n\geq 0}$ is a standard sequence of functions on some domain D, and for any *standard* n, and any x near-standard in D, $f(x) = f_0(x) + f_1(x)\varepsilon + \cdots + f_n(x)\varepsilon^n + \varepsilon^n\phi$. The expansion is *regular* if the functions f_n are C^∞, and for any *standard* p, $f^{(p)} = \sum f_n^{(p)}\varepsilon^n$. For example, if F is standard and C^∞ with respect to (x,e) and if $\varepsilon > 0$ is some infinitesimal, then Taylor's theorem shows that for $f(x) := F(x,\varepsilon)$, one has $f = \sum F_{e^n}^{(n)}(.,\varepsilon)\varepsilon^n/n!$. But there exist useful examples which are not of that kind, for example, $f(x) = \Phi(x,\varepsilon,a)$, $\Phi \in C^\infty$ standard, and a is equal to the sum of the smallest term of some diverging expansion, as for the canard values of the Van der Pol equation [3]. If f has a regular ε-shadow expansion, and $\overline{y}$ is a slow solution of $\varepsilon y' = f(x,y)$ that follows a branch of the slow curve $\{{}^of(x,y) = 0\}$ with no critical point, then $\overline{y}$ has also an ε-shadow expansion, and this expansion is the same for all slow solutions following the same branch [8].

It is possible to deal with external sets, provided one takes care not to apply (blindly) classical theorems to them. We say that a set (i.e., a subcollection defined using the extended language of some classical set) is *external*, if some classical result is wrong for it. The set ${}^{st}\mathbb{N}$ of standard integers is external (it is bounded by ω but has no least upper-bound), and so is the *principal galaxy* $\mathbb{G}$ *of* $\mathbb{R}$, the set of all limited real numbers (the previous example could be deduced from it just by intersecting it with the standard set $\mathbb{N}$), or the α-galaxy $x_0 + \alpha\mathbb{G}$ of any x_0 for any $\alpha \neq 0$. One has specific results for external sets, which can be used as *permanence* principles (or overspill principles). The most obvious one is the "Cauchy principle" (named after Cauchy's original statement on continuity of the limit of a sequence of continuous functions)

that just states that an external set is not internal. So, an internal property cannot hold only on an external set (otherwise this statement would define the external set, and there would be a contradiction in classical mathematics) and the property must overspill to other points. An external set is a *halo* if it is the *intersection* on all standard indices of an internal family of sets (think of $\mathrm{hal}(0) := \bigcap_{st\mathbb{N}}[-\frac{1}{n}, \frac{1}{n}]$), and it is a *galaxy* if it is the *union* on all standard indices of such a family (think of $\mathbb{G} := \bigcup_{st\mathbb{N}}[-n, n]$). So the external domain where a function is limited or appreciable is a galaxy, and the external domain where it is infinitesimal or unlimited is a halo. Fehrele's principle states that no halo is a galaxy, so, for example, two functions cannot be infinitely close to each other only on a galaxy. This is a generalization of an easy but nice result called Robinson's lemma. The typical use of this permanence result is when it is necessary to overspill an "up to an infinitesimal" estimate from all limited values of the variable up to some infinitely large ones: the domain where this type of estimate holds is a halo, whereas the external set of limited numbers is a galaxy; so the halo must be strictly larger than this galaxy.

We refer to [9] for proofs, further results, and bibliography.

REFERENCES

[1] R. C. Ackerberg and R. E. O'Malley, *Boundary layer problems exhibiting resonance*, Stud. Appl. Math., 49 (1970), pp. 277–295.

[2] E. Benoit, *Loupes variables*, in Analyse Non Standard et Représentation du Réel, M. Diener and C. Lobry, eds., Editions CNRS/OPU, 1984, pp. 93–102.

[3] E. Benoit, J. L. Callot, F. Diener, and M. Diener, *Chasse au canard*, Collectanea Mathematica, Barcelone, 31 (1981), pp. 37–119.

[4] F. Blais, *Fleuves généralisés*, Ph.D. thesis, Université Paris 7, 1989.

[5] J. L. Callot, *Bifurcation du portrait de phase pour les équations différentielles linéaires du second ordre ayant pour type l'équation d'Hermite*, Ph.D. thesis, d'Etat 125/TE-13, IRMA, 7, rue R. Descartes, F67084 Strasbourg Cedex, 1981.

[6] B. Candelpergher, J. C. Nosmas, and F. Pham, *Approche de la résurgence*, Hermann, Paris, 1992.

[7] F. Diener, *Fleuves et variétés centrales*, in Singularités des équations différentielles, Dijon 1985, Astérisque 150-151, Société Mathématique de France, 1987, pp. 59–66.

[8] F. Diener and M. Diener, *Some asymptotic results in ordinary differential equation*, in Non Standard Analysis and its Applications, N. Cutland, ed., Cambridge University Press, London, 1988, pp. 282–297.

[9] F. Diener and G. Reeb, *Analyse Non Standard*, Hermann, Paris, 1989.

[10] M. Diener, *Canards et bifurcations*, in Outils et modèles mathématiques pour l'automatique, l'analyse des systèmes et le traitement du signal, tome 3, I. D. Landau, ed., Editions du CNRS, 1983, pp. 315–328.

[11] M. Diener and G. Reeb, *Champs polynômiaux: nouvelles trajectoires remarquables*, Bull. Soc. Math. Belgique, 38 (1987), pp. 131–150.

[12] J. Dieudonné, *Calcul Infinitésimal*, Hermann, Paris, 1980.

[13] J. Ecalle, *Les fonctions résurgentes,* tomes 1,2,3, prépublication, Publications mathématiques d'Orsay, Université de Paris Sud, Département de Mathématiques, bat. 425, F91425 Orsay, 1981–1985.

[14] N. Kopell, *A geometric approach to boundary layer problems exhibiting resonance*, SIAM J. Appl. Math., 37 (1979), pp. 436–458.

[15] S. Lefschetz, *Algebraic Geometry*, Princeton University Press, Princeton, NJ, 1953.

[16] E. F. Mishchenko and N. K. Rosov, *Differential Equations with Small Parameters and Relaxation Oscillations*, Plenum Press, New York, 1980.

[17] J. P. Ramis, *Les séries k-sommables et leurs applications*, Springer Lecture Notes in Physics, 126, 1980.

[18] Y. Sibuya, *A theorem concerning uniform simplification at a transition point and the problem of resonance*, SIAM J. Math. Anal., 12 (1981), pp. 653–668.

[19] I. P. Van den Berg, *Nonstandard Asymptotic Analysis*, vol. 1249, Lecture Notes in Mathematics, Springer-Verlag, New York, 1987.

[20] I. P. VAN DEN BERG, *On solutions of polynomial growth of ordinary differential equations*, J. Differential Equations, 81 (1989), pp. 368–402.
[21] W. WASOW, *Linear Turning Point Theory*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 174–178, January 1994

006

A NEW STANDARD ISOMETRY OF DEVELOPABLE SURFACES IN CAD/CAM*

ERWIN KREYSZIG†

Abstract. This paper discusses a recent exact method by Clements and Leon [*Marine Tech.*, 18 (1981), pp. 227–233] for the isometric mapping of developable surfaces into the plane, as needed in computer-aided design and manufacturing (CAD/CAM), and some practical shortcomings of this method. A new exact method for that purpose is then presented that is free of those deficiencies. Of equal practical importance is the fact that, whereas the method by Clements and Leon requires the numerical solution of a second-order nonlinear ordinary differential equation (or of an equivalent first-order system), the present method involves only the evaluation of two single integrals.

Key words. developable surface, isometric mapping into the plane

AMS subject classifications. 53, 65

1. Introduction.

Computer-aided design and manufacturing (CAD/CAM), robotics, computer graphics, and pattern recognition are among the fields that provide various, novel, and practical problems and applications of the differential-geometric theory of curves and surfaces in space [1], [6]. Software reduction calls for standardization to a relatively small number of well-documented portable codes corresponding to fast and efficient algorithms. With respect to curves this means a preference for cubic splines, Bezier curves (named after P. Bezier, of the French Renault Automobile Company), or B-splines, instead of the host of special algebraic and transcendental curves explored during the 18th and 19th centuries. With respect to surfaces the situation is similar, although more complex. For a simple introduction into these matters, see [6].

In construction work, if shape and mechanical stability permit it, one often chooses (portions of) developable surfaces because then the prescribed design surface S can be obtained by cutting a suitable portion S^* from plane material and bending S^* to give it the form of S. Bending is an isometry, a length-preserving mapping [4], and *standardization* here means to design, once and for all, standard isometric mappings $S \to S^*$ and corresponding software applicable to any of the usual practical problems as they arise in ship building, where large steel plates are to be cut, in roof constructions, in airfoil design, and in numerous other tasks; see the references in [2] and [3].

Two not quite satisfactory approximation methods for setting up an isometry $S \to S^*$ into the plane are mentioned in [3]. The paper [3] itself seems to contain the earliest method for that purpose based on an exact differential-geometric theory. In §2 we outline the basics of this method. In §3 we show that it is essentially local and discuss its theoretical and practical limitations. In §4 we present a new method that is free of those limitations. Moreover, our method is numerically much simpler because, instead of the numerical solution of a nonlinear second-order differential equation (or an equivalent first-order system) required in [3], it involves only the (generally numerical) evaluation of two single integrals. A proof of isometry of the mapping given by the algorithm in §4 is provided in the last two sections.

*Received by the editors October 10, 1992; accepted for publication January 11, 1993.

†Department of Mathematics and Statistics, Carleton University, Ottawa, Canada K1S 5B6.

2. Clements–Leon method [3]. An isometry is a geodesic mapping; hence it maps every geodesic on a developable surface S into a straight line in a plane, called the pq-plane. Accordingly, it is natural to determine on a given developable surface

$$S: \quad \mathbf{x}(s,t) = \mathbf{y}(s) + t\mathbf{z}(s),$$

a geodesic $G: t = \tilde{t}(s)$, and map G into a convenient straight line in the pq-plane, say, into a straight-line segment $0 \le p \le P$ on the p-axis.

Theoretically, this is simple. The essential formulas are as follows. First we have

$$G: \quad \tilde{\mathbf{x}}(s) = \mathbf{x}(s, \tilde{t}(s)) = \mathbf{y}(s) + \tilde{t}\mathbf{z}(s).$$

The second derivative with respect to s is

$$\tilde{\mathbf{x}}'' = \mathbf{y}'' + \tilde{t}''\mathbf{z} + 2\tilde{t}'\mathbf{z}' + \tilde{t}\mathbf{z}''.$$

G is geodesic on S if $\kappa_g = 0$, thus $\tilde{\mathbf{x}}'' \cdot \mathbf{w} = 0$, where $\mathbf{w} = \tilde{\mathbf{x}}' \times \tilde{\mathbf{n}}$ and $\tilde{\mathbf{n}}$ is a normal vector of S along G (cf. [4, p. 155]). We thus obtain for $\tilde{t}(s)$ the nonlinear differential equation

$$(\tilde{t}''\mathbf{z} + 2\tilde{t}'\mathbf{z}' + \tilde{t}\mathbf{z}'' + \mathbf{y}'') \cdot \mathbf{w} = 0.$$

Except for the notation in [3], this equation is written in the form

$$\tilde{t}'' = f(s, \tilde{t}(s), \tilde{t}'(s))$$

or as an equivalent first-order system of two differential equations, as usual. For details on differentiability conditions, etc., we refer to [3]. This is the theory. Practically, this equation or system is now solved *numerically*, subject to two suitable initial conditions by which the geodesic G is uniquely determined. Then, using conformality of an isometry, one constructs the plane image S^* of S by means of the images of the generators of S, which are straight-line segments of the same length as their inverse images on S. These are the essential ideas in [3].

3. Restrictions of the method in [3]. In the method just described, one must tacitly impose the condition that on S there exists at least one geodesic that has a point in common with each generator on S. Furthermore, in choosing initial conditions as mentioned in §2, one must be skillful in order to obtain a geodesic of that kind and resulting from it an isometric image of the *entire* given surface S; also, such a geodesic must not meet an edge of regression or any other singularity that S may have.

However, that condition tacitly assumed need not hold even in very simple cases. For instance, think of the surface of a straight circular cone whose plane image (obtained after cutting along a generator) has at the image of the apex an angle $\alpha \ge \pi$. Accordingly, in this sense the method is *local*, and it may be necessary to proceed "piecewise," that is, work with portions of several geodesics; this appears to be a detour. Remarks at the beginning of §4 on p. 970 in [3] seem to indicate that the authors were aware of those practical difficulties.

In the next section we present a method that gives a differential geometrically exact standard mapping that is *global* and makes direct use of the given representation of S in the sense that the auxiliary construction of a geodesic is avoided, and the integration of two single integrals is all that is needed numerically, as was mentioned before.

4. Global standard isometry of a developable surface into the plane. Given a developable surface

$$S: \quad \mathbf{x}(s,t) = \mathbf{y}(s) + t\mathbf{z}(s) \qquad [t_1(s) \le t \le t_2(s),\ a \le s \le b], \tag{1}$$

where

$$\mathbf{z}(s) = \sum_{j=1}^{3} \alpha_j(s)\mathbf{v}_j(s), \quad \sum_{j=1}^{3} \alpha_j^2(s) = 1, \quad \alpha_3(s) \neq 0, \tag{2}$$

$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is the trihedron (cf. [4, p. 36] and [5, p. 469]), of the curve $C: \mathbf{y}(s)$ with arc length s and curvature $\kappa(s) > 0$.

Note that $\alpha_3(s) \neq 0$ is no restriction of the geometric shape of S. Geometrically, it means that the representation of S is chosen so that S has no singularities along C, and calls for a change of the representation if it is violated somewhere (or everywhere) along C.

ALGORITHM ISOM(S, $\Pi(\epsilon)$). For a developable surface S given by (1) and (2), this algorithm computes an isometric image S^* in the pq-plane given by (9).

Input. Developable surface S given by (1), (2), partition $\Pi(\epsilon)$ of $[a, b]$ fine enough so that the integration rule chosen (e.g., Simpson's rule) will give values (5) and (7), whose absolute errors do not exceed a given $\epsilon > 0$.

Output. Isometric image S^* of S given by (9) in the pq-plane.

(i) Calculate the curvature $\kappa(s)$ of C, a unit normal vector $\mathbf{n}$ of S along C,

$$\mathbf{n}(s) = \frac{1}{|\mathbf{v}_1 \times \mathbf{z}|}\mathbf{v}_1 \times \mathbf{z} \qquad (a \le s \le b), \tag{3}$$

and the geodesic curvature κ_g of C on S,

$$\kappa_g = \mathbf{v}_3 \cdot \mathbf{n} \qquad (a \le s \le b). \tag{4}$$

(ii) Evaluate the integral

$$L(s) = \int_0^s \kappa_g(u)\, du \qquad [s \in \Pi(\epsilon)]. \tag{5}$$

(iii) Compute

$$\mathbf{v}_1^*(s) = \begin{bmatrix} \cos L(s) \\ \sin L(s) \end{bmatrix}, \quad \mathbf{v}_2^*(s) = \begin{bmatrix} -\sin L(s) \\ \cos L(s) \end{bmatrix} \quad [s \in \Pi(\epsilon)]. \tag{6}$$

(iv) Evaluate the integral

$$\mathbf{y}^*(s) = \int_0^s \mathbf{v}_1^*(\sigma)\, d\sigma \qquad [s \in \Pi(\epsilon)]. \tag{7}$$

(v) Compute

$$\mathbf{z}^*(s) = \alpha_1 \mathbf{v}_1^*(s) + (1 - \alpha_1^2)^{1/2}\mathbf{v}_2^*(s) \qquad [s \in \Pi(\epsilon)]. \tag{8}$$

(vi) Compute the isometric image S^* of S given by

$$\mathbf{x}^*(s,t) = \mathbf{y}^*(s) + t\mathbf{z}^*(s) \qquad [t_1(s) \le t \le t_2(s),\ s \in \Pi(\epsilon)]. \tag{9}$$

End.

5. Auxiliary formulas for the isometry proof. For easy reference, we continue the equation numbering. From (6),

$$|\mathbf{v}_1^*| = |\mathbf{v}_2^*| = 1, \qquad \mathbf{v}_1^* \cdot \mathbf{v}_2^* = 0. \tag{10}$$

From this and (7),

$$|\mathbf{y}^{*\prime}| = 1; \tag{11}$$

hence s is the arc length of the image $C^* : \mathbf{y}^*(s)$ of C. Also, by (8),

$$|\mathbf{z}^*|^2 = \alpha_1^2 + 1 - \alpha_1^2 = 1. \tag{12}$$

For the unit normal vector $\mathbf{n}$ of S along C we get, by (1) and (2),

$$\mathbf{n} = \frac{\mathbf{x}_s \times \mathbf{x}_t}{|\mathbf{x}_s \times \mathbf{x}_t|} = \frac{\mathbf{v}_1 \times \mathbf{z}}{|\mathbf{v}_1 \times \mathbf{z}|} = \frac{\alpha_2 \mathbf{v}_3 - \alpha_3 \mathbf{v}_2}{(1-\alpha_1^2)^{1/2}}. \tag{13}$$

From this we get by the first Frenet formula (cf. [4, pp. 41, 155])

$$\kappa_g = |\mathbf{y}'\ \mathbf{y}''\ \mathbf{n}| = (\mathbf{v}_1 \times \kappa \mathbf{v}_2) \cdot \mathbf{n} = \kappa \mathbf{v}_3 \cdot \mathbf{n} = \frac{\kappa \alpha_2}{(1-\alpha_1^2)^{1/2}}. \tag{14}$$

From (6) we also have

$$\mathbf{v}_1^{*\prime} = \kappa_g \mathbf{v}_2^*, \qquad \mathbf{v}_2^{*\prime} = -\kappa_g \mathbf{v}_1^*. \tag{15}$$

6. Proof that S and S^* are isometric. We have an isometry if and only if corresponding coefficients of the first fundamental forms of S and S^* are equal [4, p. 176]). With the usual notation $E = \mathbf{x}_s \cdot \mathbf{x}_s$, $F = \mathbf{x}_s \cdot \mathbf{x}_t$, $G = \mathbf{x}_t \cdot \mathbf{x}_t$, we have for S by (1), (2), and $\mathbf{z}' \cdot \mathbf{z} = 0$,

$$E = 1 + 2t\mathbf{y}' \cdot \mathbf{z}' + t^2 \mathbf{z}' \cdot \mathbf{z}', \quad F = \mathbf{v}_1 \cdot \mathbf{z}, \quad G = 1.$$

Similarly for S^* by (9), (11), (12),

$$E^* = 1 + 2t\mathbf{y}^{*\prime} \cdot \mathbf{z}^{*\prime} + t^2 \mathbf{z}^{*\prime} \cdot \mathbf{z}^{*\prime}, \quad F^* = \mathbf{v}_1^* \cdot \mathbf{z}^*, \quad G^* = 1.$$

Accordingly, we have to show that

$$\mathbf{v}_1^* \cdot \mathbf{z}^* = \mathbf{v}_1 \cdot \mathbf{z}, \tag{16}$$

$$\mathbf{v}_1^* \cdot \mathbf{z}^{*\prime} = \mathbf{v}_1 \cdot \mathbf{z}', \tag{17}$$

$$\mathbf{z}^{*\prime} \cdot \mathbf{z}^{*\prime} = \mathbf{z}' \cdot \mathbf{z}'. \tag{18}$$

Now (16) follows from (2) and (8):

$$\mathbf{v}_1^* \cdot \mathbf{z}^* = \alpha_1 = \mathbf{v}_1 \cdot \mathbf{z}. \tag{19}$$

We prove (17). Differentiating (19), we first have

$$\mathbf{v}_1^{*\prime} \cdot \mathbf{z}^* + \mathbf{v}_1^* \cdot \mathbf{z}^{*\prime} = \mathbf{v}_1' \cdot \mathbf{z} + \mathbf{v}_1 \cdot \mathbf{z}'$$

and obtain (17) by proving that the first terms on both sides are equal. Now by (15), the first term on the left equals $\kappa_g \mathbf{v}_2^* \cdot \mathbf{z}^*$, which equals $\kappa_g(1-\alpha_1^2)^{1/2}$ by (8) and (10), and this equals $\alpha_2\kappa$ by (14). But the first term on the right equals $\kappa\mathbf{v}_2 \cdot \mathbf{z}$ by the first Frenet formula, hence $\alpha_2\kappa$ by (2).

We prove (18). From $\alpha_3(s) \neq 0$ in (2) we have $|\alpha_1(s)| < 1$; hence $\mathbf{v}_1$ and $\mathbf{z}$ are linearly independent, so that the developability condition $|\mathbf{z}\ \mathbf{z}'\mathbf{v}_1| = 0$ (cf. [4, p. 182]) implies that

$$\mathbf{z}' = \beta_1\mathbf{v}_1 + \beta_2\mathbf{z}, \quad \text{thus } 0 = \mathbf{z} \cdot \mathbf{z}' = \beta_1\alpha_1 + \beta_2$$

(with β_1 and β_2 depending on s), as well as

$$\mathbf{v}_1 \cdot \mathbf{z}' = \beta_1 + \beta_2\alpha_1 = \beta_1(1-\alpha_1^2). \tag{20}$$

By Lagrange's identity and (8), (10), and (12),

$$|\mathbf{v}_1^* \times \mathbf{z}^*|^2 = 1 - \alpha_1^2 \neq 0,$$

so that $\mathbf{v}_1^*$ and $\mathbf{z}^*$ are linearly independent and we can write

$$\mathbf{z}^{*\prime} = \beta_1^*\mathbf{v}_1^* + \beta_2^*\mathbf{z}^*, \quad \text{thus } 0 = \mathbf{z}^* \cdot \mathbf{z}^{*\prime} = \beta_1^*\alpha_1 + \beta_2^*$$

(with β_1^* and β_2^* depending on s), as well as, by (8) and (10),

$$\mathbf{v}_1^* \cdot \mathbf{z}^{*\prime} = \beta_1^* + \beta_2^*\alpha_1 = \beta_1^*(1-\alpha_1^2). \tag{21}$$

But $\mathbf{v}_1 \cdot \mathbf{z}' = \mathbf{v}_1^* \cdot \mathbf{z}^{*\prime}$ by (17), so that (20) and (21) imply that $\beta_1^* = \beta_1$, and then $\beta_2^* = \beta_2$. Together,

$$\mathbf{z}' = \beta_1\mathbf{v}_1 + \beta_2\mathbf{z}, \qquad \mathbf{z}^{*\prime} = \beta_1\mathbf{v}_1^* + \beta_2\mathbf{z}^*,$$

and (18) now follows because of (10) and (12). This proves isometry.

REFERENCES

[1] R. E. BARNHILL AND R. F. RIESENFELD, *Computer-Aided Geometric Design*, Academic Press, New York, 1972.

[2] J. C. CLEMENTS, *A computer system to derive developable hull surfaces and table of offsets*, Marine Tech., 18 (1981), pp. 227–233.

[3] J. C. CLEMENTS AND L. J. LEON, *A fast, accurate algorithm for the isometric mapping of a developable surface*, SIAM J. Math. Anal. 18 (1987), pp. 966–971.

[4] E. KREYSZIG, *Differential Geometry*, Dover, New York, 1991.

[5] ———, *Advanced Engineering Mathematics*, 7th ed., John Wiley, New York, 1993.

[6] M. E. MORTENSON, *Geometric Modeling*, John Wiley, New York, 1987.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 179–196, Jnauary 1994

007

EXISTENCE OF A HOMOCLINIC ORBIT OF THE LORENZ SYSTEM BY PRECISE SHOOTING*

BRIAN HASSARD† AND JIANHE ZHANG†

Abstract. A proof of the existence of a homoclinic orbit of the Lorenz system is given, based on shooting using precise numerical integrations. This makes rigorous an argument given by Sparrow [*Appl. Math. Sci.* 41, Springer-Verlag, New York, 1982], and improves greatly on the accuracy of the estimate of the value R^* for which a homoclinic orbit exists, in the Hastings–Troy proof [*Bull. Amer. Math. Soc.*, 27 (1992), pp. 128–131, and *J. Diff. Eqns.*, 1993, to appear].

Key words. Lorenz system, homoclinic orbit, precise integration

AMS subject classifications. 34A50, 34C37, 65L70

1. Introduction.

The Lorenz system [5]

$$x' = s(y - x), \tag{1.1}$$

$$y' = (R - z)x - y, \tag{1.2}$$

$$z' = xy - qz, \tag{1.3}$$

is often cited as a simple ordinary differential system which exhibits complex dynamical behavior, including chaos. However, remarkably little of this behavior has been actually proven to exist. Our knowledge of the more interesting phenomena in the system, derives from traditional numerical integrations lacking rigorous control of truncation and roundoff error.

Only recently has even existence of a homoclinic orbit for the system been proven. Hastings and Troy [2], [3] show for each (s, q) in some neighborhood of (10,1), there is an R in the range (1,1000) for which (1.1) has a homoclinic orbit.

Here we give a proof which is essentially a rigorous version of the numerical shooting argument in Sparrow [6] that a homoclinic orbit exists. The numerical part of our proof uses a precise numerical integrator which produces true error bounds. As in [2] and [3], the existence of a homoclinic orbit is inferred as a consequence of properties of the just two computed orbits (shots). As in [6], the estimate of a value of R for which a homoclinic orbit exists is sharpened by additional shooting.

Each orbit we compute has one of the following properties.

Property P. As $t \to -\infty, (x, y, z)^T \to (0, 0, 0)^T$. Also, there exist numbers $\tau_1 < s_1 < t_1$ such that

$$\begin{aligned}
&x' > 0 \quad \text{on } (-\infty, \tau_1),\\
&x' < 0 \quad \text{on } (\tau_1, t_1],\\
&x(t_1) = 0,\\
&y' < 0 \quad \text{on } [\tau_1, s_1],\\
&y(s_1) = 0,\\
&y < 0 \quad \text{on } (s_1, t_1],\\
&x > 0 \quad \text{on } (-\infty, t_1).
\end{aligned}$$

*Received by the editors July 27, 1992; accepted for publication February 12, 1993.

†Department of Mathematics, State University of New York at Buffalo, Buffalo, New York, 14214.

Property Q. As $t \to -\infty, (x, y, z)^T \to (0, 0, 0)^T$. There exists a τ_1 such that $x' > 0$ on $(-\infty, \tau_1)$. Also, the orbit $(x, y, z)^T$ does not have Property P.

In §2, we show how the existence of orbits having Properties P and Q leads to the existence of a homoclinic orbit. We incorporate more information about orbits into Property P than in the hypotheses of [2, Lemma 2.2] so that we can separate a central argument (Lemma 2.1) from the body (Lemma 2.2) of the proof, and organize the proofs around the concept of sequences of orbits having Property P.

In §§3 and 4 we construct orbits with Properties P and Q. On $(-\infty, 0]$, orbits are approximated by polynomials in $e^{\lambda_3 t}$, where λ_3 is the positive eigenvalue of the Jacobian at $(0, 0, 0)^T$. Initial conditions for the numerical integrations derive from these approximations at $t = 0$. The origin of t is defined such that on the time interval $(-\infty, 0]$, the orbits are in neighborhood of $(0, 0, 0)^T$ sufficiently small that local theory (Lemma 3.4) applies to show existence and to construct highly accurate approximations. High accuracy at $t = 0$ is necessary to compensate for the large loss of precision during the integrations for $t \geq 0$.

In §4, we describe precise numerical integrations of using an algorithm due to Aberth [1]. The actual code we used is based on Aberth's C++ program called "difsys.cc," which generates precise solutions of initial value problems of ordinary differential systems for which the component functions are elementary. In §4 we first describe the modifications we had to make to Aberth's algorithm to achieve our goals, and then we present the results in the form of Lemma 4.1. Finally, we state Theorem 4.2.

2. Conditions for existence of a homoclinic orbit. We fix $q > 0$, $s > 0$ and restrict our attention to some interval $[R_a, R_b]$ of R values, $R_a > 1$. The Jacobian matrix for (1.1)–(1.3) at the origin has two negative and one positive eigenvalues. The unstable manifold at the origin has components $\gamma^+(R)$, $\gamma^-(R)$, where locally $x > 0$ on γ^+, $x < 0$ on γ^-. We will be concerned only with γ^+.

We let $p^+(R) = (\sqrt{q(R-1)}, \sqrt{q(R-1)}, R-1)^T$ denote the stationary point in the positive octant. Let $p(t; R) = (x(t; R), y(t; R), z(t; R))^T$ denote a family of orbits of the Lorenz system, one for each $R \in [R_a, R_b]$, jointly smooth in t and R, known to exist on the interval $(-\infty, 0]$, and such that $\lim_{t \to -\infty} p(t; R) = (0, 0, 0)^T$ and $x'(t, R) > 0$ on $(-\infty, 0]$. Such orbits will be constructed in §3. Then $p(.; R) \in \gamma^+(R)$. It is well known [5, Appendix C] that for fixed parameters q,R,s, there is a region V, the interior of an ellipse, which contains the origin and is a (forward) invariant set for the flow. Because of the invariant region, each orbit $p(t; R)$ on $(-\infty, 0]$ may be continued to the entire line $(-\infty, \infty)$. Furthermore, the partial derivatives of x, y, and z with respect to t of any fixed order, are bounded for all t and bounds which are uniform in R may be constructed.

The main task in establishing the existence of a homoclinic orbit is to show that for the sequence $\{p(.; R_j)\}$ of orbits with Property P constructed in Lemma 2.2, the sequence $\{t_{1,j}\}$ is unbounded. This will be done using the following lemma.

LEMMA 2.1. *Suppose* $0 < q < 2s$ *and* $1 < R_a < R < R_b$. *Let* $R_j \in (R_a, R_b)$ *and suppose each orbit* $p(.; R_j) = (x(.; R_j), y(.; R_j), z(.; R_j))^T$ *has Property P. Let* $\tau_{1,j} < s_{1,j} < t_{1,j}$ *be from Property P such that* $x'(\tau_{1,j}; R_j) = 0$, $y(s_{1,j}; R_j) = 0$ *and* $x(t_{1,j}; R_j) = 0$. *If the sequences* $\{\tau_{1,j}\}, \{s_{1,j}\}, \{t_{1,j}\}$ *have finite limits* τ_1, s_1, t_1, *then the orbit* $(x(.; R), y(.; R), z(.; R))$ *also has Property P.*

The basic existence result is then Lemma 2.2.

LEMMA 2.2. *Suppose* $0 < q < 2s$ *and* $1 < R_a < R_b$. *Suppose that for* $R = R_b$, *there is an orbit with Property P and for* $R = R_a$ *there is an orbit with Property Q.*

Suppose further that for all $R \in [R_a, R_b]$, $p^+(R)$ *is linearly stable. Then there is a value* R^*, $R_a < R^* < R_b$ *for which there is an orbit* $p(.; R^*) = (x, y, z)^T$ *with the following properties.*

As $t \to \pm\infty$, $(x, y, z)^T \to (0, 0, 0)^T$. *Also, there exist* τ_1, s_1, $-\infty < \tau_1 < s_1 < \infty$ *such that*

$$
\begin{aligned}
&x' > 0 \quad \textit{on } (-\infty, \tau_1),\\
&x' < 0 \quad \textit{on } (\tau_1, \infty),\\
&y' < 0 \quad \textit{on } [\tau_1, s_1],\\
&y(s_1) = 0,\\
&y < 0 \quad \textit{on } (s_1, \infty),\\
&x > 0 \quad \textit{on } (-\infty, \infty).
\end{aligned}
$$

The remainder of this section gives the proofs.

Proof of Lemma 2.1. As τ_1, s_1, t_1 are the limits of $\{\tau_{1,j}\}$, $\{s_{1,j}\}$, $\{t_{1,j}\}$, it follows from Property P that

$$
\begin{aligned}
&\tau_1 \le s_1 \le t_1,\\
&x' \ge 0 \quad \text{on } (-\infty, \tau_1),\\
&x' \le 0 \quad \text{on } (\tau_1, t_1],\\
&x(t_1) = 0,\\
&y' \le 0 \quad \text{on } [\tau_1, s_1],\\
&y(s_1) = 0,\\
&y \le 0 \quad \text{on } (s_1, t_1],\\
&x \ge 0 \quad \text{on } (-\infty, t_1),
\end{aligned}
$$

where $x(t) = x(t; R)$ and $y(t) = y(t; R)$.

To establish that $p(.; R)$ has Property P, we only need to show

(a) $\tau_1 \neq s_1$,

(b) $s_1 \neq t_1$,

(c) $x \neq 0$ on $(-\infty, t_1)$,

(d) $x' \neq 0$ on $(-\infty, \tau_1)$,

(e) $y' \neq 0$ on (τ_1, s_1),

(f) $x' \neq 0$ on (τ_1, t_1),

(g) $y'(\tau_1) \neq 0$,

(h) $y \neq 0$ on (s_1, t_1),

(i) $y'(s_1) \neq 0$,

(j) $y(t_1) \neq 0$,

(k) $x'(t_1) \neq 0$.

We first derive an inequality at τ_1 which shows $p(.;R)$ is nontrivial and will be useful below. Let $Q_j(t) = z(t;R_j) - (x(t;R_j)^2/2s)$. Then

$$\begin{aligned} Q_j' &= -qz(t;R_j) + \left(1 - \frac{q}{2s}\right) x(t;R_j)^2, \\ Q_j'' + qQ_j' &= 2\left(1 - \frac{q}{2s}\right) x(t;R_j)x'(t;R_j), \\ Q_j'(t) &= 2\left(1 - \frac{q}{2s}\right) \int_{-\infty}^{t} e^{q(t-\tau)} x(\tau;R_j)x'(\tau;R_j)d\tau, \end{aligned}$$

and from Property P for $p(.;R_j)$, $Q_j'(\tau_{1,j}) > 0$. Thus $z'(\tau_{1,j};R_j) = Q_j'(\tau_{1,j}) > 0$. From (1.1), $y(\tau_{1,j};R_j) = x(\tau_{1,j};R_j)$ so from (1.3), $x^2(\tau_{1,j};R_j) > qz(\tau_{1,j};R_j)$. By P, $y'(\tau_{1,j};R_j) < 0$ so from (1.2) $(R_j - z(\tau_{1,j};R_j) - 1)x(\tau_{1,j};R_j) < 0$ which implies $z(\tau_{1,j};R_j) > R_j - 1$. Then $x^2(\tau_{1,j};R_j) > q(R_j - 1)$. In the limit $j \to \infty$,

$$x(\tau_1;R) = y(\tau_1;R) \geq (q(R-1))^{\frac{1}{2}} > 0. \tag{2.1}$$

We now establish (a)–(k) in order.

(a) If $\tau_1 = s_1$, then $y(\tau_1) = y(s_1) = 0$, contradicting (2.1). Therefore, $\tau_1 < s_1$.

(b) If $s_1 = t_1$, then $x(t_1) = y(t_1) = 0$ and (1.1), (1.2) imply $x(t) = y(t) = 0$ for all t, contradicting (2.1). Therefore, $s_1 < t_1$.

(c) Suppose there is a $t_0 < t_1$ such that $x(t_0) = 0$. By (2.1), $x(\tau_1) \neq 0$ so either $t_0 < \tau_1$ or $t_0 > \tau_1$. If $t_0 < \tau_1$, then as $x'(t) \geq 0$ on $(-\infty, \tau_0]$, $\int_{-\infty}^{t_0} x'(t)dt = x(t_0) = 0$ and x' is continuous, it follows that $x'(t) = 0$ on $(-\infty, t_0]$. Then $x(t) = 0$ on $(-\infty, t_0]$ and (1.1),(1.2) imply $x(t) = y(t) = 0$ for all t, contradicting (2.1).

If $\tau_1 < t_0 < t_1$, then as $x'(t) \leq 0$ on (τ_1, t_1), $\int_{-\infty}^{t_0} x'(t)dt = x(t_1) = 0$ and x' is continuous, it follows that $x'(t) = 0$ on $[t_0, t_1]$. Then (1.1),(1.2) imply $x(t) = y(t) = 0$ for all t, contradicting (2.1).

It follows that $x > 0$ on $(-\infty, t_1)$.

(d) Suppose there is a $\tau_0 < \tau_1$ such that $x'(\tau_0) = 0$.

We claim that $x''(0) = 0$. For, if $x''(0) \neq 0$, $x'(t)$ changes sign at τ_0 and this is impossible since $x' \geq 0$ on $(-\infty, \tau_1)$. We claim further that $x'''(\tau_0) \geq 0$. For, if $x'''(\tau_0) < 0$, then $x'(t) < 0$ for t near τ_0; again, this is impossible.

So now $x'(\tau_0) = x''(\tau_0) = 0$ and $x'''(\tau_0) \geq 0$. Equation (1.1) gives $y(\tau_0) = x(\tau_0)$, (1.1)$'$ gives $y'(\tau_0) = 0$, (1.1)$''$ gives $y''(\tau_0) = x'''(\tau_0)/s \geq 0$, and (1.2) then implies $(R - z(\tau_0) - 1)x(\tau_0) = 0$. But from (c), $x(\tau_0) > 0$ and so $z(\tau_0) = R - 1$. From (1.2)$'$, $y''(\tau_0) = -x(\tau_0)z'(\tau_0) \geq 0$ which implies $z'(\tau_0) \leq 0$.

Let $Q = z - (x^2/2s)$. Then as above $Q'(\tau_0) = (1 - \frac{q}{2s}) \int_{-\infty}^{\tau_0} 2e^{q(t-\tau)}x(\tau)x'(\tau)d\tau$. Since $x(\tau_0) > 0$, $x'(t) > 0$ for at least some $t < \tau_0$. From (c), $x(t) > 0$ on $(-\infty, \tau_0]$, and so $Q'(\tau_0) > 0$. But then $z'(\tau_0) = Q'(\tau_0) > 0$, and so we have a contradiction.

It follows that $x' > 0$ on $(-\infty, \tau_1)$.

(e) Suppose there is a σ_2, $\tau_1 < \sigma_2 < s_1$ such that $y'(\sigma_2) = 0$. We claim that $y''(\sigma_2) = 0$. For if $y''(\sigma_2) \neq 0$, $y'(t)$ changes sign at σ_2 which is impossible since $y' \leq 0$ on (τ_1, s_1).

We also claim that $y(\sigma_2) > 0$. For if $y(\sigma_2) = -\int_{\sigma_1}^{s_1} y'(\tau)d\tau = 0$, then, since $y' \leq 0$ on $[\tau_1, s_1]$ it follows that $y'(t) = 0$ for all $t \in [\sigma_2, s_1]$, so then $y(t) = 0$ on $[\sigma_2, s_1]$. Then (1.2) and (c) imply $z(t) = R$ on $[\sigma_2, s_1]$, and (1.3) gives $-qR = 0$, a contradiction.

So now $y'(\sigma_2) = y''(\sigma_2) = 0$ and $y(\sigma_2) > 0$. From (1.2), $y'(\sigma_2) = (R - z(\sigma_2))x(\sigma_2) - y(\sigma_2) = 0$. From (c), $x(\sigma_2) > 0$ so $z(\sigma_2) < R$. From (1.2)$'$, $y''(\sigma_2) =$

$(R-z)x'-xz'=0$ and it follows that $z'(\sigma_2)\le 0$. From (1.3)', $z''+qz'=x'y+xy'$, and so for all $t\in(\sigma_2,s_1)$, $e^{qt}z'(t)=e^{q\sigma_2}z'(\sigma_2)+\int_{\sigma_2}^{t}e^{q\tau}(x'y+xy')d\tau\le 0$. Therefore $z(s_1)\le z(\sigma_2)<R$. But at s_1, $y'(s_1)=(R-z)x\le 0$, so $z(s_1)\ge R$ a contradiction.

It follows that $y'<0$ on (τ_1,s_1).

(f) Suppose there is a $\tau_2\in(\tau_1,t_1)$ such that $x'(\tau_2)=0$.

If $\tau_2\in(\tau_1,s_1)$, then by the same argument as in (d), $x''(\tau_2)=0$. Then from (1.1)', $y'(\tau_2)=0$ contradicting (e).

If $\tau_2\in[s_1,t_1)$, $y(\tau_2)\le 0$. From (1.1) and (c), $y(\tau_2)=x(\tau_2)>0$, a contradiction.

If $\tau_2=t_1$, then (1.1), (1.2) imply $x(t)=y(t)=0$ for all t, contradicting (2.1).

It follows that $x'<0$ on (τ_1,t_1).

(g) Suppose $y'(\tau_1)=0$. Then from (1.1)', $x''(\tau_1)=0$.

We claim that $x'''(\tau_1)=0$. For otherwise x' does not change sign at τ_1, which we know happens by (d) and (f).

So now $x'(\tau_1)=x''(\tau_1)=x'''(\tau_1)=0$. From (1.1)'', $y''(t_1)=0$ as well. Then from (1.2)', $z'(\tau_1)=0$. But letting Q be as in (d), $z'(\tau_1)=Q'(\tau_1)=(1-\frac{q}{2s})\int_{-\infty}^{\tau_1}2e^{q(t-\tau)}x(\tau)x'(\tau)d\tau>0$ a contradiction.

Since we have shown that $y'(\tau_1)\neq 0$, and from (e) $y'(t)<0$ for t near τ_1, $t>\tau_1$, it follows that $y'(\tau_1)<0$.

(h) Suppose there is a point s_2, $s_1<s_2<t_1$ such that $y(s_2)=0$.

We claim that $y'(s_2)=0$ and $y''(s_2)\le 0$. For otherwise y changes sign at s_2, contradicting $y\le 0$ on (s_1,t_1).

So now $y(s_2)=y'(s_2)=0$ and $y''(s_2)\le 0$. From (1.2), $(R-z(s_2))x(s_2)=0$. By (c), $x(s_2)>0$ so $z(s_2)=R$. From (1.2)', $y''(s_2)=-z'(s_2)x(s_2)\le 0$ and so $z'(s_2)\ge 0$. But from (1.3), $z'(s_2)=-qz(s_2)=-qR<0$, a contradiction.

It follows that $y<0$ on (s_1,t_1).

(i) Suppose that $y'(s_1)=0$.

We claim that $y''(s_1)=0$. For if $y''(s_1)<0$, then $y'(t)>0$ for t near s_1, $t<s_1$ contradicting (e). If $y''(s_1)>0$, then $y(t)>0$ for t near s_1, $t>s_1$ contradicting (h).

So now $y(s_1)=y'(s_1)=y''(s_1)=0$. From (1.2), $(R-z(s_1))x(s_1)=0$. By (c), $x(s_1)>0$, so $z(s_1)=R$. Then from (1.2)', $z'(s_1)=0$. From (1.3), $z'(s_1)=-qz(s_1)=0$, so $z(s_1)=0$, a contradiction.

It follows that $y'(s_1)<0$.

(j) If $y(t_1)=0$, then since $x(t_1)=0$, (1.1), (1.2) give $x(t)=y(t)=0$ for all t contradicting (2.1). Therefore $y(t_1)\neq 0$.

It follows that $y(t_1)<0$.

(k) We have $x'(t_1)=s(y(t_1)-x(t_1))=sy(t_1)<0$ from (j). □

Proof of Lemma 2.2. First we construct a family of orbits $p(t;R)$, $R\in[R_a,R_b]$. Take $J=1$ in the hypotheses of Lemma 3.4. Then choose $\eta>0$ small enough so that the conditions of Lemma 3.4 are satisfied and furthermore $0<x(0;R)<y(0;R)$ and $z(0;R)<R-1$ for all $R\in[R_a,R_b]$ so the conditions of Lemma 3.5 are satisfied. Then by Lemmas 3.4 and 3.5, there exists a family of orbits $p(t;R), t\in(-\infty,0]$ such that $\lim_{t\to-\infty}p(t;R)=0$ and $x'(0;R)>0$ on $(-\infty,0]$. Furthermore $p(t;R)$ is jointly smooth in t and R.

By uniqueness of the unstable manifold, the hypothesized orbit for $R=R_b$ differs from $p(.;R_b)$ only by a phase shift, so $p(.;R_b)$ has Property P. Similarly, $p(.;R_a)$ has Property Q.

If an orbit $p(.;R), R>1$ has Property P, then we claim that all orbits $p(.;\rho)$ for ρ in some neighborhood of R also have Property P. One establishes this claim by first using the implicit function theorem and $x''(\tau_1(R);R)=sy'(\tau_1(R);R)<0$,

$y'(s_1(R);R) < 0$ and $x'(t_1(R);R) < 0$ to show the existence of $\tau_1(\rho)$, $s_1(\rho)$, and $t_1(\rho)$ such that $x'(\tau_1(\rho);\rho) = y(s_1(\rho);\rho) = x(t_1(\rho);\rho) = 0$. As $x'(t;\rho) > 0$ on $(-\infty, 0]$, $\tau_1(\rho) > 0$ and we may restrict our attention to the bounded set $[0, t_1(R)+1] \times [R_a, R_b]$. We omit further details.

Let

$$R^* = \inf\{R, R_a < R < R_b \text{ such that } p(.;R) \text{ has Property P}\}.$$

Then $p(.;R^*)$ does not have Property P. If $R^* = R_a$ this is immediate. If $R^* > R_a$ and $p(.;R^*)$ has Property P, then so do $p(.;\rho)$ for ρ sufficiently close to R^*,$\rho < R^*$ contradicting the definition of R^*.

Let $\{R_j\}$ be decreasing towards R^* and such that each $p(.;R_j)$ has Property P. Let $\{\tau_{1,j}\}$, $\{s_{1,j}\}$, and $\{t_{1,j}\}$ be the associated values such that $x'(\tau_{1,j};R_j) = y(s_{1,j};R_j) = x(t_{1,j};R_j) = 0$. By construction, $x'(t;R_j) > 0$ on $(-\infty, 0]$ and so $0 < \tau_{1,j} < s_{1,j} < t_{1,j}$ for each j.

We now claim $\{t_{1,j}\}$ is unbounded. For, if $\{t_{1,j}\}$ is bounded, so are $\{\tau_{1,j}\}$ and $\{s_{1,j}\}$ and we may extract convergent subsequences $\{t_{1,k}\}$, $\{\tau_{1,k}\}$, $\{s_{1,k}\}$, and define t_1, τ_1, s_1 as the limits. By Lemma 2.1, the orbit for $R = R^*$ would then have Property P, a contradiction.

So now $\{t_{1,j}\}$ is unbounded. Let $\{t_{1,j'}\}$ be a subsequence such that $t_{1,j'} \to \infty$. We claim that $\{\tau_{1,j'}\}$ is bounded. For suppose not. Then taking the limit $k' \to \infty$ over a subsequence of $\{j'\}$ such that $\tau_{1,k'} \to \infty$, the orbit $p(.;R^*)$ has the property that x is nondecreasing for all t. Since $0 < x(0;R^*)$, $p(.;R^*)$ is nontrivial and the only possibility is that $\lim_{t\to\infty} p(t;R^*) = p^+(R^*)$. By hypothesis, for all $R \in [R_a, R_b]$, $p^+(R)$ is a linearly stable stationary point. If $p(.;R^*)$ is a heteroclinic connection between $(0,0,0)^T$ and $p^+(R^*)$ and $x(t;R^*) > 0$ for all t, then for all R sufficiently close to R^*, $p(.;R)$ is a heteroclinic connection between $(0,0,0)^T$ and $p^+(R)$, and $x(t;R) > 0$ for all t. Then for all sufficiently large k' $p(.;R_{k'})$ does not have Property P. This contradicts the definition of R^*.

So now $\{\tau_{1,j'}\}$ is bounded. Let $\{\tau_{1,j''}\}$ be a subsequence such that

$$\tau_1 = \lim_{j''\to\infty} \tau_{1,j''}$$

exists. As $x'(t,R^*) = \lim_{j''\to\infty} x'(t;R_{j''})$, $x'(t) \le 0$ for $t > \tau_1$. So $x(t,R^*)$ is monotone when $t > \tau_1$ which implies $\lim_{t\to\infty} p(t,R^*) = p_\infty$ exists and is an equilibrum point. As $x(.;R^*) \ge 0$, the only possibilities are $p_\infty = p^+(R^*)$ and $p_\infty = (0,0,0)^T$. If $p_\infty = p^+(R^*)$, then $p(.;R^*)$ is a heteroclinic connection between $(0,0,0)^T$ and $p^+(R^*)$ which contradicts the definition of R^* by the same argument as above. Therefore $\lim_{t\to\infty} p(t,R^*) = (0,0,0)^T$, that is, $p(.;R^*)$ is a homoclinic orbit.

We now claim $\{s_{1,j''}\}$ is bounded. For if not, let $\{s_{1,k''}\}$ be a subsequence such that $s_{1,k''} \to \infty$, then $y'(t,R^*) = \lim_{k''\to\infty} y'(t,R_{k''}) \le 0$ for $t > \tau_1$. On the other hand, $y' = (z-R^*)x - y \ge (z-R^*-1)x$ when $t > \tau_1$. From $\lim_{t\to\infty} p(t,R^*) = (0,0,0)^T$, $z < R^* - 1$ when t is large. So one can find large t such that $y'(t) > 0$, which is a contradiction.

What remains to be shown are:

$$\begin{aligned} x' &\neq 0 \quad \text{on } (-\infty, \tau_1), \\ x' &\neq 0 \quad \text{on } (\tau_1, \infty), \\ y' &\neq 0 \quad \text{on } (\tau_1, s_1], \\ y &\neq 0 \quad \text{on } (s_1, \infty), \\ x &\neq 0 \quad \text{on } (-\infty, \infty). \end{aligned}$$

These inequalities are established the same way as in the proof of Lemma 2.1. □

3. Local approximation of the unstable manifold. Our objective in this section is to construct highly accurate approximations to solutions $p(t)$ of (1.1)–(1.3) with the property $\lim_{t\to-\infty} p(t) = (0,0,0)^T$. The construction also shows existence. The first step is a coordinate transformation.

Let $s' = s - 1$ and $d = ((s')^2 + 4sR)^{\frac{1}{2}}$. The eigenvalues of the Jacobian matrix for (1.1) at the origin are then $\lambda_1 = -q < 0$, $\lambda_2 = ((-(s+1) - d)/2) < 0$, and $\lambda_3 = ((-(s+1) + d)/2) > 0$. Let Λ denote the matrix whose columns are the eigenvectors $(0,0,1)^T$, $(1,c_2,0)^T$, and $(1,c_3,0)^T$ corresponding to λ_1, λ_2, and λ_3. Here $c_2 = ((s'-d)/2s) < 0$ and $c_3 = ((s'+d)/2s) > 0$. Let $c_1 = s/d > 0$. Setting $(x,y,z)^T = \Lambda w$, (1.1)–(1.3) becomes

$$\begin{aligned} \frac{dw_1}{dt} &= \lambda_1 w_1 + f_1 = \lambda_1 w_1 + (c_2 w_2 + c_3 w_3)(w_2 + w_3), \\ \frac{dw_2}{dt} &= \lambda_2 w_2 + f_2 = \lambda_2 w_2 + c_1 w_1 (w_2 + w_3), \\ \frac{dw_3}{dt} &= \lambda_3 w_3 + f_3 = \lambda_3 w_3 - c_1 w_1 (w_2 + w_3). \end{aligned} \tag{3.1}$$

We shall find solutions of (3.1) with the property $\lim_{t\to-\infty} w = (0,0,0)^T$, by solving the system of integral equations

$$\begin{aligned} w_1(t) &= \int_{\tau=-\infty}^{t} e^{\lambda_1(t-\tau)}(c_2 w_2(\tau) + c_3 w_3(\tau))(w_2(\tau) + w_3(\tau))d\tau, \\ w_2(t) &= \int_{\tau=-\infty}^{t} e^{\lambda_2(t-\tau)} c_1 w_1(\tau)(w_2(\tau) + w_3(\tau))d\tau, \\ w_3(t) &= \eta e^{\lambda_3 t} - \int_{\tau=t}^{0} e^{\lambda_3(t-\tau)} c_1 w_1(\tau)(w_2(\tau) + w_3(\tau))d\tau. \end{aligned} \tag{3.2}$$

To start, we define some functionals.

DEFINITION 3.1. For bounded continuous R^3-valued functions $u(t), v(t)$ on $(-\infty, 0]$ and for real η, let $\beta(u,v) = (\beta_1, \beta_2, \beta_3)^T$, $\varphi(u)$ and $\psi(u;\eta)$ denote the functionals

$$\begin{aligned} \beta_1(u,v)(t) &= \int_{-\infty}^{t} e^{\lambda_1(t-\tau)}(c_2 u_2(\tau) + c_3 u_3(\tau))(v_2(\tau) + v_3(\tau))d\tau, \\ \beta_2(u,v)(t) &= \int_{-\infty}^{t} e^{\lambda_2(t-\tau)} c_1 u_1(\tau)(v_2(\tau) + v_3(\tau))d\tau, \\ \beta_3(u,v)(t) &= -\int_{t}^{0} e^{\lambda_3(t-\tau)} c_1 u_1(\tau)(v_2(\tau) + v_3(\tau))d\tau, \end{aligned} \tag{3.3}$$

$$\varphi(u) = \beta(u,u), \tag{3.4}$$

$$\psi(u;\eta) = (0,0,\eta e^{\lambda_3 t})^T + \varphi(u). \tag{3.5}$$

The system (3.2) then becomes simply

$$w = \psi(w;\eta). \tag{3.6}$$

DEFINITION 3.2. Let $W^1 = (0, 0, \eta e^{\lambda_3 t})^T$, and for each $J \geq 2$, let

$$W^J(t) = W^{J-1}(t) + \frac{\eta^J}{J!} \frac{\partial^J \varphi(W^{J-1})(t)}{\partial \eta^J} \Big|_{\eta=0} . \tag{3.7}$$

We use W^J for some fixed J as the approximation to the unstable manifold for (3.1) at the origin. Our objective in the remainder of this section is to establish the error bound in Lemma 3.4. (U^J defined by $U^1 = W^1$, $U^j = \psi(U^{j-1}; \eta)$ for $2 \leq j < J$, has similar approximation properties to W^J but is impractical here since U^J is a polynomial of degree 2^{J-1} in η.)

LEMMA 3.1. *For each $J \geq 1$, the component functions $W_i^J, i = 1, 2, 3$ of W^J have the form*

$$W_i^J = \sum_{j=1}^{J} \eta^j \sum_{k=1}^{j} W_{ijk} \xi^k, \tag{3.8}$$

where $\xi = e^{\lambda_3 t}$ and the coefficients W_{ijk} are constants given recursively by (3.9), (3.10), *and* (3.12) *below.*

Proof. The components of W^1 are of the form (3.8): explicitly,

$$W_{111} = W_{211} = 0 \quad \text{and} \quad W_{311} = 1. \tag{3.9}$$

For $J \geq 2$, assume for induction that W^{J-1} has components of the form (3.8). From (3.1) the components of $f(W^{J-1})$ are then

$$f_i(W^{J-1}) = \sum_{j=1}^{2J-2} \eta^j \sum_{k=1}^{j} \xi^k f_{ijk}^{J-1},$$

where for $1 \leq i \leq 3$, $1 \leq j \leq 2J-2$, and $1 \leq k \leq j$,

$$\begin{aligned} f_{1jk}^{J-1} &= \sum_{j'+j''=j} \sum_{k'+k''=k} (c_2 W_{2j'k'} + c_3 W_{3j'k'})(W_{2j''k''} + W_{3j''k''}), \\ f_{2jk}^{J-1} &= \sum_{j'+j''=j} \sum_{k'+k''=k} c_1 W_{1j'k'} (W_{2j''k''} + W_{3j''k''}), \\ f_{3jk}^{J-1} &= -f_{2jk}^{J-1}, \end{aligned} \tag{3.10}$$

and $1 \leq j', j'', k', k'' \leq J-1$ in the summations and $f_{ijk}^{J-1} = 0$ when no terms are present. From (3.7), (3.4), and (3.2),

$$\begin{aligned} W_i^J &= W_i^{J-1} + \eta^J \int_{-\infty}^{t} e^{\lambda_i (t-\tau)} \sum_{k=2}^{J} e^{k\lambda_3 \tau} f_{iJk}^{J-1} d\tau \\ &= W_i^{J-1} + \eta^J \sum_{k=2}^{J} e^{k\lambda_3 t} \frac{f_{iJk}^{J-1}}{k\lambda_3 - \lambda_i}, \qquad i = 1, 2, \\ W_3^J &= W_3^{J-1} - \eta^J \int_{t}^{0} e^{\lambda_3 (t-\tau)} \sum_{k=2}^{J} e^{k\lambda_3 \tau} f_{3Jk}^{J-1} d\tau \\ &= W_3^{J-1} + \eta^J \sum_{k=2}^{J} (e^{k\lambda_3 t} - e^{\lambda_3 t}) \frac{f_{3Jk}^{J-1}}{k\lambda_3 - \lambda_3}. \end{aligned} \tag{3.11}$$

Replacing $e^{\lambda_3 t}$ with ξ in (3.11) shows that the functions W_i^J are of the form (3.8), where the new coefficients are

$$
\begin{aligned}
W_{iJk} &= \frac{f_{iJk}^{J-1}}{k\lambda_3 - \lambda_i}, \qquad 1 \le i \le 2,\ 1 \le k \le J, \\
W_{3J1} &= -\sum_{k=2}^{J} \frac{f_{3Jk}^{J-1}}{k\lambda_3 - \lambda_3}, \\
W_{3Jk} &= \frac{f_{3Jk}^{J-1}}{k\lambda_3 - \lambda_3}, \qquad 2 \le k \le J.
\end{aligned} \tag{3.12}
$$

This completes the induction. □

It is easy to show that the components of $\psi(W^J;\eta)$ are of the general form

$$\psi_i(W^J;\eta) = \sum_{j=1}^{2J} \eta^j \sum_{k=1}^{j} \xi^k \psi_{ijk}^J.$$

The purpose of the following two lemmas is to show that in fact $\psi_{ijk}^J = W_{ijk}$ for $1 \le j \le J$ and $1 \le k \le j$. The "$O(\eta^{J+1})$" reasoning is simply a quick way to achieve this result. The actual bounding argument does not appear until Lemma 3.4.

LEMMA 3.2. *For each $J \ge 2$, $\psi(W^{J-1};\eta) = W^J + O(\eta^{J+1})$ for any fixed $t \le 0$.*

Proof. For $J = 2$ this is true since $W^2 = \psi(W^1;\eta)$. Consider $J \ge 3$. For each $2 \le j \le J$, since $W^{j-1} = W^{J-1} + O(\eta^j)$,

$$
\begin{aligned}
\varphi(W^{j-1}) &= \beta(W^{J-1} + O(\eta^j), W^{J-1} + O(\eta^j)) = \beta(W^{J-1}, W^{J-1}) + O(\eta^{j+1}) \\
&= \varphi(W^{j-1}) + O(\eta^{j+1}),
\end{aligned}
$$

so

$$\frac{\partial^j}{\partial \eta^j} |_{\eta=0}\, \psi(W^{j-1}) = \frac{\partial^j}{\partial \eta^j} |_{\eta=0}\, \psi(W^{J-1}).$$

Now

$$\psi(W^{J-1};\eta) - W^J = \varphi(W^{J-1}) + W^1 - W^J = \varphi(W^{J-1}) - \sum_{j=2}^{J} \frac{\eta^j}{j!} \frac{\partial^j}{(\partial \eta)^j} |_{\eta=0}\, \varphi(W^{J-1}).$$

Since $W^{J-1} = O(\eta)$ the linear term in the Taylor expansion of $\varphi(W^{J-1})$ about $\eta = 0$ vanishes, and $\psi(W^{J-1};\eta) - W^J = O(\eta^{J+1})$ as desired. □

LEMMA 3.3. *For each $J \ge 2$,$\psi(W^J;\eta) = W^J + O(\eta^{J+1})$ for any fixed $t \le 0$, and the components of $\psi(W^J;\eta)$ are given by* (3.14)–(3.15) *below.*

Proof. Since $W^J = W^{J-1} + O(\eta^J)$,

$$
\begin{aligned}
\psi(W^J;\eta) &= \psi(W^{J-1} + O(\eta^J);\eta) = W^1 + \beta(W^{J-1} + O(\eta^J), W^{J-1} + O(\eta^J)) \\
&= W^1 + \beta(W^{J-1}, W^{J-1}) + O(\eta^{J+1}) = \psi(W^{J-1};\eta) + O(\eta^{J+1})
\end{aligned}
$$

so

$$\psi(W^J;\eta) - W^J = \psi(W^{J-1};\eta) - W^J + O(\eta^{J+1}) = O(\eta^{J+1})$$

follows from Lemma 3.2.

From (3.1) and (3.8) the components of $f(W^J)$ are of the form

$$f_i(W^J) = \sum_{j=1}^{2J} \eta^j \sum_{k=1}^{j} \xi^k f_{ijk}^J,$$

where for $1 \le i \le 3$, $1 \le j \le 2J$ and $1 \le k \le j$, f_{ijk}^J is given by the right sides of (3.10) but with $1 \le j', j'', k', k'' \le J$ in the summations. Using (3.2)–(3.7) and $\psi(W^J;\eta) = W^J + O(\eta^{J+1})$,

$$\begin{aligned}
\psi_i(W^J;\eta) &= \sum_{j=1}^{2J} \eta^j \int_{-\infty}^{t} e^{\lambda_i(t-\tau)} \sum_{k=1}^{j} e^{k\lambda_3\tau} f_{ijk}^J d\tau \\
&= W_i^J + \sum_{j=J+1}^{2J} \eta^j \sum_{k=1}^{j} e^{k\lambda_3 t} \frac{f_{ijk}^J}{k\lambda_3 - \lambda_i}, \qquad i = 1,2, \\
\psi_3(W^J;\eta) &= -\sum_{j=1}^{2J} \eta^j \int_{t}^{0} e^{\lambda_3(t-\tau)} \sum_{k=1}^{j} e^{k\lambda_3\tau} f_{3jk}^J d\tau \\
&= W_3^j + \sum_{j=J+1}^{2J} \eta^j \sum_{k=1}^{j} (e^{k\lambda_3 t} - e^{\lambda_3 t}) \frac{f_{3jk}^J}{k\lambda_3 - \lambda_3}.
\end{aligned} \tag{3.13}$$

Setting $e^{\lambda_3 t} = \xi$ in (3.13) gives

$$\psi_i(W^J;\eta) = W_i^J + \sum_{j=J+1}^{2J} \eta^j \sum_{k=1}^{j} \xi^k \psi_{ijk}^J, \qquad 1 \le i \le 3, \tag{3.14}$$

where for $J+1 \le j \le 2J$,

$$\begin{aligned}
\psi_{ijk}^J &= \frac{f_{ijk}^J}{k\lambda_3 - \lambda_i}, \qquad 1 \le i \le 2, 1 \le k \le j, \\
\psi_{3j1}^J &= -\sum_{k=2}^{J} \frac{f_{3jk}^J}{k\lambda_3 - \lambda_3}, \\
\psi_{3jk}^J &= \frac{f_{3jk}^J}{k\lambda_3 - \lambda_3}, \qquad 2 \le k \le J. \qquad \square
\end{aligned} \tag{3.15}$$

Definition of norms. For $u \in R^3$, we write $|u| = |u_1| + |u_2| + |u_3|$, and for continuous, R^3-valued functions $u(t)$ on $(-\infty, 0]$ we write $\|u\| = \sup_{t \le 0} |u(t)|$.

LEMMA 3.4. *Fix $J \ge 1$. Let $\rho > 0$ be such that $0 < K(\rho) < \frac{1}{2}$, where $K(\rho) = 2\rho((c_3/|\lambda_1|) + (c_1/|\lambda_2|) + (c_1/\lambda_3))$. Choose $\eta_0 > 0$ sufficiently small so that for all $|\eta| \le \eta_0$, the conditions $\|w^0\| \le \frac{\rho}{2}$ and $\|\psi(w^0;\eta) - w^0\| \le \frac{\rho}{4}$ are satisfied, where $w^0 = W^J(t,\eta)$. Then the sequence $w^n = \psi(w^{n-1};\eta)$, $n = 1,2,\ldots$, converges uniformly in t and η to a continuous, bounded solution $w(t)$ of the integral equation $w = \psi(w;\eta)$. Moreover, for each $t \le 0$ and $|\eta| \le \eta_0$,*

$$|w(t) - w^0(t)| \le (1 - K(\rho))^{-1} \sum_{i=1}^{3} \sum_{j=J+1}^{2J} |\eta|^j \sum_{k=1}^{j} \xi^k |\psi_{ijk}^J|, \tag{3.16}$$

where $\xi = e^{\lambda_3 t}$ *and the coefficients* ψ^J_{ijk} *are given by* (3.15).

Proof. If $u, v \in R^3$ obey $|u| \le \rho$ and $|v| \le \rho$, then since $|c_2| < c_3$,

$$\begin{aligned}|f_1(u) - f_1(v)| &= |(u_2+u_3)(c_2u_2+c_3u_3-c_2v_2-c_3v_3)+(u_2+u_3-v_2-v_3)(c_2v_2+c_3v_3)| \\ &\le |u|c_3|u-v| + |u-v|c_3|v| \le 2c_3\rho|u-v|,\end{aligned}$$

and

$$\begin{aligned}|f_2(u) - f_2(v)| = |f_3(u) - f_3(v)| &\le c_1|u_1(u_2+u_3-v_2-v_3)+(u_1-v_1)(v_2+v_3)| \\ &\le c_1(|u_1||u-v| + |u-v||v|) \le 2c_1\rho|u-v|.\end{aligned}$$

Using these inequalities (3.2) and (3.4), if $u(t)$ and $v(t)$ are continuous R^3-valued functions on $(-\infty, 0]$ and obey $\|u\| \le \rho$, $\|v\| \le \rho$, then for each $t \le 0$

$$\begin{aligned}|\varphi_1(u) - \varphi_1(v)| &\le \frac{2c_3\rho}{|\lambda_1|}\|u-v\|, \\ |\varphi_2(u) - \varphi_2(v)| &\le \frac{2c_1\rho}{|\lambda_2|}\|u-v\|, \\ |\varphi_3(u) - \varphi_3(v)| &\le \frac{2c_1\rho}{|\lambda_3|}\|u-v\|,\end{aligned}$$

so

$$\|\varphi(u) - \varphi(v)\| \le K(\rho)\|u-v\|.$$

By hypothesis $\|w^0\| \le \frac{\rho}{2}$ and $\|w^1 - w^0\| \le \frac{\rho}{4}$, so $\|w^1\| \le \|w^0\| + \|w^1 - w^0\| \le \frac{3\rho}{4}$. Then $\|w^2 - w^1\| \le K(\rho)\|w^1 - w^0\| < \frac{\rho}{8}$ which implies $\|w^2\| \le \|w^1\| + \frac{\rho}{8} \le \frac{7\rho}{8}$.

Assume for induction that $\|w^k\| \le \rho(1 - 2^{-(k+1)})$ for $0 \le k \le n$. Then

$$\begin{aligned}\|w^{n+1} - w^n\| = \|\varphi(w^n) - \varphi(w^{n-1})\| &\le K(\rho)\|w^n - w^{n-1}\| \\ &\le K(\rho)^n\|w^1(t) - w^0(t)\| \le 2^{-(n+2)}\rho,\end{aligned}$$

and so $\|w^{n+1}\| \le \rho(1 - 2^{-(n+2)})$ completing the induction. Therefore $\|w^n\| < \rho$ for all n. For any pair $m \ge n$,

$$\|w^m - w^n\| \le \sum_{k=n}^{m-1} \|w^{k+1} - w^k\| \le \sum_{k=n}^{m-1} 2^{-(k+2)}\rho < 2^{-(n+1)}\rho,$$

so the sequence is Cauchy uniformly for $t \le 0$ and $|\eta| \le \eta_0$. Define w as the limit of the sequence. Then $w(t)$ is bounded, continuous in t and η, and

$$\|w^{n+1} - w^0\| \le \sum_{k=0}^{n} \|w^{k+1} - w^k\| \le \sum_{k=0}^{n} K(\rho)^k\|w^1 - w^0\| < (1 - K(\rho))^{-1}\|w^1 - w^0\|$$

so

$$\|w - w^0\| \le (1 - K(\rho))^{-1}\|w^1 - w^0\| \le \frac{\rho}{2}. \tag{3.17}$$

Since $w^{n+1} = \psi(w^n; \rho)$, $\|w^n\| \le \rho$ and $\|w\| \le \|w^0\| + \|w - w^0\| \le \rho$,

$$\|w - \psi(w;\eta)\| \le \|w - w^{n+1}\| + \|\psi(w^n;\eta) - \psi(w)\| \le \|w - w^{n+1}\| + K(\rho)\|w^n - w\|$$

from which $w = \psi(w;\eta)$. Immediately, w solves (3.1), $p = \Lambda w$ solves (1.1)–(1.3) and

$$\lim_{t\to-\infty} p(t) = V(\lim_{t\to-\infty} w(t)) = (0,0,0)^T.$$

Now from (3.14),

$$w_i^1 - w_i^0 = \psi_i(W^J;\eta) - W_i^J = \sum_{j=J+1}^{2J} \eta^j \sum_{k=1}^{j} \xi^k \psi_{ijk}^J, \qquad 1 \le i \le 3, \tag{3.18}$$

where $\xi = e^{\lambda_3 t}$ and the coefficients ψ_{ijk} are given by (3.15). The bound

$$|w - W^J| \le (1 - K(\rho))^{-1} \sum_{i=1}^{3} \sum_{j=J+1}^{2J} |\eta|^j \sum_{k=1}^{j} \xi^k |\psi_{ijk}^J| \tag{3.19}$$

then follows using (3.17). □

To apply Lemma 3.4, J is chosen and the coefficients W_{ijk}, $1 \le i \le 3$, $1 \le j \le J$, $1 \le k \le j$ and ψ_{ijk}^J, $1 \le i \le 3$, $J+1 \le j \le 2J$, $1 \le k \le j$ are computed. Then η is chosen, and the expressions

$$B_0 = \sum_{i=1}^{3} \sum_{j=1}^{J} |\eta|^j \sum_{k=1}^{j} |W_{ijk}|,$$

$$B_1 = \sum_{i=1}^{3} \sum_{j=J+1}^{2J} |\eta|^j \sum_{k=1}^{j} |\psi_{ijk}^J|$$

are computed. These bound $\|w^0\|$ and $\|w^1 - w^0\|$, respectively. Then for $\rho = \max(2B_0, 4B_1)$, $K(\rho)$ is evaluated. If $K(\rho) \ge \frac{1}{2}$, it is necessary to choose a smaller η. Assuming $K(\rho) < \frac{1}{2}$, the expressions

$$W_i^J(0) = \sum_{j=1}^{J} \eta^j \sum_{k=1}^{j} W_{ijk}, \qquad 1 \le i \le 3$$

are evaluated, and then

$$w_i(0) \in \bar{w}_i = \left[W_i^J(0) - \frac{B_1}{1-K(\rho)}, W_i^J(0) + \frac{B_1}{1-K(\rho)} \right], \qquad 1 \le i \le 3.$$

For example, if $s = 10$, $R = 15$, and $q = \frac{8}{3}$, one finds that $\lambda_1 = -2.666667$, $\lambda_2 = -18.548$, $\lambda_3 = 7.548$, $c_1 = 0.3832, c_2 = -0.8548$, and $c_3 = 1.7548$, the numeric values in this example being precise in the sense that the absolute errors do not exceed $\frac{1}{2}$ unit in the last decimal place shown. Then for $J = 2$,

$$w^0 = (0.0988\eta^2\xi^2, 0, \eta\xi)^T$$

$$\psi(w^0;\eta) = w^0 + (0, 0.000919\eta^3\xi^3, 0.00251\eta^3(1-\xi^3))^T.$$

Choosing $\eta = \frac{1}{10}$, we find that $\|w^0\| \le B_0 = 0.10098815$ and $\|w^1 - w^0\| \le B_1 = 0.0000059367$. Then for $\rho = \max(2B_0, 4B_1) = 0.2019763$, $K(\rho) = 0.2947 < \frac{1}{2}$, so Lemma 3.4 applies. Now $w^0(0) = (0.000988, 0, 0.1)^T$, and $w(0)$ obeys $|w(0) - w^0(0)| \le B_1/(1 - K(\rho)) \le 0.0000084171$, from which $x(0) = w_2(0) + w_3(0) \in \bar{x} = [0.0999, 0.1001]$, $y(0) = c_2 w_2(0) + c_3 w_3(0) \in \bar{y} = [0.1754, 0.1756]$ and $z(0) = w_1(0) \in \bar{z} = [0.0009, 0.0011]$.

Lemma 3.4 shows existence of solutions $p(t) = (x, y, z)^T = \Lambda(R)w$ of (1.1)–(1.3), defined on $(-\infty, 0]$ and such that $\lim_{t\to-\infty} p(t) = 0$. Then, thanks to the invariant set V, the solutions $p(t)$ may be continued to $(-\infty, \infty)$. To have Property P or Q, the solutions must also satisfy $x' > 0$ for $t < \tau_1$, where $x'(\tau_1) = 0$. By choosing η sufficiently small positive, we can arrange that $x' > 0$ on $(-\infty, 0]$ as in the following lemma, which implies that any τ_1 such that $x'(\tau_1) = 0$ must be positive.

LEMMA 3.5. *Let* $p(t) = (x(t), y(t), z(t))^T$, $t \in (-\infty, 0]$ *be a solution of* (1.1)–(1.3) *such that* $0 < x(0) < y(0)$, $z(t) < R - 1$ *on* $(-\infty, 0]$ *and* $x(t) \to 0$ *as* $t \to -\infty$. *Then* $x' > 0$ *on* $(-\infty, 0]$.

Proof. From (1.1), $x'(0) = s(y(0) - x(0)) > 0$, so $x' > 0$ for $t < 0$ near 0.

Suppose for contradiction that there exist point(s) $\tau < 0$ at which $x'(\tau) = 0$. At any such point, $x(\tau) \neq 0$, for if $x(\tau) = 0$ the solution of (1.1) and (1.2) would be just $x(t) = y(t) = 0$ for all t and inconsistent with $x(0) < y(0)$. Therefore $x''(\tau) = sy'(\tau) = sx(\tau)(R - 1 - z(\tau)) \neq 0$, $\operatorname{sgn}(x(\tau)) = \operatorname{sgn}(x''(\tau)) \neq 0$, and x' changes sign at τ. Let

$$\tau^* = \sup\{\tau < 0 | x'(\tau) = 0\}.$$

Then x' changes sign at τ^* and is of one sign (positive) on $(\tau^*, 0)$. Since $x''(\tau^*) \neq 0$ and $x'(t) > 0$ for $t > \tau^*$ near τ^*, necessarily $x''(\tau^*) > 0$ and so $x(\tau^*) > 0$. For $t < \tau^*$ near τ^*, $x'(t) < 0$. Now $x'(t)$ must change sign on $(-\infty, \tau^*)$, for otherwise $x(t)$ would be monotone decreasing on $(-\infty, \tau^*)$, inconsistent with $\lim_{t\to-\infty} x(t) = 0$ and $x(\tau^*) > 0$. Let

$$\tau^{**} = \sup\{t < \tau^*, x'(t) = 0\}.$$

Then x' is of one sign (negative) on (τ^{**}, τ^*) and changes sign at τ^{**}. Necessarily $x''(\tau^{**}) < 0$, which implies $x(\tau^{**}) < 0$. This is a contradiction, since one cannot have $x(\tau^{**}) < 0, x(\tau^*) > 0$, and $x'(t) < 0$ on (τ^{**}, τ^*).

We have shown that x' has no zeros on $(-\infty, 0]$. Since $x'(0) > 0$, it follows that $x' > 0$ on $(-\infty, 0]$. □

4. The solutions for $t > 0$. In [1], Aberth describes an algorithm for computing the solutions of initial value problems "precisely." The algorithm is a Taylor series method using interval arithmetic, and applies to ordinary differential systems in which the functions defining the derivatives are elementary. In [1], algorithms and programs are given for first- and second-order systems. The generalization to systems of n first-order equations is more recent and is coded as program difsys.cc; see the discussion. Modifications of difsys were required to obtain the information needed for present purposes. The basic task in difsys to advance the solution to $t_{k'} = t_k + h$ and determine bounds for $x(t_{k'}), y(t_{k'}), z(t_{k'})$. Before a trial interval $[t_k, t_k + h]$ is accepted as a "bounding interval," an associated containment region is constructed. The first two modifications we made to the algorithm were as follows:

M_1: reject a trial interval unless the containment region guarantees
at least one of x, x' is of one sign on the interval, and
at least one of y, y' is of one sign on the interval, and

at least one of z, z' is of one sign on the interval.

M_2: reject a trial interval unless the containment region guarantees
at least one of x', x'' is of one sign on the interval, and
at least one of y', y'' is of one sign on the interval, and
at least one of z', z'' is of one sign on the interval.

We also added code to compute intervals $I_k^0 \supset x(t_k)x(t_{k'})$, $J_k^0 \supset y(t_k)y(t_{k'})$, $K_k^0 \supset z(t_k)z(t_{k'})$, $I_k^1 \supset x'(t_k)x'(t_{k'})$, $J_k^1 \supset y'(t_k)y'(t_{k'})$,and $K_k^1 \supset z'(t_k)z'(t_{k'})$. The last two modifications are

M_3: if $I_k^0 < 0$, print a message "x changes sign,"
if $J_k^0 < 0$, print a message "y changes sign,"
if $K_k^0 < 0$, print a message "z changes sign,"
if I_k^0, J_k^0 or K_k^0 overlaps 0, print a message that this happens.

M_4: if $I_k^1 < 0$, print a message "x' changes sign,"
if $J_k^1 < 0$, print a message "y' changes sign,"
if $K_k^1 < 0$, print a message "z' changes sign,"
if I_k^1, J_k^1 or K_k^1 overlaps 0, print a message that this happens.

Here $I_k^0 > 0$ (respectively, < 0) means every point in I_k^0 is positive (respectively, negative), and "I_k^0 overlaps zero" means $0 \in I_k^0$.

Suppose that the modified code has accepted a bounding interval $[t_k, t_{k'}]$ and has successfully computed an approximate solution on this interval, and no messages from M_3 appear. Then we claim that $x(t)$ is either strictly positive or strictly negative on $[t_k, t_{k'}]$. For, as M_3 issued no message, neither $I_k^0 < 0$ nor I_k^0 overlaps zero, so $I_k^0 > 0$. Therefore, $x(t_k)x(t_{k'}) > 0$ and x does not vanish at either endpoint. Since the interval $[t_k, t_{k'}]$ passed the test M_1, the containment region guarantees that at least one of x, x' is of one sign on the interval. If x is of one sign, then $\operatorname{sgn}(x(t)) = \operatorname{sgn}(x(t_k)) \neq 0$ for all $t \in [t_k, t_{k'}]$ and we are done. If x' is of one sign, so then $x(t)$ is monotonic (increasing or decreasing) from $x(t_k)$ to $x(t_{k'})$, where $x(t_k)$ and $x(t_{k'})$ are of the same sign, and so again $\operatorname{sgn}(x(t)) = \operatorname{sgn}(x(t_k))$ for all $t \in [t_k, t_{k'}]$. Similarly if no messages from M_3 appear, $y(t)$ and $z(t)$ are strictly positive or negative on $[t_k, t_{k'}]$.

If the message "x changes sign" appears, then $x(t_k)x(t_{k'}) < 0$ so $x(t_k)$ and $x(t_{k'})$ are of opposite signs and $x(t)$ has at least one zero in the interval. In this case, since the interval passed the test M_1, necessarily x' is of one sign on the interval and it follows that $x(t)$ changes sign at exactly one point, an interior point, of the interval. Messages "y changes sign" and "z changes sign" are similarly interpreted. If no messages from M_4 appear for a bounding interval $[t_k, t_{k'}]$, by a similar argument $x'(t), y'(t)$ and $z'(t)$ are each of one sign on the interval. If a message "x' changes sign" appears, then x' is strictly monotonic on the interval and has a zero at an interior point. Messages "y' changes sign" and "z' changes sign" are interpreted similarly. If any messages that $I_k^0, J_k^0, K_k^0, I_k^1, J_k^1$ or K_k^1 overlap 0 should appear, it is necessary to repeat the computation at a higher precision.

We now state the numerical results as follows.

LEMMA 4.1. *For each set of values s, q, R in Table 1, the Lorenz system has an orbit $p(t)$ with the Property Q, and for each set of values s, q, R in Table 2, the Lorenz system has an orbit $p(t)$ with Property P.*

TABLE 1

Orbits with Property Q.

$R = 13.926$ (exact)
$s = 10$ (exact)
$q = \frac{8}{3}$
$J = 25$
$\eta = 0.1$ (exact)
$B_0 = 0.10101874785080090852561365779491980762459271357300$
$B_1 = 2.40820490426304850175700 0E - 51$
$x(0)$ in $0.10000101387916306034307402007221470577302262 \sim$
$y(0)$ in $0.17129644412294703784847058977871 0542299221 \sim$
$z(0)$ in $0.00101205806115328023910366159554912883232786 \sim$
$n = 40$
y' changes sign on the interval 0.6938476563 to 0.6943359375
x' changes sign on the interval 0.7558593750 to 0.7578125000
z' changes sign on the interval 0.8242187500 to 0.8251953125
y changes sign on the interval 0.9521484375 to 0.9560546875
y' changes sign on the interval 1.0224609375 to 1.0263671875
y changes sign on the interval 1.4863281250 to 1.4873046875
x' changes sign on the interval 1.5744628906 to 1.5747070313
$t_f = 1.6$ (exact)

$R = 13.9265$ (exact)
$s = 10$ (exact)
$q = \frac{8}{3}$
$J = 25$
$\eta = 0.1$ (exact)
$B_0 = 0.10101873551056010840559137742027929504345305451900$
$B_1 = 2.40659919095945520117401190 00E - 51$
$x(0)$ in $0.10000101383096035494527865462507460119921357739 4 \sim$
$y(0)$ in $0.17129842358345975564593357177092041637153684 \sim$
$z(0)$ in $0.00101204608358902851663018738676 1869868147368 \sim$
$n = 25$
y' changes sign on the interval 0.6938476563 to 0.6943359375
x' changes sign on the interval 0.7558593750 to 0.7578125000
z' changes sign on the interval 0.8242187500 to 0.8251953125
y changes sign on the interval 0.9521484375 to 0.9560546875
y' changes sign on the interval 1.0224609375 to 1.0263671875
y changes sign on the interval 1.5902099609 to 1.5903320313
x' changes sign on the interval 1.6719360352 to 1.6719970703
$t_f = 1.72$ (exact)

Proof. For each triple (s, q, R) shown in Tables 1 and 2,[1] we chose J and and computed intervals $\bar{x}(0)$, $\bar{y}(0)$, $\bar{z}(0)$ guaranteed by Lemma 3.4 to contain a point $x(0)$, $y(0)$, $z(0)$ of a trajectory of the Lorenz system such that $x(t) \to 0$ as $t \to -\infty$. This was done with a short program in C++ using the Aberth–Schaefer interval arithmetic package. Input data consists of the parameters q, R, s, the integer J, $\eta = w_3(0) > 0$, and an integer determining the precision of the arithmetic. Output consists of the information, whether or not η was chosen sufficiently small for the hypotheses of

[1]The notation "$\sim$" in the tables indicates intervals having midpoint of the given decimal number and half-width one unit of the last decimal place. The absolute error in decimal values given does not exceed $\frac{1}{2}$ unit in the last decimal place.

TABLE 2

Orbits with Property P.

$R = 13.93$ (exact)
$s = 10$ (exact)
$q = \frac{8}{3}$
$J = 25$
$\eta = 0.1$ (exact)
$B_0 = 0.10101864915078526087165201466934523185017094722900$
$B_1 = 2.39539098865440343175485240000E - 51$
$x(0)$ in $0.10000101349365681675988598297465044410099459892 1 \sim$
$y(0)$ in $0.17131227893846698671597596429068023867274433 \sim$
$z(0)$ in $0.00101196226158061216784055167835988349168981 \sim$
$n = 25$
y' changes sign on the interval 0.6938476563 to 0.6943359375
x' changes sign on the interval 0.7558593750 to 0.7578125000
z' changes sign on the interval 0.8242187500 to 0.8251953125
y changes sign on the interval 0.9521484375 to 0.9560546875
y' changes sign on the interval 1.0224609375 to 1.0263671875
x changes sign on the interval 1.4384765625 to 1.4404296875
y' changes sign on the interval 1.4511718750 to 1.4521484375
$t_f = 1.46$ (exact)

$R = 13.927$ (exact)
$s = 10$ (exact)
$q = \frac{8}{3}$
$J = 25$
$\eta = 0.1$ (exact)
$B_0 = 0.10101872317110199349682497404003229580910620678400$
$B_1 = 2.40499461474461557372980370000E - 51$
$x(0)$ in $0.10000101378276177196766847317506571067572971416 2 \sim$
$y(0)$ in $0.17130040301294877808616205243979664937742171 \sim$
$z(0)$ in $0.00101203410677283095541051420081399240699564 1 \sim$
$n = 25$
y' changes sign on the interval 0.6938476563 to 0.6943359375
x' changes sign on the interval 0.7558593750 to 0.7578125000
z' changes sign on the interval 0.8242187500 to 0.8251953125
y changes sign on the interval 0.9521484375 to 0.9560546875
y' changes sign on the interval 1.0224609375 to 1.0263671875
x changes sign on the interval 1.5346679688 to 1.5351562500
$t_f = 1.54$ (exact)

Lemma 3.4 to be satisfied, and if they are, the initial condition intervals $\bar{x}(0), \bar{y}(0), \bar{z}(0)$ and bounds B_0, B_1 on $\|w^0\|$ and $\|w^1 - w^0\|$. All this data is shown in the tables. Examining $\bar{x}(0)$ and $\bar{y}(0)$, since $x(0) \in \bar{x}(0)$ and $y(0) \in \bar{y}(0)$ it follows that $0 < x(0) < y(0)$. Since $|z(t)| = |w_1(t)|$, it follows that $|z(t)| < \|p(t)\| \leq \rho = \max(2B_0, 4B_1) = 2B_0 < R - 1$ for all $t < 0$. From Lemma 3.5 we conclude that $x'(t) > 0$ on $(-\infty, 0]$, which implies $x(t) > 0$ on $(-\infty, 0]$.

Using Aberth's code difsys with modifications M_1 through M_4 described above, we computed trajectories of the Lorenz system for $t > 0$, taking $\bar{x}(0), \bar{y}(0), \bar{z}(0)$ as interval-valued initial conditions. The integrations were carried out until either

(1) the sign change message showed that the orbit had Property P;

(2) the sign change message showed that the orbit had Property Q;

(3) the computation failed due to determine whether the orbit had Property P or Q.

Data corresponding to successful computations was then entered in the Tables 1 and 2. □

We now state our result concerning a homoclinic orbit for (1.1).

THEOREM 4.2. *For* $s = 10$, $q = \frac{8}{3}$ *and some* R^*, $R_a < R_* < R_b$, *where* $R_a = 13.9265$ *is the maximum of the* R *values in Table* 1 *and* $R_b = 13.927$ *is the minimum of the values in Table* 2, *the Lorenz system has an orbit* $p(t)$ *such that* $\lim_{t\to-\infty} p(t) = \lim_{t\to\infty} p(t) = (0,0,0)^T$, *and there are numbers* $\tau_1 < s_1$ *such that*

$$
\begin{aligned}
&x' > 0 \quad \text{on } (-\infty, \tau_1),\\
&y(\tau_1) < 0,\\
&x' < 0 \quad \text{on } (\tau_1, \infty),\\
&y' < 0 \quad \text{on } [\tau_1, s_1],
\end{aligned}
$$

$$
\begin{aligned}
&y(s_1) = 0,\\
&y < 0 \quad \text{on } (s_1, \infty),\\
&x > 0 \quad \text{on } (-\infty, \infty).
\end{aligned}
$$

Proof. It is easily verified that $p^+(R)$ is linearly stable for all $R \in [R_a, R_b]$. The proof then follows immediately from Lemmas 2.2 and 4.1. □

5. Discussion. The present methods do not obtain an approximation to the homoclinic orbit valid on $(-\infty, \infty)$; to do this, one could approximate trajectories in the stable manifold much as in §3.

We expect that some details of the "homoclinic explosion" in the Lorenz system could be established rigorously by methods similar to those used here.

The present work is ostensibly numerical, while [2] and [3] are analytical. However, this distinction is blurry. Hastings and Troy use calculator computations, and allow margins in their bounds to account for possible roundoff errors. They approximate the unstable manifold, and shoot, generating bounds on the trajectories. Here we do the same, except that we use an interval-arithmetic package and methods of precise numerical analysis to obtain the bounds automatically.

The condition in Lemma 2.2 that $p^+(R)$ is linearly stable could be dropped if we adopted part of the argument in [2,3]. We retain the condition because it simplifies the argument and holds for the present parameter values.

Together with Hastings and Troy [4] we have recently demonstrated the existence of an infinite set of solutions of (1.1) with certain properties of chaos, for $R = 76$, $s = 10$, $q = 9$. The verification of condition A in [4] used the present methods.

Aberth has made the Aberth–Schaefer interval arithmetic package, including the code difsys.cc, publicly available by anonymous ftp from math.tamu.edu, as file pub/range/range.tar.Z.

Acknowledgments. We wish to thank S. P. Hastings for suggesting the problem, O. Aberth for his most useful interval analysis codes, and both for helpful discussions.

REFERENCES

[1] O. ABERTH, *Precise Numerical Analysis*, William C. Brown Publishers, Dubuque, Iowa, 1988.
[2] S. HASTINGS AND W. TROY, *A shooting approach to the Lorenz equations*, Bull. Amer. Math. Soc, 27 (1992), pp. 128–131.
[3] ———, *A proof that the Lorenz equations have a homoclinic orbit*, J. Diff. Eqns., (1993), to appear.
[4] B. HASSARD, S. P. HASTINGS, W. C. TROY, AND J. ZHANG, *A computer proof that the Lorenz equations have "chaotic" solutions*, Appl. Math. Lett., (1993), to appear.
[5] E. N. LORENZ, *Deterministic non-periodic flow*, J. Atmospheric Sci., 20 (1963), pp. 130–141.
[6] C. SPARROW, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, Appl. Math. Sci. 41, Springer-Verlag, New York, 1982.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 197–217, January 1994

008

THE ADDITION FORMULA FOR CONTINUOUS q-LEGENDRE POLYNOMIALS AND ASSOCIATED SPHERICAL ELEMENTS ON THE $SU(2)$ QUANTUM GROUP RELATED TO ASKEY–WILSON POLYNOMIALS*

H. T. KOELINK†

Abstract. The known interpretation of a two-parameter family of Askey–Wilson polynomials as spherical elements on the $SU(2)$ quantum group is extended to an interpretation of a three-parameter (one discrete parameter) family of Askey–Wilson polynomials as associated spherical elements on the $SU(2)$ quantum group. An abstract addition formula, i.e., an expression involving noncommuting variables, for the two-parameter class of Askey–Wilson polynomials is obtained by this interpretation. Specialization gives an abstract addition formula for the continuous q-Legendre polynomials from which the ordinary Rahman–Verma addition formula for continuous q-Legendre polynomials is obtained.

Key words. quantum group, $SU(2)$, associated spherical elements, Askey–Wilson polynomials, continuous q-Legendre polynomials, addition formula

AMS subject classifications. 33A65, 33A75, 22E70

1. Introduction.

The interpretation of special functions of q-hypergeometric type on quantum groups, cf. [8] for a survey, started with the interpretation of the little q-Jacobi polynomials as matrix elements of irreducible unitary representations of the quantum group $SU_q(2)$; see Vaksman and Soibelman [15], Masuda et al. [11], [12], and Koornwinder [7]. The Schur orthogonality relations for the quantum group $SU_q(2)$ are equivalent to the orthogonality relations for the little q-Jacobi polynomials. Koornwinder [9] then used this quantum group interpretation to obtain an addition formula for the little q-Legendre polynomials which was not known until then. Later Koornwinder [8], [10] gave an interpretation of a two-parameter family of Askey–Wilson polynomials on the quantum group $SU_q(2)$. For a suitable choice of parameters we get a quantum group interpretation of the continuous q-Legendre polynomials for which an addition formula has been proved analytically by Rahman and Verma [14]. The main goal of this paper is to show that this addition formula can also be proved by use of the quantum group $SU_q(2)$.

This paper is closely related to Koornwinder's paper [10], and we assume the reader is familiar with §§2–5 of [10]. In §2 we start with the introduction of so-called associated (σ,τ)-spherical elements in the quantized algebra $\mathcal{A}_q$ of polynomials on $SU(2)$. These elements are invariant under the left action of some fixed element of the quantized universal enveloping algebra $\mathcal{U}_q$, and they transform in some nice ways under the right action of another fixed element of $\mathcal{U}_q$. It is proved that an associated (σ,τ)-spherical element multiplied in $\mathcal{A}_q$ from the right by a (σ,τ)-spherical element, cf. [8] and [10], is again an associated (σ,τ)-spherical element of the same sort. Then we can write an associated (σ,τ)-spherical element as some sort of minimal associated (σ,τ)-spherical element times a polynomial in $\rho_{\sigma,\tau}$, the generator of the algebra of (σ,τ)-spherical elements on $SU_q(2)$. For some suitable choice of associated (σ,τ)-spherical elements these polynomials yield systems of orthogonal polynomials.

*Received by the editors August 8, 1990; accepted for publication February 10, 1993.

†Departement of Mathematics, Catholic University of Leuven, Celestijnenlaan 200B, B-3001 Heverlee, Belgium.

In §3 we determine these polynomials, which turn out to be Askey–Wilson polynomials of two continuous and one discrete parameter. This interpretation of Askey–Wilson polynomials on $SU_q(2)$ is the second main result of this paper. The Haar functional and the Schur orthogonality relations are used to identify the polynomials mentioned in the previous paragraph as these Askey–Wilson polynomials. In this section we also show how to obtain an abstract addition formula, i.e., an identity involving noncommuting variables for a two-parameter class of Askey–Wilson polynomials by application of the comultiplication on such an Askey–Wilson polynomial in $\rho_{\sigma,\tau}$. Actually, we obtain an expansion in terms of associated (σ,τ)-spherical elements and thus in terms of Askey–Wilson polynomials of two continuous and one discrete parameter.

The difficult part is to obtain an addition formula in commuting variables from this abstract addition formula, and until now we can only do this for the special case of the continuous q-Legendre polynomial. In §4 we show how it can be done, by exploiting the fact that the abstract addition formula is essentially a development in associated (σ,τ)-spherical elements. Apart from this, only one-dimensional representations of the algebra $\mathcal{A}_q$ are used, which contrast with the quantum group theoretic proof of the addition formula for the little q-Legendre polynomials (cf. [9]), where only infinite-dimensional representations of the algebra $\mathcal{A}_q$ are used. For the continuous q-Legendre polynomials the addition formula has already been proved analytically by Rahman and Verma [14], where it occurs as a special case of the addition formula for the continuous q-ultraspherical polynomials. Section 5 is devoted to the limit case $q \uparrow 1$ of the proof presented in §4. The proof reduces to evaluation of functions on $SU(2)\times SU(2)$, which is used in [16, Chap. 3.4] to prove the addition formula for Legendre polynomials from a group theoretic point of view.

In this paper a lot of constants have to be calculated, and the reader is urged to skip these straightforward, but sometimes tedious calculations at first reading.

The results contained in §§2 and 3 have also been obtained by Noumi and Mimachi in a slightly different but more general setting. At the end of each section we compare our work with the results of their announcement [13].

Notation: $\mathbf{Z}_+$ denotes the nonnegative integers $\{0,1,2,3,\ldots\}$.

2. Associated (σ,τ)-spherical elements on $SU_q(2)$. In this section associated (σ,τ)-spherical elements on $SU_q(2)$ are defined. These elements can be expressed explicitly in terms of the matrix elements $t^l_{m,n}$ of the irreducible unitary representations of $SU_q(2)$ using dual q-Krawtchouk polynomials. Next we prove that all associated (σ,τ)-spherical elements can be expressed as some minimal associated (σ,τ)-element times a polynomial in $\rho_{\sigma,\tau}$. These polynomials form a system of orthogonal polynomials with respect to some moment functional.

Recall from [10, §4] the definitions

$$X_\sigma = iq^{\frac{1}{2}}B - iq^{-\frac{1}{2}}C - \frac{q^{-\sigma}-q^{\sigma}}{q^{-1}-q}(A-D) \in \mathcal{U}_q, \qquad \sigma \in \mathbf{R},$$

and

$$\begin{aligned}\rho_{\sigma,\tau} =& \tfrac{1}{2}\big(\alpha^2+\delta^2+q\gamma^2+q^{-1}\beta^2+i(q^{-\sigma}-q^{\sigma})(q\delta\gamma+\beta\alpha)\\ &-i(q^{-\tau}-q^{\tau})(\delta\beta+q\gamma\alpha)+(q^{-\sigma}-q^{\sigma})(q^{-\tau}-q^{\tau})\beta\gamma\big),\end{aligned}$$

as well as that for all polynomials p we have

$$X_\sigma.p(\rho_{\sigma,\tau}) = 0 = p(\rho_{\sigma,\tau}).X_\tau.$$

DEFINITION 2.1. *An element $b \in \mathcal{A}_q$ is an associated (σ,τ)-spherical element if there exists $\lambda \in \mathbf{C}$ so that*

$$X_\sigma.b = 0, \qquad b.X_\tau = \lambda b.D. \tag{2.1}$$

For $\lambda = 0$ we obtain the (σ,τ)-spherical elements as defined by Koornwinder [8], and we will first investigate the relation between (σ,τ)-spherical elements and associated (σ,τ)-spherical elements.

PROPOSITION 2.2. *Suppose $b \in \mathcal{A}_q$ satisfies (2.1), and let $a \in \mathcal{A}_q$ be a (σ,τ)-spherical element. Then ba satisfies (2.1) with the same λ.*

Proof. Because of [10, eq. (3.10), Lem. 3.1] and (2.1) we find

$$X_\sigma.(ba) = (A.b)(X_\sigma.a) + (X_\sigma.b)(D.a) = 0.$$

Similarly we find

$$(ba).X_\tau = (b.A)(a.X_\tau) + (b.X_\tau)(a.D) = \lambda(b.D)(a.D),$$

and this equals $\lambda(ba).D$ because of [10, eq. (3.10)] and $\Delta(D) = D \otimes D$. □

PROPOSITION 2.3. *If $b \in \mathcal{A}_q$ is an associated (σ,τ)-spherical element with $\lambda \in \mathbf{R}$, then b^*b is a (σ,τ)-spherical element.*

Proof. To prove this proposition we will first consider $X_\sigma.b^*$ and $b^*.X_\tau$. To do this we need the duality between the Hopf $*$-algebras $\mathcal{A}_q$ and $\mathcal{U}_q$. In particular, we have (cf. [10, eq. (3.6)])

$$\langle X, a^* \rangle = \overline{\langle S(X)^*, a \rangle} \quad \forall\, X \in \mathcal{U}_q \quad \forall\, a \in \mathcal{A}_q. \tag{2.2}$$

It easily follows from [10, eqs. (3.3) and (3.4)] that $S(X_\sigma)^* = -X_\sigma$ and $S(A)^* = D$. If we write

$$\Delta(b) = \sum_{(b)} b_{(1)} \otimes b_{(2)},$$

then we have by [10, eq. (3.8)], (2.2), and (2.1),

$$\begin{aligned} X_\sigma.b^* &= \sum_{(b)} b^*_{(1)} \langle X_\sigma, b^*_{(2)} \rangle = -\sum_{(b)} b^*_{(1)} \overline{\langle X_\sigma, b_{(2)} \rangle} \\ &= -\left(\sum_{(b)} b_{(1)} \langle X_\sigma, b_{(2)} \rangle \right)^* = -(X_\sigma.b)^* = 0. \end{aligned} \tag{2.3}$$

Similarly we find

$$\begin{aligned} b^*.X_\tau &= -(b.X_\tau)^* = -\overline{\lambda}(b.D)^* = -\overline{\lambda} \sum_{(b)} \overline{\langle D, b_{(1)} \rangle} b^*_{(2)} \\ &= -\overline{\lambda} \sum_{(b)} \langle A, b^*_{(1)} \rangle b^*_{(2)} = -\overline{\lambda} b^*.A. \end{aligned} \tag{2.4}$$

Now we are able to prove the proposition. Using [10, eq. (3.10)], (2.1), and (2.3), we have

$$X_\sigma.(b^*b) = (A.b^*)(X_\sigma.b) + (X_\sigma.b^*)(D.b) = 0,$$

and similarly, using (2.4) instead of (2.3), we obtain

$$(b^*b).X_\tau = (b^*.A)(b.X_\tau) + (b^*.X_\tau)(b.D) = (\lambda - \overline{\lambda})(b^*.A)(b.D).$$

This yields zero for $\lambda \in \mathbf{R}$. □

In a moment we will see that there are no nonzero-associated (σ,τ)-spherical elements with nonreal λ in (2.1), so Proposition 2.3 applies to all associated (σ,τ)-spherical elements on $SU_q(2)$.

Now [10, eqs. (3.18) and (3.19)] shows that $X_\sigma.$ and $.X_\tau$ preserve the linear span of the matrix elements $t^l_{n,m}$, $n,m \in \{-l,\dots,l\}$, and hence it is sufficient to look for associated (σ,τ)-spherical elements b of the form

$$\sum_{n,m=-l}^{l} \gamma_{nm} t^l_{n,m}.$$

In this case (2.1) is equivalent to (cf. [10, Lem. 3.4])

$$t^l(X_\sigma)\sum_{m=-l}^{l}\gamma_{nm}e^l_m = 0 \quad \forall n, \qquad t^l(X_\tau A)\sum_{n=-l}^{l}\overline{\gamma_{nm}}e^l_n = \overline{\lambda}\sum_{n=-l}^{l}\overline{\gamma_{nm}}e^l_n \quad \forall m, \tag{2.5}$$

since $(X_\tau A)^* = X_\tau A$. From [10, Lem. 4.6] we know that $\ker(t^l(X_\sigma))$ has dimension zero if $l+\frac{1}{2} \in \mathbf{Z}_+$, while for $l \in \mathbf{Z}_+$ $t^l(X_\sigma)$ it has a one-dimensional null space spanned by $\sum_{m=-l}^{l} q^{-m/2}c^{l,\sigma}_m e^l_m$ with

$$c^{l,\sigma}_m = \frac{i^m q^{-(l+\sigma)m}q^{m^2/2}}{(q^2;q^2)^{1/2}_{l+m}(q^2;q^2)^{1/2}_{l-m}}\,{}_3\varphi_2\left(\begin{matrix} q^{-2l+2m}, q^{-2l}, -q^{-2l-2\sigma} \\ q^{-4l},\ 0 \end{matrix}; q^2, q^2\right). \tag{2.6}$$

The eigenvalues and eigenvectors of $t^l(X_\tau A)$, $l \in \mathbf{Z}_+$, have also been determined in [8] and [10, Thm. 4.3 and Lem. 4.4]. The spectrum of $t^l(X_\tau A)$, $l \in \mathbf{Z}_+$, is simple and consists of $2l+1$ points

$$\lambda_j = \frac{q^{-2j+\tau} - q^{2j-\tau} + q^{-\tau} - q^\tau}{q^{-1}-q}, \qquad j = -l,\dots,l. \tag{2.7}$$

So the eigenspaces of $t^l(X_\tau A)$ are one-dimensional, and the corresponding eigenvectors are given in terms of dual q-Krawtchouk polynomials (cf. [10, Thm. 4.3]),

$$t^l(X_\tau A)\sum_{n=-l}^{l} p_{l-n}(\lambda_j;l)e^l_n = \lambda_j \sum_{n=-l}^{l} p_{l-n}(\lambda_j;l)e^l_n,$$

with

$$p_n(\lambda_j;l) = i^{-n}q^{n\tau}q^{\frac{1}{2}n(n-1)}\left(\frac{(q^{4l};q^{-2})_n}{(q^2;q^2)_n}\right)^{\frac{1}{2}} R_n(q^{-2l-2j}-q^{2j-2l-2\tau};q^{2\tau},2l \mid q^2), \tag{2.8}$$

where R_n is defined in [10, eq. (2.17)]. Note that for $l \le m$ spectrum$(t^l(X_\tau A)) \subset$ spectrum$(t^m(X_\tau A))$.

These results prove the following proposition.

PROPOSITION 2.4. *For $l \in \frac{1}{2} + \mathbf{Z}_+$ the space of associated (σ,τ)-spherical elements in* span $\{t^l_{mn}\}$ *is zero. If $l \in \mathbf{Z}_+$, then λ must be of the form λ_j for $j \in \{-l,\dots,l\}$, and the space of associated (σ,τ)-spherical elements in* span$\{t^l_{nm}\}$ *corresponding to λ_j is one-dimensional and spanned by*

$$\sum_{n,m=-l}^{l} q^{-\frac{m}{2}} c^{l,\sigma}_m \overline{p_{l-n}(\lambda_j;l)} t^l_{n,m}.$$

Note that Proposition 2.4 implies that all possible λ's in (2.1) are real for nonzero-associated (σ,τ)-spherical elements; hence Proposition 2.3 applies to all associated (σ,τ)-spherical elements.

We will renormalize the element spanning the one-dimensional space of associated (σ,τ)-spherical elements in span$\{t^l_{nm}\}$ corresponding to λ_j by

$$b^l_j(\sigma,\tau) = \sum_{n,m=-l}^{l} q^{-\frac{m}{2}} c^{l,\sigma}_m a^{l,j}_n(\tau) t^l_{n,m}, \tag{2.9}$$

with $a^{l,j}_n(\tau) = C^{l,j}(\tau)\overline{p_{l-n}(\lambda_j;l)}$, so that

$$\sum_{n=-l}^{l} a^{l,j}_n(\tau)\overline{a^{l,i}_n(\tau)} = \delta_{ij} = \sum_{n=-l}^{l} a^{l,n}_j(\tau)\overline{a^{l,n}_i(\tau)}. \tag{2.10}$$

From [10, eq. (2.18)] it immediately follows that

(2.11)

$$\begin{aligned} C^{l,j}(\tau) =& q^{-(l+j)} \begin{bmatrix} 2l \\ l+j \end{bmatrix}_{q^2}^{\frac{1}{2}} \left(\frac{(-q^{4l+2\tau};q^{-2})_{l+j}}{(-q^{2-2\tau};q^2)_{l+j}}\right)^{\frac{1}{2}} \\ & \times \left(\frac{1+q^{4j-2\tau}}{1+q^{-4l-2\tau}}\right)^{\frac{1}{2}} - (q^{2\tau};q^2)_{2l}^{-\frac{1}{2}} \\ =& q^{l-j} \begin{bmatrix} 2l \\ l+j \end{bmatrix}_{q^2}^{\frac{1}{2}} \left(\frac{1+q^{4j-2\tau}}{1+q^{-2\tau}}\right)^{\frac{1}{2}} \left(\frac{1}{(-q^{2-2\tau};q^2)_{l+j}(-q^{2+2\tau};q^2)_{l-j}}\right)^{\frac{1}{2}}. \end{aligned}$$

The associated (σ,τ)-spherical element $b^l_j(\sigma,\tau)$, $j \geq 0$, can be related to $b^j_j(\sigma,\tau)$ in a very simple manner. To see this we start with an arbitrary polynomial s_r of degree r, and we consider $b^j_j(\sigma,\tau)s_r(\rho_{\sigma,\tau})$. Now [6, Thm. 3.4] or [18, Thm. 5.11] (cf. [10, Prop. 3.5]) implies that we can write

$$b^j_j(\sigma,\tau)s_r(\rho_{\sigma,\tau}) = \sum_{k=|j-r|}^{j+r} b_k, \tag{2.12}$$

where $b_k \in \text{span}\{t^k_{n,m} : n,m = -k,\dots,k\}$. However, by Proposition 2.2, this is an associated (σ,τ)-spherical element with λ replaced by λ_j in (2.1), and so every b_k in (2.12) must be an associated (σ,τ)-spherical element because $X_\sigma.$ and $.X_\tau$ preserve span$\{t^l_{n,m}\}$. Hence $b_k = 0$ for $k < j$ by Proposition 2.4. Rewriting (2.12) as

$$b^j_j(\sigma,\tau)s_r(\rho_{\sigma,\tau}) = \sum_{k=j}^{j+r} c_k b^k_j(\sigma,\tau),$$

we see that we are left with $r+1$ constants c_k depending on the polynomial s_r of degree r. Application of the one-dimensional representation $\pi^1_{\theta/2}$ shows that the mapping $s_r \mapsto b^j_j s_r(\rho_{\sigma,\tau})$ is injective.

PROPOSITION 2.5. *There exist two systems of orthogonal polynomials p_{l-j}, q_{l-j} of degree $l-j$ so that for $j \geq 0$ we have*

$$b^l_j(\sigma,\tau) = b^j_j(\sigma,\tau)p_{l-j}(\rho_{\sigma,\tau})$$

and

$$b^l_{-j}(\sigma,\tau) = b^j_{-j}(\sigma,\tau)q_{l-j}(\rho_{\sigma,\tau}).$$

Proof. The previous remarks prove that for some polynomial p_{l-j} of degree $l-j$ we have

$$b^l_j(\sigma,\tau) = b^j_j(\sigma,\tau)p_{l-j}(\rho_{\sigma,\tau}). \tag{2.13}$$

To prove that the polynomials p_n form a system of orthogonal polynomials we need the Haar functional h; cf. [17, §4]. For the quantum group $SU_q(2)$ the Schur orthogonality relations [17, Thm. 5.7] have been calculated explicitly (cf. [7, §5] and [10, (3.21)]):

$$h((t^l_{m,n})^* t^k_{r,s}) = \delta_{k,l}\delta_{m,r}\delta_{n,s}q^{2(l-m)}\frac{1-q^2}{1-q^{2(2l+1)}}. \tag{2.14}$$

Now (2.9) and (2.14) imply the orthogonality of the associated (σ,τ)-spherical elements $b^l_j(\sigma,\tau)$ with respect to the Haar functional,

$$h(b^l_j(\sigma,\tau)^* b^k_j(\sigma,\tau)) \begin{cases} = 0, & l \neq k, \\ > 0, & l = k. \end{cases}$$

Because of Proposition 2.3 [8, Thm. 8.2], [10, Prop. 4.7] $b^j_j(\sigma,\tau)^* b^j_j(\sigma,\tau) = P_j(\rho_{\sigma,\tau})$ for some nonzero polynomial P_j of degree smaller or equal to j. Now (2.13) implies

$$h(\overline{p_{l-j}}(\rho_{\sigma,\tau})P_j(\rho_{\sigma,\tau})p_{k-j}(\rho_{\sigma,\tau})) \begin{cases} = 0, & l \neq k, \\ > 0, & l = k, \end{cases}$$

since $\rho_{\sigma,\tau}$ is selfadjoint. Using the explicit form for the Haar functional acting on a polynomial in $\rho_{\sigma,\tau}$ [8, Thm. 8.4], [10, Thm. 5.3] we find that the polynomials p_{l-j} form a system of orthogonal polynomials for the positive-definite moment functional $\mathcal{L}_j$ (cf. [2, Chap. 1]) defined by

$$\mathcal{L}_j[x^n] = \int_{-\infty}^{\infty} x^n P_j(x) dm_{a,b,c,d;\,q^2}(x) < \infty,$$

where $dm_{a,b,c,d;\,q^2}(x)$ denotes the orthogonality measure for the Askey–Wilson polynomials $p_n(x;a,b,c,d \mid q^2)$ with $a = -q^{\sigma+\tau+1}$, $b = -q^{-\sigma-\tau+1}$, $c = q^{\sigma-\tau+1}$, and $d = q^{\tau-\sigma+1}$.

The statements for the polynomials q_{l-j} are proved analogously. □

Remark. In their announcement [13], Noumi and Mimachi obtain results using a similar approach. For a matrix

$$g = \begin{pmatrix} a & -\bar{c} \\ c & \bar{a} \end{pmatrix} \in SU(2) \times \mathbf{R}_{>0}$$

subject to $|a|^2 - |c|^2 q^{2k} \neq 0$ for all $k \in \mathbf{Z}$, they define

$$\theta(g) = \bar{a}cq^{1/2}B + a\bar{c}q^{-1/2}C - \frac{|a|^2 - |c|^2}{q - q^{-1}}(A - D) \in \mathcal{U}_q.$$

Note that for $\bar{a}c = i$ we get $\theta(g) = X_\tau$ with $q^\tau = |a|^2$. For two such matrices g_1, g_2 they obtain matrix elements $\psi^j_{m,n}(g_1, g_2)$ in $\mathcal{A}_q$ satisfying

$$\begin{cases} \theta(g_1).\psi^j_{m,n}(g_1, g_2) = \lambda_n(g_2)A.\psi^j_{m,n}(g_1, g_2), \\ \psi^j_{m,n}(g_1, g_2).\theta(g_2) = \lambda_m(g_1)\psi^j_{m,n}(g_1, g_2).A, \end{cases} \tag{2.15}$$

where $\lambda_n(g_2)$, $\lambda_m(g_1)$ are eigenvalues of $t^l(D\theta(g_2))$, $t^l(D\theta(g_1))$ similar in form to λ_j in (2.7). So their results are more general, since they also take the action from the left into consideration. However, Definition 2.1 does not fit into (2.15) since $\theta(g)D \neq X_\tau A$ for all g. But from (2.3) and (2.4) it is easily seen that b^* fits into (2.15), whenever b is an associated (σ, τ)-spherical element and g_1 and g_2 are chosen properly.

3. Associated (σ, τ)-spherical elements related to Askey–Wilson polynomials. As we saw in the previous section there is a natural way to relate associated (σ, τ)-spherical elements to systems of orthogonal polynomials, and in this section we will show that these polynomials are Askey–Wilson polynomials by investigating the moment functional defined in the proof of Proposition 2.5. These Askey–Wilson polynomials will have two continuous and one discrete parameter. Application of the comultiplication on a (σ, τ)-spherical element in $\operatorname{span}\{t^l_{n,m}\}$ yields an abstract addition formula for a two-parameter class of Askey–Wilson polynomials involving the Askey–Wilson polynomials of two continuous and one discrete variable. From this result an expansion for this two-parameter class of Askey–Wilson polynomials of argument $\cos(\theta + \varphi)$ is obtained.

To find the polynomials p_n and q_m as described in Proposition 2.5, we investigate the corresponding moment functional. In order to do this, we need to know for which polynomials P_j, Q_j we have $b^j_j(\sigma, \tau)^* b^j_j(\sigma, \tau) = P_j(\rho_{\sigma,\tau})$ and $b^j_{-j}(\sigma, \tau)^* b^j_{-j}(\sigma, \tau) = Q_j(\rho_{\sigma,\tau})$. The easiest way is to apply a one-dimensional $*$-representation of $\mathcal{A}_q$ to these equalities and the following lemma gives the heart of the solution. Although this lemma is not stated in its most elegant way, it has the right form for the applications later in this section.

LEMMA 3.1. *For arbitrary real σ, τ we have*

$$\begin{aligned}(q^{-\sigma-\tau+1}e^{i\theta}; q^2)_j(-q^{-\tau+\sigma+1}e^{i\theta}; q^2)_j &= e^{ij\theta}q^{\sigma j}q^{\frac{1}{2}j(j+1)}(q^2; q^2)^{\frac{1}{2}}_{2j} \\ &\times \sum_{n=-j}^{j} e^{-in\theta}q^{-\tau(j-n)}q^{\frac{1}{2}(j-n)(j-n-1)}\left(\frac{(q^{4j}; q^{-2})_{j-n}}{(q^2; q^2)_{j-n}}\right)^{\frac{1}{2}} \frac{q^{-\frac{n}{2}}q^{-(j+\sigma)n}q^{\frac{n^2}{2}}}{(q^2; q^2)^{\frac{1}{2}}_{j-n}(q^2; q^2)^{\frac{1}{2}}_{j+n}} \\ &\times {}_3\varphi_2\left(\begin{matrix} q^{-2(j-n)}, q^{-2j}, -q^{-2j-2\sigma} \\ 0, \ q^{-4j} \end{matrix}; q^2, q^2\right),\end{aligned}$$

$$\begin{aligned}(-q^{\sigma+\tau+1}e^{i\theta}; q^2)_j(q^{\tau-\sigma+1}e^{i\theta}; q^2)_j &= e^{ij\theta}q^{-\sigma j}q^{\frac{1}{2}j(j+1)}(q^2; q^2)^{\frac{1}{2}}_{2j} \\ &\times \sum_{n=-j}^{j} e^{-in\theta}q^{\tau(j-n)}q^{\frac{1}{2}(j-n)(j-n-1)}(-1)^{j-n}\left(\frac{(q^{4j}; q^{-2})_{j-n}}{(q^2; q^2)_{j-n}}\right)^{\frac{1}{2}} \\ &\times \frac{q^{-\frac{n}{2}}q^{-(j-\sigma)n}q^{\frac{n^2}{2}}}{(q^2; q^2)^{1/2}_{j-n}(q^2; q^2)^{\frac{1}{2}}_{j+n}} {}_3\varphi_2\left(\begin{matrix} q^{-2(j-n)}, q^{-2j}, -q^{-2j+2\sigma} \\ 0, \ q^{-4j} \end{matrix}; q^2, q^2\right).\end{aligned}$$

Proof. Recall from [1, eq. (1.15)] and [10, §2] the definition of the Askey–Wilson polynomials

$$p_n(y;a,b,c,d \mid q) = a^{-n}(ab,ac,ad;q)_n {}_4\varphi_3\left(\begin{matrix} q^{-n}, abcdq^{n-1}, az, a/z \\ ab, ac, ad \end{matrix}; q, q\right), \tag{3.1}$$

where $y = \frac{1}{2}(z+z^{-1})$, and let us consider the following generating function for the Askey–Wilson polynomials (cf. [5, eq. (1.9)]):

$$\sum_{n=0}^{\infty} \frac{p_n(y;a,b,c,d \mid q)t^n}{(ab,dc,q;q)_n} = {}_2\varphi_1\left(\begin{matrix} a/z, b/z \\ ab \end{matrix}; q, zt\right) {}_2\varphi_1\left(\begin{matrix} cz, dz \\ cd \end{matrix}; q, t/z\right) \tag{3.2}$$

with $y = \frac{1}{2}(z+z^{-1})$.

To obtain a generating function for the dual q-Krawtchouk polynomials from (3.2) we put $a = iq^{-1/2(N+\sigma)}$, $b = c = 0$, $d = -iq^{1/2(\sigma-N)}$, and $z = iq^{x-1/2(N+\sigma)}$ for $N \in \mathbf{N}$ and $x = 0, 1, \ldots, N$. For $0 \leq n \leq N$ and these choices of a, b, c, and d we get

$$p_n(y;a,b,c,d \mid q) = i^{-n}q^{\frac{n}{2}(N+\sigma)}(q^{-N};q)_n R_n(q^{-x} - q^{x-N-\sigma}; q^{\sigma}, N \mid q).$$

Note that $ad = q^{-N}$, $a/z = q^{-x}$ for $N \in \mathbf{N}$, $x = 0, 1, \ldots, N$ implies $p_n(y;a,b,c,d \mid q) = 0$ for $n > N$. So for these values of a, b, c, d and z the left-hand side of (3.2) reduces to a finite sum involving dual q-Krawtchouk polynomials, whereas the right-hand side yields a product of two ${}_1\varphi_0$'s, which can be summed using the q-binomial theorem; cf. [4, eq. (1.3.14)]. Hence,

$$\begin{aligned} \sum_{n=0}^{N} i^{-n}q^{\frac{n}{2}(N+\sigma)} \frac{(q^{-N};q)_n}{(q;q)_n} R_n(q^{-x} - q^{x-N-\sigma}; q^{\sigma}, N \mid q)t^n \\ = (itq^{-\frac{1}{2}(N+\sigma)};q)_x(-itq^{\frac{1}{2}(\sigma-N)};q)_{N-x}. \end{aligned} \tag{3.3}$$

In this generating function for the dual q-Krawtchouk polynomials we specialize $x = j$, $N = 2j$ and replace q by q^2 and n by $j-n$. Finally, put $t = iq^{2j+2\tau+1}e^{i\theta}$ to obtain the last statement of the lemma. The first statement follows from the last by replacing q^{τ} and σ by $-q^{-\tau}$ and $-\sigma$. □

Recall the definition of $a_n^{l,j}(\tau)$ given immediately after (2.10). It easily follows from [10, eq. (2.17)] that

$$a_n^{j,-j}(\tau) = C^{j,-j}(\tau)i^{j-n}q^{(j-n)\tau}q^{\frac{1}{2}(j-n)(j-n-1)}\left(\frac{(q^{4j};q^{-2})_{j-n}}{(q^2;q^2)_{j-n}}\right)^{\frac{1}{2}}, \tag{3.4}$$

since the dual q-Krawtchouk polynomial reduces to 1. However, in the other extreme case we can evaluate the dual q-Krawtchouk polynomial as well. In this case the ${}_3\varphi_2$ occurring in [10, eq. (2.17)] is actually a ${}_2\varphi_1$, since q^{-4j} appears as upper and lower parameter. This ${}_2\varphi_1$ can be summed by the q-Chu–Vandermonde formula [4, eq. (1.5.3)]. This yields $R_{j-n}(q^{-4j} - q^{-2\tau}; q^{2\tau}, 2j|q^2) = (-q^{-2\tau})^{j-n}$. Hence we find

$$a_n^{j,j}(\tau) = C^{j,j}(\tau)i^{j-n}q^{-(j-n)\tau}(-1)^{j-n}q^{\frac{1}{2}(j-n)(j-n-1)}\left(\frac{(q^{4j};q^{-2})_{j-n}}{(q^2;q^2)_{j-n}}\right)^{\frac{1}{2}}. \tag{3.5}$$

By $\pi^1_{\theta/2}$ we denote the one-dimensional $*$-representation as defined in [10, eq. (3.22)] for $\theta \in \mathbf{R}$. From [7, Thm. 5.3 or p. 108] and [10, eq. (3.23)] it follows that $\pi^1_{\theta/2}(t^j_{n,m})$

$= \delta_{nm}e^{-in\theta}$. Application of $\pi^1_{\theta/2}$ on $b^j_j(\sigma,\tau)$ and $b^j_{-j}(\sigma,\tau)$ reduces the double sum in (2.9) to a single sum of the type considered in Lemma 3.1. Thus we have shown that

$$\begin{aligned}\pi^1_{\theta/2}(b^j_j(\sigma,\tau)) = {} & C^{j,j}(\tau)q^{-\frac{1}{2}j(j+1)}q^{-\sigma j}i^j(q^2;q^2)^{-\frac{1}{2}}_{2j} \\ & \times e^{-ij\theta}(-q^{-\sigma-\tau+1}e^{i\theta};q^2)_j(q^{\sigma-\tau+1}e^{i\theta};q^2)_j\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}\pi^1_{\theta/2}(b^j_{-j}(\sigma,\tau)) = {} & C^{j,-j}(\tau)q^{-\frac{1}{2}j(j+1)}q^{-\sigma j}i^j(q^2;q^2)^{-\frac{1}{2}}_{2j} \\ & \times e^{-ij\theta}(-q^{\sigma+\tau+1}e^{i\theta};q^2)_j(q^{\tau-\sigma+1}e^{i\theta};q^2)_j.\end{aligned} \tag{3.7}$$

From (3.6) and (3.7) it follows immediately that

$$\begin{aligned}P_j(\cos\theta) &= \pi^1_{\theta/2}((b^j_j(\sigma,\tau))^* b^j_j(\sigma,\tau)) \\ &= D_j(-q^{-\sigma-\tau+1}e^{i\theta}, -q^{-\sigma-\tau+1}e^{-i\theta}, q^{\sigma-\tau+1}e^{i\theta}, q^{\sigma-\tau+1}e^{-i\theta};q^2)_j\end{aligned} \tag{3.8}$$

and

$$\begin{aligned}Q_j(\cos\theta) &= \pi^1_{\theta/2}((b^j_{-j}(\sigma,\tau))^* b^j_{-j}(\sigma,\tau)) \\ &= D_{-j}(-q^{\sigma+\tau+1}e^{i\theta}, -q^{\sigma+\tau+1}e^{-i\theta}, q^{-\sigma+\tau+1}e^{i\theta}, q^{-\sigma+\tau+1}e^{-i\theta};q^2)_j\end{aligned} \tag{3.9}$$

for nonzero constants D_j, $j \in \mathbf{Z}$. Now we can give an interpretation of a subclass of Askey–Wilson polynomials (cf. (3.1)) on the quantum group $SU_q(2)$.

THEOREM 3.2. *For $j \geq 0$ we have*

$$\begin{aligned}b^l_{-j}(\sigma,\tau) = {} & C^l_{-j}(\sigma,\tau)b^j_{-j}(\sigma,\tau) \\ & \times p_{l-j}(\rho_{\sigma,\tau}; -q^{\sigma+\tau+1+2j}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{\tau-\sigma+1+2j} \mid q^2)\end{aligned}$$

and

$$\begin{aligned}b^l_j(\sigma,\tau) = {} & C^l_j(\sigma,\tau)b^j_j(\sigma,\tau) \\ & \times p_{l-j}(\rho_{\sigma,\tau}; -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1+2j}, q^{\sigma-\tau+1+2j}, q^{\tau-\sigma+1} \mid q^2)\end{aligned}$$

with

$$C^l_{-j}(\sigma,\tau) = q^{\frac{1}{2}(j-l+j^2-l^2)}i^{l-j}q^{\sigma(j-l)}(q^{4l};q^{-2})^{-\frac{1}{2}}_{2(l-j)}(q^{4l};q^{-2})^{-1}_{l-j}\frac{C^{l,-j}(\tau)}{C^{j,-j}(\tau)},$$

$$C^l_j(\sigma,\tau) = q^{\frac{1}{2}(j-l+j^2-l^2)}i^{l-j}q^{\sigma(j-l)}(q^{4l};q^{-2})^{-\frac{1}{2}}_{2(l-j)}(q^{4l};q^{-2})^{-1}_{l-j}\frac{C^{l,j}(\tau)}{C^{j,j}(\tau)}.$$

Proof. Let us consider general Askey–Wilson polynomials $p_n(x;a,b,c,d|q)$. These polynomials form a system of orthogonal polynomials with respect to the orthogonality measure $dm_{a,b,c,d;\,q}(x)$, which consists of an absolutely continuous part on $[-1,1]$ and possibly a finite number of discrete mass points off $[-1,1]$; see [1, Thm. 2.5] and [4, §7.5]. A straightforward calculation shows that the orthogonality measure $dm_{aq^j,b,c,dq^j;\,q}(x)$ is equivalent to $b(x)dm_{a,b,c,d;\,q}(x)$, where

$$b\left(\frac{y+y^{-1}}{2}\right) = (ay, ay^{-1}, dy, dy^{-1};q)_j.$$

Apart from the constant the theorem now follows from the proof of Proposition 2.5, (3.8), (3.9), and the fact that the Askey–Wilson polynomial is symmetric in its parameters; cf. [1, p. 6].

The fact that the constant is correct remains to be proved. Apply $\pi^1_{\theta/2}$ to the second formula of the theorem and compare the coefficients of $e^{-il\theta}$ on both sides. The coefficient of $e^{-il\theta}$ in the left-hand side is

$$q^{-l/2}c_l^{l,\sigma}a_l^{l,j}(\tau) = C^{l,j}(\tau)q^{-l/2}\frac{i^l q^{-(l+\sigma)l}q^{l^2/2}}{(q^2;q^2)_{2l}^{1/2}}. \tag{3.10}$$

To calculate the coefficient of $e^{-il\theta}$ in the right-hand side we need the coefficient of $e^{-ij\theta}$ in $\pi^1_{\theta/2}(b_j^j(\sigma,\tau))$, which is (3.10) with l replaced by j, and the coefficient of $e^{-i(l-j)\theta}$ of

$$\begin{aligned} &p_{l-j}(\cos\theta;\ -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1+2j}, q^{\sigma-\tau+1+2j}, q^{\tau-\sigma+1}|q^2) \\ &\quad = 2^{l-j}(q^{2l+2j+2};q^2)_{l-j}(\cos\theta)^{l-j} + \text{lower-order terms} \end{aligned}$$

(cf. [1, p. 5]), which is $(q^{2l+2j+2};q^2)_{l-j} = (q^{4l};q^{-2})_{l-j}$. Comparing these coefficients yields $C_j^l(\sigma,\tau)$. The other constant can be calculated similarly. □

In order to formulate the next theorem we have to introduce a map $\Theta\colon \mathcal{A}_q \to \mathcal{A}_q$, which is defined on the generators by

$$\Theta(\alpha) = q^{-\frac{1}{2}}\alpha, \quad \Theta(\beta) = q^{\frac{1}{2}}\gamma, \quad \Theta(\gamma) = q^{-\frac{1}{2}}\beta, \quad \Theta(\delta) = q^{\frac{1}{2}}\delta,$$

and extended to $\mathcal{A}_q$ as an antilinear multiplicative map. It is easily checked that Θ preserves the commutation relations [10, eq. (3.1)] of $\mathcal{A}_q$, so it is well defined.

Now we consider the image of a matrix element $t^l_{n,m}$ under Θ. First we consider the matrix elements $t^l_{n,m} = t^l_{n,m}(\alpha,\beta,\gamma,\delta)$ as polynomials in α, β, γ, and δ. From [7, Thm. 5.3] we not only know that these polynomials have real coefficients, but it also gives

$$\Theta(t^l_{n,m}(\alpha,\beta,\gamma,\delta)) = t^l_{n,m}(q^{-\frac{1}{2}}\alpha, q^{\frac{1}{2}}\gamma, q^{-\frac{1}{2}}\beta, q^{\frac{1}{2}}\delta) = q^m t^l_{n,m}(\alpha,\gamma,\beta,\delta).$$

Proposition 4.1 of [7], which states that interchanging β and γ in $t^l_{n,m}$ is the same as interchanging m and n, now yields $\Theta(t^l_{n,m}) = q^m t^l_{m,n}$. Using this information and (2.9) gives

$$\Theta(b_j^l(\sigma,\tau)) = \sum_{n,m=-l}^{l} q^{m/2}\overline{c_m^{l,\sigma}a_n^{l,j}(\tau)}t^l_{m,n}. \tag{3.11}$$

THEOREM 3.3 (abstract addition formula). *The following identity in $\mathcal{A}_q \otimes \mathcal{A}_q$ is valid*:

$$\begin{aligned} &p_l(\Delta(\rho_{\sigma,\tau});\ -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1}|q^2) \\ &\quad = \sum_{j=1}^{l} A(l,-j)\Theta(b^j_{-j}(\tau,\tau))p_{l-j}(\Theta(\rho_{\tau,\tau});\ -q^{2\tau+1+2j}, -q^{-2\tau+1}, q, q^{1+2j}|q^2) \\ &\qquad \otimes b^j_{-j}(\sigma,\tau)p_{l-j}(\rho_{\sigma,\tau};\ -q^{\sigma+\tau+1+2j}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1+2j}|q^2) \\ &\qquad + A(l,0)p_l(\Theta(\rho_{\tau,\tau});\ -q^{2\tau+1}, -q^{-2\tau+1}, q, q|q^2) \\ &\qquad \otimes p_l(\rho_{\sigma,\tau};\ -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1}|q^2) \\ &\qquad + \sum_{j=1}^{l} A(l,j)\Theta(b_j^j(\tau,\tau))p_{l-j}(\Theta(\rho_{\tau,\tau});\ -q^{2\tau+1}, -q^{-2\tau+1+2j}, q^{1+2j}, q|q^2) \\ &\qquad \otimes b_j^j(\sigma,\tau)p_{l-j}(\rho_{\sigma,\tau};\ -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1+2j}, q^{\sigma-\tau+1+2j}, q^{-\sigma+\tau+1}|q^2), \end{aligned}$$

where the constants $A(l,j)$ are defined for $j \geq 0$ by

$$A(l,j) = q^{j-l}\left(c_j^{j,\sigma}\overline{c_j^{j,\tau}}\right)^{-1}\frac{(q^{2l+2};q^2)_l}{(q^{4l};q^{-2})_{l-j}^2}\frac{|C^{l,j}(\tau)|^2}{|C^{j,j}(\tau)|^2},$$

$$A(l,-j) = q^{j-l}\left(c_j^{j,\sigma}\overline{c_j^{j,\tau}}\right)^{-1}\frac{(q^{2l+2};q^2)_l}{(q^{4l};q^{-2})_{l-j}^2}\frac{|C^{l,-j}(\tau)|^2}{|C^{j,-j}(\tau)|^2}.$$

Proof. Application of the comultiplication on the (σ,τ)-spherical element contained in the span$\{t_{n,m}^l\}$ yields

$$\begin{aligned}
&\frac{c_l^{l,\sigma}\overline{c_l^{l,\tau}}}{(q^{2l+2};q^2)_l}p_l(\Delta(\rho_{\sigma,\tau});\ -q^{\sigma+\tau+1},-q^{-\sigma-\tau+1},q^{\sigma-\tau+1},q^{\tau-\sigma+1}\mid q)\\
&\quad=\Delta\left(\sum_{n,m=-l}^{l}q^{(n-m)/2}c_m^{l,\sigma}\overline{c_n^{l,\tau}}t_{n,m}^l\right)\\
&\quad=\sum_{n,m,k,r=-l}^{l}q^{(n-m)/2}c_m^{l,\sigma}\overline{c_n^{l,\tau}}\delta_{kr}t_{n,r}^l\otimes t_{k,m}^l\\
&\quad=\sum_{j=-l}^{l}\left(\sum_{n,r=-l}^{l}q^{n/2}\overline{c_n^{l,\tau}a_r^{l,j}(\tau)}t_{n,r}^l\right)\otimes\left(\sum_{k,m=-l}^{l}q^{-m/2}c_m^{l,\sigma}a_k^{l,j}(\tau)t_{k,m}^l\right)\\
&\quad=\sum_{j=-l}^{l}\Theta(b_j^l(\tau,\tau))\otimes b_j^l(\sigma,\tau)
\end{aligned}\tag{3.12}$$

by [10, Thm. 5.2, Prop. 4.7], [10, eq. (3.13)], (2.9), (2.10), and (3.11). In (3.12) we use Theorem 3.2 and (2.6). □

Of course we would like to have an addition formula in commuting variables, so let us apply $\pi_{\theta/2}^1\otimes\pi_{\varphi/2}^1$ to this identity in $\mathcal{A}_q\otimes\mathcal{A}_q$. It is easily checked that

$$\begin{aligned}
&\pi_{\theta/2}^1\otimes\pi_{\varphi/2}^1(\Delta(\rho_{\sigma,\tau}))=\cos(\theta+\varphi),\qquad \pi_{\varphi/2}^1(\rho_{\sigma,\tau})=\cos\varphi,\\
&\pi_{\theta/2}^1(\Theta(\rho_{\tau,\tau}))=\frac{q^{-1}e^{i\theta}+qe^{-i\theta}}{2}.
\end{aligned}\tag{3.13}$$

In view of (3.6) and (3.7) it remains to calculate $\pi_{\theta/2}^1(\Theta(b_{\pm j}^j(\sigma,\tau)))$. From the definition of Θ we see that for $a\in\mathcal{A}_q$ with $\pi_{\theta/2}^1(a)=\sum_k c_k e^{ik\theta}$ we have $\pi_{\theta/2}^1(\Theta(a))=\sum_k\overline{c_k}q^{-k}e^{ik\theta}$. This proves

$$\begin{aligned}
\pi_{\theta/2}^1(\Theta(b_{-j}^j(\sigma,\tau)))&=\overline{C^{j,-j}(\tau)}i^{-j}q^{-\frac{1}{2}j(j-1)}q^{-\sigma j}(q^2;q^2)_{2j}^{-\frac{1}{2}}\\
&\quad\times e^{-ij\theta}(-q^{\sigma+\tau}e^{i\theta};q^2)_j(q^{-\sigma+\tau}e^{i\theta};q^2)_j\\
\pi_{\theta/2}^1(\Theta(b_j^j(\sigma,\tau)))&=\overline{C^{j,j}(\tau)}i^{-j}q^{-\frac{1}{2}j(j-1)}q^{-\sigma j}(q^2;q^2)_{2j}^{-\frac{1}{2}}\\
&\quad\times e^{-ij\theta}(-q^{-\sigma-\tau}e^{i\theta};q^2)_j(q^{-\tau+\sigma}e^{i\theta};q^2)_j.
\end{aligned}\tag{3.14}$$

Now (3.13), (3.14), (3.6), (3.7), and Theorem 3.3 imply the following formula:

$$
\begin{aligned}
&p_l(\cos(\theta+\varphi);\, -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1}|q^2)\\
&\quad=\sum_{j=1}^{l} B(l,-j)e^{-ij\theta}e^{-ij\varphi}(-q^{2\tau}e^{i\theta}, e^{i\theta}, -q^{\sigma+\tau+1}e^{i\varphi}, q^{\tau-\sigma+1}e^{i\varphi}; q^2)_j\\
&\qquad\times p_{l-j}\left(\frac{q^{-1}e^{i\theta}+qe^{-i\theta}}{2};\, -q^{2\tau+1+2j}, -q^{-2\tau+1}, q, q^{1+2j}|q^2\right)\\
&\qquad\times p_{l-j}(\cos\varphi;\, -q^{\sigma+\tau+1+2j}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1+2j}|q^2)\\
&\qquad+B(l,0)p_l\left(\frac{q^{-1}e^{i\theta}+qe^{-i\theta}}{2};\, -q^{2\tau+1}, -q^{-2\tau+1}, q, q|q^2\right)\\
&\qquad\times p_l(\cos\varphi;\, -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1}|q^2)\\
&\qquad+\sum_{j=1}^{l} B(l,j)e^{-ij\theta}e^{-ij\varphi}(-q^{-2\tau}e^{i\theta}, e^{i\theta}, -q^{-\sigma-\tau+1}e^{i\varphi}, q^{\sigma-\tau+1}e^{i\varphi}; q^2)_j\\
&\qquad\times p_{l-j}\left(\frac{q^{-1}e^{i\theta}+qe^{-i\theta}}{2};\, -q^{2\tau+1}, -q^{-2\tau+1+2j}, q^{1+2j}, q|q^2\right)\\
&\qquad\times p_{l-j}(\cos\varphi;\, -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1+2j}, q^{\sigma-\tau+1+2j}, q^{-\sigma+\tau+1}|q^2),
\end{aligned}
\tag{3.15}
$$

where for $j \geq 0$,

$$
\begin{aligned}
B(l,j) &= q^{j-l}\frac{(q^{2l+2};q^2)_l}{(q^{4l};q^{-2})^2_{l-j}}|C^{l,j}(\tau)|^2,\\
B(l,-j) &= q^{j-l}\frac{(q^{2l+2};q^2)_l}{(q^{4l};q^{-2})^2_{l-j}}|C^{l,-j}(\tau)|^2.
\end{aligned}
$$

The value for the $B(l,j)$ immediately follows from Theorem 3.3, (3.14), (3.6), (3.7), and the following observation:

$$
\left(c_j^{j,\sigma}\overline{c_j^{j,\tau}}\right)^{-1} = (q^2;q^2)_{2j}q^{j^2}q^{j(\sigma+\tau)}.
$$

Remark. In their announcement [13] Noumi and Mimachi obtained a result similar to Theorem 3.2 in their Theorem 3 for an element $\psi^j_{m,n}(g_1,g_2)$ (cf. remark at the end of the previous section). $\psi^j_{m,n}(g_1,g_2)$ is expressed as an Askey–Wilson polynomial times a minimal element with respect to j, which still satisfies (2.15).

The formula (3.15) can be obtained from their Theorem 4 by putting $z = e^{i\theta}$, $w = qe^{i\varphi}$, $s = u = q^\tau$, $t = q^\sigma$, and replacing q by q^2. The extra parameter u in their Theorem 4 can also be easily obtained in this context if we replace $\overline{a_r^{l,j}(\tau)}a_k^{l,j}(\tau)$ in (3.12) by $\overline{a_r^{l,j}(\mu)}a_k^{l,j}(\mu)$ for some $\mu \in \mathbf{R}$. This also generalizes Theorem 3.3, which can be obtained from Noumi and Mimachi's paper by combining their Theorem 3 and Proposition 2c) for $m = n = 0$.

4. The addition formula for the continuous q-Legendre polynomials. In Theorem 3.3 we obtained an abstract addition formula for a two-parameter family of Askey–Wilson polynomials. In this section we consider the case $\sigma = \tau = 0$, so that we obtain an abstract addition formula for the continuous q-Legendre polynomials. The continuous q-Legendre polynomial is a special case of the continuous q-ultraspherical polynomial for which Rahman and Verma obtained an addition formula; cf. [14]. We show how to prove the result of Rahman and Verma [14, eq. (1.24)], [4, Ex. 8.11]

with $a = q^{1/4}$ from our abstract addition formula using the fact that the right-hand side of (4.6) consists of associated $(0,0)$-spherical elements. Finally, we note that the method we use is quite different from Koornwinder's as used in [9] to obtain an addition formula for the little q-Legendre polynomials, since we only use one-dimensional representations of $\mathcal{A}_q$, whereas Koornwinder uses infinite-dimensional representations. Koornwinder's result is also mentioned in [4, Ex. 7.41].

Formulas (4.1)–(4.5) are just restatements of the results obtained in the previous sections. The calculations involved are straightforward but tedious, so the reader is invited to skip these at first reading.

From [1, eqs. (4.20) and (4.2)] we see that

(4.1)
$$p_{l-j}(x; -q^{1+2j}, -q, q, q^{1+2j} \mid q^2) = p_{l-j}(x; -q, -q^{1+2j}, q^{1+2j}, q \mid q^2)$$
$$= \frac{(q^{4j+4};q^4)_{l-j}(q^4;q^4)_{l-j}}{(q^{4j+2};q^2)_{l-j}} C_{l-j}(x; q^{2+4j} \mid q^4),$$

where $C_n(x;\beta \mid q)$ denotes a continuous q-ultraspherical polynomial defined by

$$C_n(x;\beta \mid q) = \frac{(\beta^2;q)_n}{\beta^{n/2}(q;q)_n} {}_4\varphi_3\left(\begin{matrix} q^{-n}, q^n\beta^2, \beta^{1/2}e^{i\theta}, \beta^{1/2}e^{-i\theta} \\ \beta q^{1/2}, -\beta q^{1/2}, -\beta \end{matrix}; q, q\right)$$

with $x = \cos\theta$; cf. [1, §4]. Note that $C^{l,j}(0) = C^{l,-j}(0)$ (see (2.11)), and if we put for $j \geq 0$,

(4.2)
$$b^l_{\pm j} \overset{\text{def}}{=} i^{-l} q^{l^2/2} \left(\frac{(q^2;q^4)_l (q^{8l};q^{-4})_l}{(-q^2;q^2)_{2l}}\right)^{\frac{1}{2}} b^l_{\pm j}(0,0)$$
$$= q^{\frac{1}{2}(l-j)} \begin{bmatrix} l+j \\ l-j \end{bmatrix}_{q^4}^{-\frac{1}{2}} b^j_{\pm j} C_{l-j}(\rho_{0,0}; q^{2+4j} \mid q^4),$$

by Theorem 3.2 and (4.1); then we have

(4.3)
$$\pi^1_{\theta/2}(b^j_j) = \pi^1_{\theta/2}(b^j_{-j}) = \frac{1}{2}\sqrt{2}\left(\frac{1+q^{4j}}{(-q^2;q^2)_{2j}}\right)^{\frac{1}{2}}$$
$$\times \left(\frac{(q^2;q^4)_j}{(q^4;q^4)_j}\right)^{\frac{1}{2}} q^{-j/2} e^{-ij\theta} (q^2 e^{2i\theta}; q^4)_j$$

and

(4.4)
$$\pi^1_{\theta/2}(\Theta(b^j_j)) = \pi^1_{\theta/2}(\Theta(b^j_{-j})) = \frac{1}{2}\sqrt{2}\left(\frac{1+q^{4j}}{(-q^2;q^2)_{2j}}\right)^{\frac{1}{2}}$$
$$\times \left(\frac{(q^2;q^4)_j}{(q^4;q^4)_j}\right)^{\frac{1}{2}} q^{j/2} e^{-ij\theta} (e^{2i\theta}; q^4)_j$$

by (3.6), (3.7), and (3.14). Use (4.2) in (3.12) to obtain the abstract addition formula for the continuous q-Legendre polynomial:

(4.5)
$$C_l(\Delta(\rho); q^2 \mid q^4)$$
$$= q^l C_l(\Theta(\rho); q^2 \mid q^4) \otimes C_l(\rho; q^2 \mid q^4)$$
$$+ \sum_{j=1}^{l} q^{l-j} \begin{bmatrix} l+j \\ 2j \end{bmatrix}_{q^4}^{-1} \big(\Theta(b^j_j) C_{l-j}(\Theta(\rho); q^{2+4j} \mid q^4) \otimes b^j_j C_{l-j}(\rho; q^{2+4j} \mid q^4)$$
$$+ \Theta(b^j_{-j}) C_{l-j}(\Theta(\rho); q^{2+4j} \mid q^4) \otimes b^j_{-j} C_{l-j}(\rho; q^{2+4j} \mid q^4)\big),$$

where $\rho = \rho_{0,0} = \frac{1}{2}(\alpha^2 + q^{-1}\beta^2 + q\gamma^2 + \delta^2) = \rho^*$. Application of $\pi^1_{\theta/2} \otimes id$ on (4.5) yields

(4.6)
$$\begin{aligned} &C_l(\rho^\theta; q^2 \mid q^4) \\ &\quad = q^l C_l\left(\frac{q^{-1}e^{i\theta} + qe^{-i\theta}}{2}; q^2 \mid q^4\right) C_l(\rho; q^2 \mid q^4) \\ &\quad + \sum_{j=1}^{l} q^{l-j} \begin{bmatrix} l+j \\ 2j \end{bmatrix}_{q^4}^{-1} \frac{1}{2}\sqrt{2}\left(\frac{1+q^{4j}}{(-q^2;q^2)_{2j}}\right)^{\frac{1}{2}} \left(\frac{(q^2;q^4)_j}{(q^4;q^4)_j}\right)^{\frac{1}{2}} q^{j/2} e^{-ij\theta}(e^{2i\theta}; q^4)_j \\ &\quad \times C_{l-j}\left(\frac{q^{-1}e^{i\theta} + qe^{-i\theta}}{2}; q^{2+4j} \mid q^4\right)(b^j_j + b^j_{-j}) C_{l-j}(\rho; q^{2+4j} \mid q^4), \end{aligned}$$

where $\rho^\theta = (\pi^1_{\theta/2} \otimes id) \circ \Delta(\rho) = \frac{1}{2}(e^{i\theta}\alpha^2 + q^{-1}e^{i\theta}\beta^2 + qe^{-i\theta}\gamma^2 + e^{-i\theta}\delta^2) = (\rho^\theta)^*$. The formula (4.6) will be our starting point for the proof of the addition formula for the continuous q-Legendre polynomials.

Put $X = X_0 = iq^{1/2}B - iq^{-1/2}C \in \mathcal{U}_q$, then we have by [10, eq. (3.9)] , (2.1), (2.7), (4.1), Proposition 2.5, and Theorem 3.2,

$$(b^j_j + b^j_{-j}) C_{l-j}(\rho; q^{2+4j} \mid q^4).(XA)^2 = \left(\frac{q^{-2j} - q^{2j}}{q^{-1} - q}\right)^2 (b^j_j + b^j_{-j}) C_{l-j}(\rho; q^{2+4j} \mid q^4).$$

Consequently,

$$\begin{aligned} (b^j_j + b^j_{-j}) C_{l-j}(\rho; q^{2+4j} \mid q^4).\prod_{i=0}^{k-1}(1 - q^{4i}((q-q^{-1})^2(XA)^2 + 2) + q^{8i}) \\ = (q^{-4j}, q^{4j}; q^4)_k (b^j_j + b^j_{-j}) C_{l-j}(\rho; q^{2+4j} \mid q^4). \end{aligned} \tag{4.7}$$

Let us define

(4.8)
$$Y = \sum_{k=0}^{\infty} q^{4k} \frac{\prod_{i=0}^{k-1}(1 - q^{4i}((q-q^{-1})^2(XA)^2 + 2) + q^{8i})(e^{i\theta+i\varphi+i\psi}, e^{i\theta+i\varphi-i\psi}; q^4)_k}{(q^4, q^2, e^{2i\theta}, q^2e^{2i\varphi}; q^4)_k}.$$

Although Y is not defined as an element of the quantized, universal enveloping algebra $\mathcal{U}_q$, we will see that the action of Y on (4.6) from the right is well defined. For the right-hand side this follows directly from (4.7) and for the left-hand side we use a similar argument, which is worked out in more detail in the proof of Proposition 4.2.

If we first apply $.Y$ to the right-hand side of (4.6) and next $\pi^1_{\varphi/2}$, then we find by (4.8), (4.7), and (4.3),

$$\begin{aligned} &q^l C_l\left(\frac{q^{-1}e^{i\theta} + qe^{-i\theta}}{2}; q^2 \mid q^4\right) C_l(\cos\varphi; q^2 \mid q^4) \\ &\quad + \sum_{j=1}^{l} q^{l-j} \begin{bmatrix} l+j \\ 2j \end{bmatrix}_{q^4}^{-1} \frac{1+q^{4j}}{(-q^2;q^2)_{2j}} \frac{(q^2;q^4)_j}{(q^4;q^4)_j} e^{-ij\theta} e^{-ij\varphi} (e^{2i\theta}; q^4)_j (q^2e^{2i\varphi}; q^4)_j \\ &\quad \times C_{l-j}\left(\frac{q^{-1}e^{i\theta} + qe^{-i\theta}}{2}; q^{2+4j} \mid q^4\right) C_{l-j}(\cos\varphi; q^{2+4j} \mid q^4) \\ &\quad \times {}_4\varphi_3\left(\begin{matrix} q^{-4j}, q^{4j}, e^{i\theta+i\varphi+i\psi}, e^{i\theta+i\varphi-i\psi} \\ q^2,\ e^{2i\theta},\ q^2e^{2i\varphi} \end{matrix}; q^4, q^4\right). \end{aligned} \tag{4.9}$$

This ${}_4\varphi_3$ series is an Askey–Wilson polynomial without any factors, which we will denote by $\tilde{p}_j(\cos\psi; e^{i\theta+i\varphi}, q^2e^{-i\theta-i\varphi}, e^{i\theta-i\varphi}, q^2e^{-i\theta+i\varphi} \mid q^4)$.

THEOREM 4.1. *We have*

$$\begin{aligned}
&C_l(\cos\psi; q^2 \mid q^4)\\
&\quad = q^l C_l\left(\frac{q^{-1}e^{i\theta}+qe^{-i\theta}}{2}; q^2 \mid q^4\right) C_l(\cos\varphi; q^2 \mid q^4)\\
&\qquad + \sum_{j=1}^{l} q^{l-j} \begin{bmatrix} l+j \\ 2j \end{bmatrix}_{q^4}^{-1} \frac{1+q^{4j}}{(-q^2;q^2)_{2j}} \frac{(q^2;q^4)_j}{(q^4;q^4)_j} e^{-ij\theta}e^{-ij\varphi}(e^{2i\theta};q^4)_j(q^2e^{2i\varphi};q^4)_j\\
&\qquad \times C_{l-j}\left(\frac{q^{-1}e^{i\theta}+qe^{-i\theta}}{2}; q^{2+4j} \mid q^4\right) C_{l-j}(\cos\varphi; q^{2+4j} \mid q^4)\\
&\qquad \times \tilde{p}_j(\cos\psi; e^{i\theta+i\varphi}, q^2e^{-i\theta-i\varphi}, e^{i\theta-i\varphi}, q^2e^{-i\theta+i\varphi} \mid q^4).
\end{aligned}$$

Remark. If we replace $e^{i\theta}$ by $qe^{i\theta}$, then we obtain the addition formula of Rahman and Verma [14, eq. (1.24)] for $a = q^{1/4}$ with q replaced by q^4. The general Rahman–Verma addition formula for the continuous q-ultraspherical polynomials, [14, eq. (1.24)], can also be obtained from Theorem 4.1, as suggested by Askey. Application of the divided difference operator δ_{q^4} defined in [1, eq. (5.3)] on the result contained in Theorem 4.1 yields the addition formula for the continuous q-ultraspherical polynomials $C_l(\cos\psi; q^{2+4k} \mid q^4)$ for all $k \in \mathbf{Z}_+$ by [1, eq. (5.6)]. Since the Rahman–Verma addition formula for the continuous q-ultraspherical polynomials $C_l(\cos\psi; \beta \mid q)$ is a rational expression in $\beta^{1/2}$, the addition formula follows for all values of β by analytic continuation.

The theorem is proved if we could show that

$$\pi^1_{\varphi/2}(C_l(\rho^\theta; q^2 \mid q^4).Y) = C_l(\cos\psi; q^2 \mid q^4) \quad \forall\, l \in \mathbf{Z}_+,$$

or equivalently,

$$\pi^1_{\varphi/2}\big(f(\rho^\theta).Y\big) = f(\cos\psi) \quad \forall \text{ polynomials } f, \tag{4.10}$$

or equivalently,

$$\pi^1_{\varphi/2}\big(f_l(\rho^\theta).Y\big) = f_l(\cos\psi) \quad \forall\, l \in \mathbf{Z}_+ \tag{4.11}$$

for some basis f_l of the space of polynomials. By (4.8) the left-hand side of (4.11) is an expansion in the polynomials

$$\cos\psi \mapsto (e^{i(\theta+\varphi)}e^{i\psi}, e^{i(\theta+\varphi)}e^{-i\psi}; q^4)_l.$$

So it is natural to choose these polynomials as a basis for the space of polynomials, and then condition (4.11) is equivalent to

$$\pi^1_{\varphi/2}\left(\left(\prod_{j=0}^{l-1}(1-2e^{i(\theta+\varphi)}q^{4j}\rho^\theta+e^{2i(\theta+\varphi)}q^{8j})\right).Y\right) = (e^{i(\theta+\varphi)}e^{i\psi}, e^{i(\theta+\varphi)}e^{-i\psi}; q^4)_l$$

for all $l \in \mathbf{Z}_+$. Now this is a direct consequence of (4.8) and the following proposition.

PROPOSITION 4.2. *For the polynomial*

$$f_l(\cos\psi) = (e^{i(\theta+\varphi)+i\psi}, e^{i(\theta+\varphi)-i\psi}; q^4)_l = \prod_{j=0}^{l-1}(1 - 2e^{i(\theta+\varphi)}q^{4j}\cos\psi + e^{2i(\theta+\varphi)}q^{8j}),$$

we have

$$\pi^1_{\varphi/2}\left(f_l(\rho^\theta).\left(\prod_{i=0}^{k-1}(1 - q^{4i}((q-q^{-1})^2(XA)^2 + 2) + q^{8i})\right)\right)$$
$$= \delta_{lk}q^{-4k}(q^2, q^4, e^{2i\theta}, q^2e^{2i\varphi}; q^4)_k.$$

To prove this proposition we need two lemmas. The first one is straightforward, using the definitions in [10, §3].

LEMMA 4.3. *For $X = iq^{1/2}B - iq^{-1/2}C \in \mathcal{U}_q$ we have*

$$\alpha^2.(XA) = i(1+q^2)\gamma\alpha,\ \alpha^2.(XA)^2 = (1+q^2)(\alpha^2 - q^{-1}\gamma^2),\ \alpha^2.A^l = q^l\alpha^2,$$
$$\beta^2.(XA) = i(1+q^2)\delta\beta,\ \beta^2.(XA)^2 = (1+q^2)(\beta^2 - q^{-1}\delta^2),\ \beta^2.A^l = q^l\beta^2,$$
$$\gamma^2.(XA) = -iq^{-1}(1+q^2)\gamma\alpha,\ \gamma^2.(XA)^2 = (1+q^2)(q^{-2}\gamma^2 - q^{-1}\alpha^2),\ \gamma^2.A^l = q^{-l}\gamma^2,$$
$$\delta^2.(XA) = -iq^{-1}(1+q^2)\delta\beta,\ \delta^2.(XA)^2 = (1+q^2)(q^{-2}\delta^2 - q^{-1}\beta^2),\ \delta^2.A^l = q^{-l}\delta^2,$$

and

$$(\gamma\alpha).XA = iq^{-1}\gamma^2 - i\alpha^2, \quad (\delta\beta).XA = iq^{-1}\delta^2 - i\beta^2, \quad (\gamma\alpha).A = \gamma\alpha, \quad (\delta\beta).A = \delta\beta.$$

COROLLARY 4.4.

$$\pi^1_{\varphi/2}(2\rho^\theta.A^{2p}XAA^{2r}XAA^{2s}) = e^{-i(\theta+\varphi)}q^{-2-2(p+s)}(1+q^2)(1-q^{4p}e^{2i\theta})(1-q^{2+4s}e^{2i\varphi}).$$

LEMMA 4.5. *For all $l, n \in \mathbf{Z}_+$ and all $a, b, c, d \in \mathbf{C}$, we have*

$$\pi^1_{\varphi/2}\left((a\alpha^2 + b\beta^2 + c\gamma^2 + d\delta^2)^l.(XA)^{2n+1}\right) = 0.$$

Proof. For $l = 0$ this follows directly from $\epsilon(XA) = 0$, and we proceed by induction with respect to l. Since

$$\Delta(XA) = A^2 \otimes XA + XA \otimes 1, \tag{4.12}$$

we find that a general term $\Delta(XA)^{2n+1}$ is of the form $Z \otimes (XA)^m$, where Z consists of $2n+1-m$ terms XA intermingled with terms A^2. If m is odd, then

$$\pi^1_{\varphi/2}\left((a\alpha^2 + b\beta^2 + c\gamma^2 + d\delta^2)^{l-1}.(XA)^m\right) = 0$$

by the induction hypothesis. If m is even, then Z contains an odd number of XA, and by Lemma 4.3 and [10, eq. (3.22)] we find

$$\pi^1_{\varphi/2}\left((a\alpha^2 + b\beta^2 + c\gamma^2 + d\delta^2).Z\right) = 0.$$

Now [10, eq. (3.10)] provides for the induction step. □

Proof of Proposition 4.2. For any polynomial r_l of degree l we have by [6, Thm. 3.4],

$$r_l(\rho^\theta) \in \bigoplus_{k=0}^{l} \operatorname{span}\{t^k_{n,m} \mid n,m = -k, -k+1, \dots, k\}.$$

It follows from §2 that we can write

$$r_l(\rho^\theta) = \sum_{j=0}^{l} z_j, \qquad z_j.(XA)^2 = \left(\frac{q^{-2j} - q^{2j}}{q^{-1} - q}\right)^2 z_j.$$

Consequently,

$$\pi^1_{\varphi/2}\left(r_l(\rho^\theta).\left(\prod_{i=0}^{k-1}(1-q^{4i}((q-q^{-1})^2(XA)^2+2)+q^{8i})\right)\right) = \sum_{j=0}^{l}(q^{-4j}, q^{4j}; q^4)_k \pi^1_{\varphi/2}(z_j),$$

and this equals 0 for $k > l$, which proves the proposition in case $k > l$.

Next we will consider the case $k < l$. We will show that for all k and l with $k < l$ we have

$$\pi^1_{\varphi/2}\Big(f_l(\rho^\theta).(XA)^{2k}\Big) = 0. \tag{4.13}$$

To do this we use induction with respect to k. If $k = 0$, then

$$\pi^1_{\varphi/2}\Big(f_l(\rho^\theta)\Big) = (1, e^{2i(\theta+\varphi)}; q^4)_l = 0$$

for $l \geq 1 > k = 0$. Now we assume (4.13) true for all l, k with $l > k$ and $k < n$. Since

$$\Delta(XA)^{2n} = A^{4n} \otimes (XA)^{2n} + \text{terms of the form } Z \otimes (XA)^p,\ p \leq 2n-1, \tag{4.14}$$

[10, (3.10)], Lemma 4.5 and the induction hypothesis imply, for $l > n$,

$$\begin{aligned}\pi^1_{\varphi/2}\Big(f_l(\rho^\theta).(XA)^{2n}\Big) =& \pi^1_{\varphi/2}\left(\left(\prod_{j=n}^{l-1}(1 - 2e^{i(\theta+\varphi)}q^{4j}\rho^\theta + e^{2i(\theta+\varphi)}q^{8j})\right).A^{4n}\right) \\ &\times \pi^1_{\varphi/2}\left(\left(\prod_{j=0}^{n-1}(1 - 2e^{i(\theta+\varphi)}q^{4j}\rho^\theta + e^{2i(\theta+\varphi)}q^{8j})\right).(XA)^{2n}\right).\end{aligned}$$

Now $.A^{4n}$ is a homomorphism of $\mathcal{A}_q$, and so the first term in this product is

$$\prod_{j=n}^{l-1}(1 - e^{i(\theta+\varphi)}q^{4j}(q^{4n}e^{i(\theta+\varphi)} + q^{-4n}e^{-i(\theta+\varphi)}) + e^{2i(\theta+\varphi)}q^{8j}) = 0.$$

The case $l = k$ remains to be considered. Put

$$\begin{aligned}A_l =& \pi^1_{\varphi/2}\left(f_l(\rho^\theta).\left(\prod_{i=0}^{l-1}(1 - q^{4i}((q-q^{-1})^2(XA)^2 + 2) + q^{8i})\right)\right) \\ =& (-1)^l(q-q^{-1})^{2l}q^{2l(l-1)}\pi^1_{\varphi/2}\Big(f_l(\rho^\theta).(XA)^{2l}\Big),\end{aligned}$$

where we used (4.13). If we work out (4.14) a bit more we find

$$\begin{aligned}\Delta(XA)^{2l} =& A^{4l}\otimes(XA)^{2l}+\sum_{p+r+s=2(l-1)} A^{2p}XAA^{2r}XAA^{2s}\otimes(XA)^{2(l-1)}\\ &+\text{terms of the form } Z\otimes(XA)^{2n},\quad n<l-1\\ &+\text{terms of the form } Z\otimes(XA)^{2m+1},\end{aligned}$$

and thus by [10, eq. (3.10)], Lemma 4.5, and the part of the proposition already proved, we find

$$\begin{aligned}A_l =& (-1)^l(q-q^{-1})^{2l}q^{2l(l-1)}\\ &\times\pi^1_{\varphi/2}\left((1-2e^{i(\theta+\varphi)}q^{4(l-1)}\rho^\theta+e^{2i(\theta+\varphi)}q^{8(l-1)}).A^{4l}\right)\pi^1_{\varphi/2}\left(f_{l-1}(\rho^\theta).(XA)^{2l}\right)\\ &-q^{4(l-1)}(q-q^{-1})^2A_{l-1}\\ &\times\sum_{p+r+s=2(l-1)}\pi^1_{\varphi/2}\left((1-2e^{i(\theta+\varphi)}q^{4(l-1)}\rho^\theta+e^{2i(\theta+\varphi)}q^{8(l-1)}).A^{2p}XAA^{2r}XAA^{2s}\right).\end{aligned}$$

In order to rewrite this equation as a recurrence relation for A_l we use

$$\begin{aligned}&(-1)^l(q-q^{-1})^{2l}q^{2l(l-1)}(XA)^{2l}\\ &\quad=\prod_{i=0}^{l-1}(1-q^{4i}((q-q^{-1})^2(XA)^2+2)+q^{8i})\\ &\qquad+(-1)^l(q-q^{-1})^{2(l-1)}q^{2l(l-1)}\left(\frac{1-q^{-4l}}{1-q^{-4}}-2l+\frac{1-q^{4l}}{1-q^4}\right)(XA)^{2(l-1)}\\ &\qquad+\text{ lower-order terms}\end{aligned}$$

to obtain

(4.15)

$$\begin{aligned}A_l = A_{l-1}\Bigg[&-q^{4(l-1)}\pi^1_{\varphi/2}\left((1-2e^{i(\theta+\varphi)}q^{4(l-1)}\rho^\theta+e^{2i(\theta+\varphi)}q^{8(l-1)}).A^{4l}\right)\\ &\times\left(\frac{1-q^{-4l}}{1-q^{-4}}-2l+\frac{1-q^{4l}}{1-q^4}\right)\\ &+(q-q^{-1})^2q^{8(l-1)}e^{i(\theta+\varphi)}\sum_{p+r+s=2(l-1)}\pi^1_{\varphi/2}\left(2\rho^\theta.A^{2p}XAA^{2r}XAA^{2s}\right)\Bigg].\end{aligned}$$

To make this recurrence relation more explicit some calculations have to be made. First note that

$$\pi^1_{\varphi/2}\left((1-2e^{i(\theta+\varphi)}q^{4(l-1)}\rho^\theta+e^{2i(\theta+\varphi)}q^{8(l-1)}).A^{4l}\right)=(1-q^{-4})(1-q^{8l-4}e^{2i(\theta+\varphi)}).$$

The sum in (4.15) can be evaluated using Corollary 4.4, where we obtain

$$\begin{aligned}&q^{4(l-1)}e^{i(\theta+\varphi)}\sum_{p+r+s=2(l-1)}\pi^1_{\varphi/2}\left(2\rho^\theta.A^{2p}XAA^{2r}XAA^{2s}\right)\\ &=(1+q^{-2})\Bigg[\frac{q^{4l}(1-2lq^{-4l+2}+(2l-1)q^{-4l})+q^{4l-2}e^{2i(\theta+\varphi)}(1-2lq^{4l-2}+(2l-1)q^{4l})}{(1-q^2)^2}\\ &\qquad-\frac{e^{2i\theta}+q^2e^{2i\varphi}}{(1-q^2)(1-q^4)}(1-q^{4l-2}+q^{8l-2}-q^{4l})\Bigg].\end{aligned}$$

Plugging these two identities into (4.15) gives an explicit recurrence relation:

$$\begin{aligned} A_l &= q^{-4}(1-q^{4l})(1-q^{4l-2})(1-q^{4l-4}e^{2i\theta})(1-q^{4l-2}e^{2i\varphi})A_{l-1} \\ &= q^{-4l}(q^2,q^4,e^{2i\theta},q^2e^{2i\varphi};q^4)_l A_0. \end{aligned}$$

Since $A_0 = 1$, the proof of the proposition is complete. □

Remark. The map

$$\pi^1_{\theta/2} \otimes (\pi^1_{\varphi/2} \circ .Y) : \mathcal{A}_q \otimes \mathcal{A}_q \to \mathbf{C},$$

used to obtain Theorem 4.1 from (4.5), is not injective.

5. The limit case $q \uparrow 1$. In the previous section the addition formula for the continuous q-Legendre polynomials was derived from the abstract addition formula (4.5) by use of the map $\pi^1_{\theta/2} \otimes \pi^1_{\varphi/2} \circ .Y$. For $q \uparrow 1$ the abstract addition formula (4.5) can be regarded as an identity for functions on $SL(2,\mathbf{C}) \times SL(2,\mathbf{C})$, and we will show in this section that $\pi^1_{\theta/2} \otimes \pi^1_{\varphi/2} \circ .Y$ reduces to evaluation at some suitable element of $SL(2,\mathbf{C}) \times SL(2,\mathbf{C})$, which brings us back to the group theoretic proof of the addition formula for Legendre polynomials as presented in [16, Chap. 3.4].

We regard $\mathcal{A}_1$ as the commutative algebra of functions on $SL(2,\mathbf{C})$, which are polynomials in the coordinate elements α, β, γ, and δ. In particular, the comultiplication Δ of $\mathcal{A}_1$ is given by the group multiplication on $SL(2,\mathbf{C})$ as follows:

$$(\Delta f)(g,h) = f(gh) \quad \forall\, f \in \mathcal{A}_1 \quad \forall\, g,h \in SL(2,\mathbf{C}). \tag{5.1}$$

The one-dimensional representation $\pi^1_{\theta/2} : \mathcal{A}_1 \to \mathbf{C}$ corresponds to evaluation of functions at a diagonal element of $SL(2,\mathbf{C})$:

$$\pi^1_{\theta/2}(f) = f\left(\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix}\right) \quad \forall\, f \in \mathcal{A}_1. \tag{5.2}$$

To take the limit in $\mathcal{U}_q$ we replace A by $e^{((q-1)/2)H}$ and let $q \uparrow 1$; then $\mathcal{U}_q$ tends to the universal enveloping algebra $\mathfrak{U}(\mathfrak{sl}(2,\mathbf{C}))$ with generators H, B, and C and relations

$$[H,B] = 2B, \qquad [H,C] = -2C, \qquad [B,C] = H.$$

Then the right action of $\mathcal{U}_q$ on $\mathcal{A}_q$ corresponds to the action of the universal enveloping algebra $\mathfrak{U}(\mathfrak{sl}(2,\mathbf{C}))$ on functions on $SL(2,\mathbf{C})$ by right invariant differential operators.

First we consider the action of Y from the right as $q \to 1$. Note that $XA \to \tilde{X} = iB - iC$ for $q \uparrow 1$, which we identify with $\left(\begin{smallmatrix} 0 & i \\ -i & 0 \end{smallmatrix}\right) \in \mathfrak{sl}(2,\mathbf{C})$. Taking termwise limits in (4.8) formally yields

$$\tilde{Y} \overset{\text{def}}{=} \lim_{q\uparrow 1} Y = {}_2F_1\left(\begin{matrix} \tilde{X}/2, -\tilde{X}/2 \\ 1/2 \end{matrix}; \frac{\cos\psi - \cos\theta\cos\varphi + \sin\theta\sin\varphi}{2\sin\theta\sin\varphi}\right),$$

where the hypergeometric function is defined by

$${}_2F_1\left(\begin{matrix} a,b \\ c \end{matrix}; x\right) = \sum_{k=0}^{\infty} \frac{(a)_n (b)_n}{n!(c)_n} x^k, \qquad (a)_n = a(a+1)\cdots(a+n-1).$$

Pick z so that

$$\cos 2z = \frac{\cos\theta\cos\varphi - \cos\psi}{\sin\theta\sin\varphi}, \tag{5.3}$$

then (cf. [3, 2.8(11)])

$${}_2F_1\left(\begin{matrix}a/2,-a/2\\1/2\end{matrix};\frac{1-\cos 2y}{2}\right) = \cos ay = \frac{e^{iay}+e^{-iay}}{2}$$

implies that formally

$$\tilde{Y} = \lim_{q\uparrow 1} Y = \frac{e^{iz\tilde{X}}+e^{-iz\tilde{X}}}{2}. \tag{5.4}$$

So (4.8) is a q-analogue of (5.4) in combination with (5.3).

From (5.4) it follows that the action of the analogue of $.Y$ on an element $f \in \mathcal{A}_1$ is given by

$$(f.\tilde{Y})(g) = \frac{1}{2}\left(f\left(\begin{pmatrix}\cos z & -\sin z\\ \sin z & \cos z\end{pmatrix}g\right) + f\left(\begin{pmatrix}\cos z & \sin z\\ -\sin z & \cos z\end{pmatrix}g\right)\right), \tag{5.5}$$

for $g \in SL(2,\mathbf{C})$, which could have been defined without using the universal enveloping algebra $\mathfrak{U}(\mathfrak{sl}(2,\mathbf{C}))$. The transition of (4.6) to (4.9) corresponds to the fact that if $f \in \mathcal{A}_1$ satisfies

$$f\left(\begin{pmatrix}\cos z & -\sin z\\ \sin z & \cos z\end{pmatrix}g\right) = e^{ijz}f(g), \qquad g \in SL(2,\mathbf{C}),$$

then $(f.\tilde{Y})(g) = \cos(jz)f(g)$.

The previous paragraph and (5.2) give a clear understanding of the action of the analogue of $\pi^1_{\theta/2}\otimes\pi^1_{\varphi/2}\circ .Y$ on the right-hand side of (4.5) for $q\uparrow 1$. For the left-hand side we consider the action of the analogue of $\pi^1_{\theta/2}\otimes\pi^1_{\varphi/2}\circ .Y$ on an element $\Delta(f)\in\mathcal{A}_1\otimes\mathcal{A}_1$. From (5.1), (5.2), and (5.5) it follows that this action is given by

$$\frac{1}{2}\left(f\begin{pmatrix}e^{i(\varphi+\theta)/2}\cos z & -e^{i(\theta-\varphi)/2}\sin z\\ e^{i(\varphi-\theta)/2}\sin z & e^{-i(\theta+\varphi)/2}\cos z\end{pmatrix} + f\begin{pmatrix}e^{i(\varphi+\theta)/2}\cos z & e^{i(\theta-\varphi)/2}\sin z\\ -e^{i(\varphi-\theta)/2}\sin z & e^{-i(\theta+\varphi)/2}\cos z\end{pmatrix}\right).$$

If f is any polynomial in $\rho = \frac{1}{2}(\alpha^2+\beta^2+\gamma^2+\delta^2)$, say $f = p(\rho)$, we can give an explicit expression for this action. Since evaluation is a homomorphism of $\mathcal{A}_1$ and because of the quadratic terms γ^2 and β^2 in ρ we find that the action of the analogue of $\pi^1_{\theta/2}\otimes\pi^1_{\varphi/2}\circ .Y$ on $\Delta(p(\rho))$ is just

$$p(\rho)\begin{pmatrix}e^{i(\varphi+\theta)/2}\cos z & -e^{i(\theta-\varphi)/2}\sin z\\ e^{i(\varphi-\theta)/2}\sin z & e^{-i(\theta+\varphi)/2}\cos z\end{pmatrix} = p(\cos\psi), \tag{5.6}$$

where z and ψ are related by (5.3). So (5.6) is easily proved, whereas its analogue (4.10) requires a lengthy proof.

Acknowledgment. The author wishes to express his gratitude to Tom Koornwinder for his support and suggestions.

REFERENCES

[1] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 54 (1985).

[2] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[3] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F.G. TRICOMI, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[4] G. GASPER AND M. RAHMAN, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, 1990.

[5] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and ${}_4\varphi_3$ polynomials*, J. Approx. Theory, 36 (1982), pp. 43–54.

[6] H. T. KOELINK AND T. H. KOORNWINDER, *The Clebsch–Gordan coefficients for the quantum group $S_\mu U(2)$ and q-Hahn polynomials*, Nederl. Akad. Wetensch. Proc. Ser. A, 92 (1989), pp. 443–456.

[7] T. H. KOORNWINDER, *Representations of the twisted $SU(2)$ quantum group and some q-hypergeometric orthogonal polynomials*, Neder. Akad. Wetensch. Proc. Ser. A, 92 (1989), pp. 97–117.

[8] ———, *Orthogonal polynomials in connection with quantum groups*, in Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., NATO ASI series C, 294, Kluwer, Norwell, MA, 1990, pp. 257–292.

[9] ———, *The addition formula for little q-Legendre polynomials and the $SU(2)$ quantum group*, SIAM J. Math. Anal., 22 (1991), pp. 195–301.

[10] ———, *Askey–Wilson polynomials as zonal spherical functions on the $SU(2)$ quantum group*, CWI report AM-R9013, 1990; SIAM J. Math. Anal., 24 (1993), pp. 795–813.

[11] T. MASUDA, K. MIMACHI, Y. NAKAGAMI, M. NOUMI, AND K. UENO, *Representations of quantum groups and a q-analogue of orthogonal polynomials*, C.R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 559–564.

[12] ———, *Representations of the quantum group $SU_q(2)$ and the little q-Jacobi polynomials*, J. Funct. Anal., 99 (1991), pp. 357–386.

[13] M. NOUMI AND K. MIMACHI, *Askey–Wilson polynomials and the quantum group $SU_q(2)$*, Proc. Japan Acad. Ser. A Math., 66 (1990), pp. 146–149.

[14] M. RAHMAN AND A. VERMA, *Product and addition formulas for the continuous q-ultraspherical polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 1461–1474.

[15] L. L. VAKSMAN AND YA. S. SOIBELMAN, *Algebra of functions on the quantum group $SU(2)$*, Functional Anal. Appl., 22 (1988), pp. 170–181.

[16] N.J. VILENKIN, *Special Functions and the Theory of Group Representations*, Transl. Math. Monographs, 22, American Mathematical Society, Providence, RI, 1968.

[17] S.L. WORONOWICZ, *Compact matrix pseudogroups*, Comm. Math. Physics, 111 (1987), pp. 613–665.

[18] ———, *Twisted $SU(2)$ group. An example of non-commutative differential calculus*, Publ. Res. Inst. Math. Sci., 23 (1987), pp. 117–181.

SIAM J. MATH. ANAL.
Vol. 25, No. 1, pp. 218–241, January 1994

009

BASIC HYPERGEOMETRIC FUNCTIONS AND THE BOREL–WEIL CONSTRUCTION FOR $U_q(3)$*

M. A. LOHE†‡ AND L. C. BIEDENHARN†

Abstract. The Borel–Weil construction of the irreducible unitary representations of the quantum group $U_q(3)$ is investigated. The representation functions are calculated explicitly and found to be expressible in terms of basic hypergeometric functions; the form of these basis functions verifies previous general results. It is shown that several identities satisfied by basic hypergeometric functions, including a special case of Watson's formula, are implied by properties of the quantum group.

Key words. quantum group, basic hypergeometric functions, irreducible representations, Borel–Weil construction

AMS subject classifications. 33A70, 20G45

1. Introduction.

Quantum groups have achieved significance in recent years as algebraic structures, which have appeared in several areas of both mathematics and physics including, for example, knot theory, gauge theories, and statistical mechanical models. Quantum groups are Hopf algebras which are neither commutative nor co-commutative; alternatively, we can regard them as deformations of the universal enveloping algebra of an underlying Lie group, so that in the limit in which the deformation parameter q approaches 1 we regain the Lie algebra. Various aspects of quantum groups, including the fundamental theory and applications to physics, are investigated in [1] (where further references may be found), and for general discussions we mention [2]–[4].

Of particular interest are the irreducible representations (irreps) of quantum groups (for generic real q), for it is through these that the quantum group manifests itself in physical applications. The explicit construction of all representations of $SU_q(2)$, the simplest of the quantum groups, can be carried out in several different ways, one of which is to express the generators in terms of finite difference operators acting in the space of homogeneous polynomials of some fixed degree [5]. For $q \to 1$ one regains the familiar harmonic oscillator representation of $SU(2)$, in which the generators are realized as differential operators, or equivalently, can be expressed in terms of boson creation and annihilation operators. This construction is, unfortunately, difficult to generalize to $U_q(n)$ in such a way as to obtain all representations.

There is, however, another method of constructing representations of $SU_q(2)$ that *does* admit a relatively simple generalization to arbitrary n, the Borel–Weil (BW) construction [6]. This method, which allows us to construct (for generic real q) *all* unitary irreps of $U_q(n)$, is recursive in that one assumes that all irreps of the subgroup $U_q(n-1) \times U(1)$ have already been constructed and is explicit to the extent that the basis vectors spanning an irreducible representation space for $U_q(n)$ are expressed in terms of those for the subgroup $U_q(n-1) \times U(1)$. For $n = 2$ this construction is straightforward and appears in [6]. Our aim in this paper is to provide explicit details of the BW construction, and its q-extension, for the case $n = 3$; this is of

* Received by the editors February 18, 1992; accepted for publication (in revised form) December 9, 1992. This research was supported in part by the National Science Foundation.

† Department of Physics, University of Texas at Austin, Austin, Texas 78712-1081.

‡ Permanent address, Northern Territory University, P.O. Box 40146, Casuarina, Northern Territory, Australia, 0811.

interest for several reasons. First, it is only for $n \geq 3$ that several complexities of the quantum group appear, such as the Serre relations, which define the quantum group (see §2), and the explicit realizations show how these relations are satisfied. Second, it will be seen that q-extensions of classical functions appear naturally in the BW construction, and that general results for the form of the basis vectors imply identities among these functions. This is true even for the simplest case $n = 2$, where we use the q-exponential function to construct q-BW states, with addition properties being related to co-multiplication in $SU_q(2)$. For the case $n = 3$, terminating basic hypergeometric functions, q-analogs of ${}_6F_5$ functions arise in our calculations of the q-BW basis states, and we reduce these to ${}_3\phi_2$ functions by using a special case of the q-analog of Whipple's transformation (Watson's formula), which reduces a ${}_8\phi_7$ basic hypergeometric function to a ${}_4\phi_3$ function. Moreover, since we know from general considerations [6] that the final form of the basis vector *must* involve a ${}_3\phi_2$ function (which is related to a q-Clebsch–Gordan coefficient), we have in effect proved the special case of Watson's formula by using quantum group properties.

It is, of course, well known that many properties of special functions can be understood as group theoretical properties of basis vectors for irreps of Lie groups (see, for example, Vilenkin [7] and [8]–[10]). It is now clear, however, that quantum groups underlie properties of q-extensions to special functions in a similar way; examples of this are the q-Clebsch–Gordan and q-Racah coefficients of $SU_q(2)$, which can be expressed as ${}_3\phi_2$ and ${}_4\phi_3$ functions, and for which properties such as the symmetries and orthogonality relations are implied by the structure of the quantum group [11], [12]. Some of these properties were, of course, derived well before the formulation of quantum groups (by Askey and Wilson [13]), but it is through the quantum group that one achieves a unifying perspective.

The plan of the paper is to first outline in §2 some basic facts about the quantum groups $U_q(2)$ and $U_q(3)$, also including a discussion on q-extensions to classical functions such as the q-exponential function, basic hypergeometric functions, and properties of the finite difference operator. Next, in §3, we show how the BW construction is applied to $U(2)$ and its quantum deformation $U_q(2)$, and then in §4 we explicitly calculate the BW basis states for $n = 3$. This calculation serves more than to merely rederive a certain form for these states, found from previous considerations [6]; the method used here is direct and the explicit calculation demonstrates that hypergeometric functions appear as coefficients of basis vectors. The general form of the basis states therefore implies the existence of identities and transformations that must be satisfied by the hypergeometric functions, including a transformation due to Whipple. In §5 we formulate q-BW states (using slightly different conventions than those of [6]) and show that the $U_q(3)$ algebra is satisfied by the realization of the quantum group generators. Finally, in §6, we present the explicit calculation of the q-BW states, in which the basic hypergeometric functions appear.

2. Quantum groups and q-extensions. The relations which define the quantum group $U_q(n)$ have been given by several authors [2], [3], [15]. In the case of $n = 2$ there are four generators, denoted $E_{ij}(i, j = 1, 2)$, satisfying the commutation relations

$$
\begin{aligned}
&[E_{11}, E_{12}] = E_{12}, \qquad [E_{11}, E_{21}] = -E_{21}, \\
&[E_{22}, E_{12}] = -E_{12}, \qquad [E_{22}, E_{21}] = E_{21},
\end{aligned} \tag{2.1}
$$

and

$$[E_{12}, E_{21}] = \frac{q^{\frac{1}{2}(E_{11}-E_{22})} - q^{-\frac{1}{2}(E_{11}-E_{22})}}{q^{\frac{1}{2}} - q^{-\frac{1}{2}}}, \tag{2.2}$$

where q is a positive real number. We will frequently use the notation

$$\begin{aligned}[n] &= \frac{q^{n/2} - q^{-n/2}}{q^{\frac{1}{2}} - q^{-\frac{1}{2}}} \\ &= q^{(n-1)/2} + q^{(n-3)/2} + \cdots q^{-(n-1)/2},\end{aligned} \tag{2.3}$$

where n is an integer, but we extend this notation so that n can also be an operator; hence (2.2) may be written $[E_{12}, E_{21}] = [E_{11} - E_{22}]$. For $q \to 1$ we have $[n] \to n$, and so in this limit (2.1) and (2.2) reduce to the usual commutation relations for $U(2)$.

The quantum group $U_q(3)$ is generated by the elements $\{e_i, f_i, h_i \mid i = 1, 2\}$, which satisfy the following equations:

$$\begin{aligned}&[h_i, h_j] = 0, \\ &[h_i, e_j] = k_{(i,j)} e_j, \qquad [h_i, f_j] = -k_{(i,j)} f_j, \\ &[e_i, f_j] = \delta_{ij}[2h_i],\end{aligned} \tag{2.4}$$

where

$$k_{(i,j)} = \begin{cases} 1, & i = j, \\ -\frac{1}{2}, & i = j + 1, \\ 0, & \text{otherwise}, \end{cases}$$

together with the $U(1)$ element h_3, which commutes with the other generators. As in [6] we will use the notation

$$\begin{aligned}&E_{12} = e_1, \quad E_{21} = f_1, \quad E_{23} = e_2, \quad E_{32} = f_2, \\ &\tfrac{1}{2}(E_{11} - E_{22}) = h_1, \quad \tfrac{1}{2}(E_{22} - E_{33}) = h_2, \quad E_{11} + E_{22} + E_{33} = h_3.\end{aligned} \tag{2.5}$$

In addition, the definition of quantum group $U_q(3)$ includes the following Serre relations:

$$\begin{aligned}&E_{12}^2 E_{23} - [2] E_{12} E_{23} E_{12} + E_{23} E_{12}^2 = 0, \\ &E_{12} E_{23}^2 - [2] E_{23} E_{12} E_{23} + E_{12} E_{23}^2 = 0,\end{aligned} \tag{2.6}$$

and the conjugate relations

$$\begin{aligned}&E_{21}^2 E_{32} - [2] E_{21} E_{32} E_{21} + E_{32} E_{21}^2 = 0, \\ &E_{21} E_{32}^2 - [2] E_{32} E_{21} E_{32} + E_{21} E_{32}^2 = 0.\end{aligned} \tag{2.7}$$

In general, $U_q(n)$ has the structure of a Hopf algebra with comultiplication Δ: $U_q(n) \to U_q(n) \otimes U_q(n)$ which is defined for the elements $\{e_i, f_i, h_i\}$ by

$$\begin{aligned}\Delta(h_i) &= h_i \otimes I + I \otimes h_i, \\ \Delta(e_i) &= e_i \otimes q^{-h_i/2} + q^{h_i/2} \otimes e_i, \\ \Delta(f_i) &= f_i \otimes q^{-h_i/2} + q^{h_i/2} \otimes f_i.\end{aligned} \tag{2.8}$$

The definition of a Hopf algebra also requires the concepts of a co-unit ε and antipode γ, which are readily defined for the quantum group:

$$\varepsilon(1) = 1, \qquad \varepsilon(e_i) = \varepsilon(f_i) = \varepsilon(h_i) = 0,$$
$$\gamma(e_i) = -q^{-\frac{1}{2}} e_i, \quad \gamma(f_i) = -q^{\frac{1}{2}} f_i, \quad \gamma(h_i) = -h_i.$$

It is well known that we can construct representations of $U(n)$ in which the generators are realized as differential operators acting on polynomials in complex variables [5]–[7] or, equivalently, we can express this construction in the language of boson creation and annihilation operators $a, \bar{a}$ acting on a vacuum state $|0\rangle$, and satisfying $[\bar{a}, a] = 1$ (see [16]). In order to construct representations in a similar way for the quantum groups, we introduce q-boson operators [17], [18] a^q and $\bar{a}^q$, satisfying

$$\bar{a}^q a^q - q^{\frac{1}{2}} a^q \bar{a}^q = q^{-\frac{N}{2}}, \tag{2.9}$$

where N (the number operator) satisfies

$$[N, a^q] = a^q, \qquad [N, \bar{a}^q] = -\bar{a}^q. \tag{2.10}$$

The q-boson vacuum $|0\rangle$ satisfies $\bar{a}^q|0\rangle = 0$, and basis states are constructed by allowing q-boson operators to act on the vacuum $|0\rangle$. Equivalently, we can realize q-boson operators as finite difference operators by defining, for suitable functions $f(z)$,

$$a^q f(z) = z f(z), \tag{2.11a}$$

$$\bar{a}^q f(z) = D_q f(z) \equiv \frac{f(zq^{\frac{1}{2}}) - f(zq^{-\frac{1}{2}})}{z(q^{\frac{1}{2}} - q^{-\frac{1}{2}})},$$

$$N\, f(z) = z\, \frac{\partial f(z)}{\partial z}. \tag{2.11b}$$

The last equation implies that $q^{-\frac{N}{2}} f(z) = f(zq^{-\frac{1}{2}})$ and the relations (2.9) and (2.10) can then be verified. Evidently, D_q acts as a finite difference operator and for $q \to 1$ becomes differentiation $\partial/\partial z$; we can regard the properties of D_q as comprising a "q-calculus."

It follows from (2.11) that

$$D_q z^n = [n] z^{n-1}, \tag{2.12}$$

where $[n]$ is defined by (2.3). Let us define the q-exponential function $\exp_q$ by

$$\exp_q(z) = \sum_{n=0}^{\infty} \frac{z^n}{[n]!}, \tag{2.13}$$

where $[n]! = [n][n-1]\cdots[1]$. Then as a result of (2.12),

$$D_q \exp_q(Az) = A \exp_q(Az), \tag{2.14}$$

where A is a constant, or an operator independent of z. The q-exponential is a q-analog of the classical exponential function, although as such it is not unique; however, it is invariant under $q \leftrightarrow q^{-1}$. The finite difference operator D_q and q-extensions to classical functions are not new to quantum groups; they were studied some time ago by Jackson [19] and the subject has been developed extensively by Askey [13], Andrews [20], and also by Milne [21] and Koornwinder [22]. Many of the results of this q-calculus have been derived by Feinsilver [23] using operator methods, in which operators satisfying (2.9) (or an equivalent relation) are postulated and q-identities are developed from algebraic considerations (see also Cigler [24]).

The q-exponential and the operators $a^q, \bar{a}^q$ appear in the BW construction of $U_q(n)$ states, and several further properties will be used extensively there, in particular the following operator equations:

$$a^q \bar{a}^q = [N], \qquad \bar{a}^q a^q = [N+1] \tag{2.15}$$

(these equations imply (2.9) directly and, with certain assumptions, can be derived from them).

Apart from the q-exponential function, we will use another q-extension of classical functions, the basic hypergeometric function, which can be defined in the following way: Let

$$(a;q)_n = \begin{cases} 1, & n = 0, \\ (1-a)(1-aq)\cdots(1-aq^{n-1}), & n > 0, \end{cases} \tag{2.16}$$

where n is an integer, then the basic hypergeometric function ${}_{p+1}\phi_p$ is defined by

$${}_{p+1}\phi_p\left(\begin{matrix} a_1 a_2 \cdots a_{p+1} \\ b_1 b_2 \cdots b_p \end{matrix}; q, z\right) = \sum_{n=0}^{\infty} \frac{(a_1;q)_n (a_2;q)_n \cdots (a_{p+1};q)_n}{(q;q)_n (b_1;q)_n \cdots (b_p;q)_n} z^n. \tag{2.17}$$

We can express the symbol $(a;q)_n$ in terms of the notation (2.3) by means of the formula

$$(a;q)_n = (q^\alpha;q)_n = (1-q)^n q^{\frac{n}{4}(n+2\alpha-3)} ([\alpha])_n, \tag{2.18}$$

where we have put $a = q^\alpha$ and

$$([\alpha])_n = [\alpha][\alpha+1]\cdots[\alpha+n-1]. \tag{2.19}$$

We can therefore write ${}_{p+1}\phi_p$ in the form

$$\begin{aligned} {}_{p+q}\phi_p\left(\begin{matrix} q^{\alpha_1} q^{\alpha_2} \cdots q^{\alpha_{p+1}} \\ q^{\beta_1} q^{\beta_2} \cdots q^{\beta_p} \end{matrix}; q, z\right) &= \sum_{n=0}^{\infty} \frac{([\alpha_1])_n ([\alpha_2])_n \cdots ([\alpha_{p+1}])_n}{[n]!([\beta_1])_n \cdots ([\beta_p])_n} (q^{\frac{\sigma}{2}} z)^n \\ &= {}_{p+1}\phi_p\left(\begin{matrix} q^{-\alpha_1} q^{-\alpha_2} \cdots q^{-\alpha_{p+1}} \\ q^{-\beta_1} \cdots q^{-\beta_p} \end{matrix}; q^{-1}, z q^\sigma\right), \end{aligned} \tag{2.20}$$

where

$$\sigma = \sum_{i=1}^{p+1} \alpha_i - \sum_{i=1}^{p} \beta_i - 1. \tag{2.21}$$

For $q = 1$ this function reduces to

$$(2.22) \qquad {}_{p+1}F_p\left(\begin{matrix}\alpha_1 \cdots \alpha_{p+1} \\ \beta_1 \cdots \beta_p\end{matrix}; z\right).$$

For the case $\sigma = -2$ and $z = q$ the function ${}_{p+1}\phi_p$ is called *balanced* or *Saalschützian.* If one of the numerator parameters $\alpha_1 \cdots \alpha_{p+1}$ is a negative integer $-m$ the series terminates, and in this case any one of the denominator parameters $\beta_1 \cdots \beta_p$ can also be negative, provided it is less than $-m$. Basic hypergeometric functions occur naturally within quantum group structures, for example, the q-Clebsch–Gordan and q-Racah coefficients can be expressed in terms of ${}_3\phi_2$ and ${}_4\phi_3$ functions, respectively [11], [12]. Basic hypergeometric functions were first introduced by Heine [25] and a recent exposition of their properties has been presented by Gasper and Rahman [26].

Finally, let us mention the q-binomial theorem, from which an addition theorem can be derived for the q-exponential function. An elegant formulation can be written in terms of quantum coordinates a, b, which satisfy

$$(2.23) \qquad ba = q\, ab,$$

and then the q-binomial theorem states, for positive integers n,

$$(2.24) \qquad (a+b)^n = \sum_{j=1}^{n} \frac{q^{j(n-j)/2}[n]!}{[j]![n-j]!} a^j b^{n-j},$$

which can be proved by induction on n. (For a discussion of this theorem and applications using q-calculus see Feinsilver [23].) We will interpret the quantum coordinates a, b as operators to be constructed from quantum group generators.

3. The BW construction for $U(2)$ and $U_q(2)$. The Borel–Weil (BW) construction of irreducible representations of the unitary groups was originally formulated in differential geometric terms as the construction of a line bundle over the homogeneous space $U(n)\backslash(U(n-1) \times U(1))$, in which the fiber carries an irreducible representation of $U(n-1) \times U(1)$ with sections that are holomorphic functions in the homogeneous space. An algebraic version of this construction can be formulated in the framework of vector coherent states (LeBlanc and Biedenharn [14]), and this procedure can be generalized directly to quantum groups [6], unlike the fiber bundle construction for which too little is known of nonplanar quantum manifolds.

Let us outline the construction of vector coherent states, which we refer to as BW states. These are described in more detail in [14]; BW states are formed from the subset of Gel'fand–Weyl basis vectors annihilated by the raising operators $E_{in}, i = 1, \ldots, n-1$ of $U(n)$. This subset is therefore given by

$$(3.1) \qquad \{|(\mu)\rangle\} \equiv \{|(m)_n\rangle : E_{in}|(m)_n\rangle = 0, i = 1, \ldots, n-1\},$$

where $|(m)_n\rangle$ is a basis vector in the irrep space of $U(n)$ labelled by $[\mathbf{m}_n] = [m_{1n}, \ldots, m_{nn}]$ and $(m)_n$ is an n-rowed Gel'fand–Weyl pattern. The linear space spanned by the vectors $\{|(\mu)\rangle\}$ carries the irrep

$$(3.2) \qquad [\mu_{1,n-1}, \mu_{2,n-1}, \ldots, \mu_{n-1,n-1}] \otimes [m_{nn}],$$

where $\mu_{i,n-1} = m_{in}, i = 1, \ldots, n-1$, which is of the subgroup $U(n-1) \times U(1)$. The BW states are vector-valued holomorphic functions defined by

$$(3.3) \qquad |(m)_n\rangle_{BW} = \sum_{(\mu)} \langle(\mu)|e^{z \cdot E}|(m)_n\rangle|(\mu)\rangle,$$

where $z \cdot E = \sum_{i=1}^{n-1} z_i E_{in}$, and $\{z_i\}$ is a set of $(n-1)$ complex variables used as coordinates for the co-set space $U(n)\backslash(U(n-1)\times U(1))$; these complex variables $\{z_i\}$ can be identified with boson creation operators $\{a_i\}$ acting on the vacuum $|0\rangle$, as mentioned in §2. The sum in (3.3) is over all Gel'fand–Weyl patterns (μ) carrying the irrep of $U(n-1)\times U(1)$ shown in (3.2).

The basis $|(m)_n\rangle_{BW}$ carries the $U(n)$ irrep $[\mathbf{m}_n]$, with the group action defined by

$$g \circ |(m)_n\rangle_{BW} \equiv \sum_{(\mu)} \langle(\mu)|e^{z\cdot E} g|(m)_n\rangle|(\mu)\rangle, \tag{3.4}$$

where $g \in U(n)$. This group action can be generalized to the quantum group, provided it is formulated in terms of the Lie algebra; the BW state realization $\Gamma(E_{ij})$ of a generator E_{ij} of $U(n)$ acts on the BW states according to

$$\Gamma(E_{ij})|(m)_n\rangle_{BW} = \sum_{(\mu)} \langle(\mu)|e^{z\cdot E} E_{ij}|(m)_n\rangle|(\mu)\rangle. \tag{3.5}$$

It can be shown [14] that the basis vectors $|(m)_n\rangle_{BW}$ take the form

$$|(m)_n\rangle_{BW} = K\begin{pmatrix}[\mathbf{m}_n]\\ [\mathbf{m}_{n-1}]\end{pmatrix} \sum_{(\mu),(\mu')} C^{[\mu][w\dot{0}][m_{n-1}]}_{(\mu)(\mu')(m_{n-1})} \left(\left| \begin{matrix}[w\ \dot{0}]\\ (\mu')\end{matrix} \right\rangle \otimes \left| \begin{matrix}[\mu]\\ (\mu)\end{matrix} \right\rangle \right), \tag{3.6}$$

where

(i) the numerical constant K depends only on the $U(n)$ irrep labels $[\mathbf{m}_n]$ and the $U(n-1)$ irrep labels $[\mathbf{m}_{n-1}]$, according to the formula

$$K\begin{pmatrix}[\mathbf{m}_n]\\ [\mathbf{m}_{n-1}]\end{pmatrix} = \left(\prod_{i=1}^{n-1} \frac{(p_{in} - p_{nn} - 1)!}{(p_{i,n-1} - p_{nn})!} \right)^{\frac{1}{2}}, \tag{3.7}$$

where $p_{ij} \equiv m_{ij} + j - i$;

(ii) the Wigner–Clebsch–Gordan coefficient $C^{\cdots}_{\cdots}$ effects the tensor coupling: $[w\ \dot{0}] \times [\mu] \to [\mathbf{m}_{n-1}]$;

(iii) The irrep vector $|\begin{smallmatrix}[w\ \dot{0}]\\ (\mu')\end{smallmatrix}\rangle$ is homogeneous and holomorphic in the boson operators $\{z_i\}$ acting on the vacuum with the $U(n-1)$ irrep labels $[w, 0\cdots 0]$, where

$$w = \sum_{i=1}^{n-1} (m_{i,n-1} - \mu_{i,n-1}); \tag{3.8}$$

(iv) the $U(n-1)$ irrep labels of the fiber vector $|\begin{smallmatrix}[\mu]\\ (\mu)\end{smallmatrix}\rangle$ are given by $\mu_{i,n-1} = m_{i,n}$ for $i = 1, 2, \ldots, n-1$. (These fiber vectors are actually tensored with a fixed $U(1)$ vector carrying the irrep $[m_{nn}]$ but this is suppressed to avoid complication.)

Let us illustrate this construction and the rather complicated result (3.6) by writing the $U(2)$ BW states explicitly, and deriving the form (3.6) for this case. It is then straightforward to determine the extension to the quantum group. The $U(2)$ BW states are defined by

$$\left| \begin{matrix} m_{12} & m_{22} \\ m_{11} \end{matrix} \right\rangle_{BW} = \left\langle \begin{matrix} m_{12} & m_{22} \\ m_{12} \end{matrix} \right| \exp(aE_{12}) \left| \begin{matrix} m_{12} & m_{22} \\ m_{11} \end{matrix} \right\rangle |0\rangle \otimes \left| \begin{matrix} m_{12} & m_{22} \\ m_{12} \end{matrix} \right\rangle, \tag{3.9}$$

where we have used the definition (3.1) in which the summation over the patterns (μ) has one term only, and where we have also used the operator notation in which a boson operator a acts on $|0\rangle$. Using the known matrix elements of E_{12},

$$(3.10)\quad E_{12}\left|\begin{matrix} m_{12} & m_{22} \\ & m_{11} \end{matrix}\right\rangle = \sqrt{(m_{12}-m_{11})(m_{11}+1-m_{22})}\left|\begin{matrix} m_{12} & m_{22} \\ & m_{11}+1 \end{matrix}\right\rangle$$

and orthonormality of the Gel'fand–Weyl basis vectors, we find

(3.11)

$$\left|\begin{matrix} m_{12} & m_{22} \\ & m_{11} \end{matrix}\right\rangle_{BW} = \left(\frac{(m_{12}-m_{22})!}{(m_{11}-m_{22})!}\right)^{\frac{1}{2}} \frac{a^{m_{12}-m_{11}}}{\sqrt{(m_{12}-m_{11})!}}|0\rangle \otimes \left|\begin{matrix} m_{12} & m_{22} \\ & m_{12} \end{matrix}\right\rangle,$$

where the factors correspond to those shown in (3.6), i.e., the K-factor and the boson polynomial.

Now let us extend this construction to the quantum group $U_q(2)$, defining the q-BW states by

(3.12)

$$\left|\begin{matrix} m_{12} & m_{22} \\ & m_{11} \end{matrix}\right\rangle_{q,BW} = \left\langle\begin{matrix} m_{12} & m_{22} \\ & m_{12} \end{matrix}\right| \exp_q(a^q E_{12}) \left|\begin{matrix} m_{12} & m_{22} \\ & m_{11} \end{matrix}\right\rangle |0\rangle \otimes \left|\begin{matrix} m_{12} & m_{22} \\ & m_{12} \end{matrix}\right\rangle,$$

where the q-boson operator and the q-exponential are defined in (2.9), (2.10), and (2.13), respectively. These states may be evaluated directly from the matrix elements of E_{12}:

$$(3.13)\quad E_{12}\left|\begin{matrix} m_{12} & m_{22} \\ & m_{11} \end{matrix}\right\rangle = \sqrt{[m_{12}-m_{11}][m_{11}+1-m_{22}]}\left|\begin{matrix} m_{12} & m_{22} \\ & m_{11}+1 \end{matrix}\right\rangle,$$

and we obtain a straightforward analog of (3.11):

(3.14)

$$\left|\begin{matrix} m_{12} & m_{22} \\ & m_{11} \end{matrix}\right\rangle_{q,BW} = \left(\frac{[m_{12}-m_{22}]!}{[m_{11}-m_{22}]!}\right)^{\frac{1}{2}} \frac{(a^q)^{m_{12}-m_{11}}}{\sqrt{[m_{12}-m_{11}]!}}|0\rangle \otimes \left|\begin{matrix} m_{12} & m_{22} \\ & m_{12} \end{matrix}\right\rangle.$$

From the definition of the q-BW states and the action of the algebra given in (3.2), we can calculate the q-BW realization $\Gamma(E_{ij})$ of the generators E_{ij}, as summarized in the following two lemmas.

LEMMA 3.1. *The realization $\Gamma(E_{ij})$ of the $U_q(2)$ generators acting on the basis $|(m)\rangle_{q,BW}$ is given by*

$$(3.15a)\qquad \Gamma(E_{11}) = E_{11} - N,$$

$$(3.15b)\qquad \Gamma(E_{22}) = E_{22} + N,$$

$$(3.15c)\qquad \Gamma(E_{12}) = \bar{a}^q,$$

$$(3.15d)\qquad \Gamma(E_{21}) = a^q[E_{11} - E_{22} - N],$$

where N is the number operator for the q-bosons, satisfying (2.10).

Proof. By using the commutation relations (2.1) we find

$$[E_{11} - N, a^q E_{12}] = 0 = [E_{22} + N, a^q E_{12}], \tag{3.16}$$

from which we obtain $(E_{11}-N)\exp_q(a^qE_{12})\,|\,0\rangle = \exp_q(a^qE_{12})E_{11}\,|\,0\rangle$, which in turn implies (3.15a), and similarly implies (3.15b). Equation (3.15c) follows upon using the q-boson form of (2.14). To obtain (3.15d) we need the relation

$$E_{21}E_{12}^n = E_{12}^n E_{21} - [E_{11} - E_{22} - n + 1][n]E_{12}^{n-1}, \qquad n \in \mathbb{Z} \tag{3.17}$$

which is proved by induction on n (the proof uses the identity

$$[a][b+c] - [b][a+c] = [a-b][c] \tag{3.18}$$

for $a = E_{11} - E_{22} - n$, $b = n$, c $= 1$). From (3.17),

$$\exp_q(a^qE_{12})E_{21}\,|\,0\rangle = (E_{21} + a^q[E_{11} - E_{22} - N])\exp_q(a^qE_{12})\,|\,0\rangle \tag{3.19}$$

follows, and now we use the fact that the state $|(\mu)\rangle$ (shown specifically in (3.12)) is of highest weight in $U(1)$, i.e., $\langle(\mu)|E_{21} = 0$ in order to obtain (3.15d). □

LEMMA 3.2. *The map* $\Gamma : g \to U_q(g)$, *where* g *is a generator of* $U_q(2)$, *given by* (3.15), *is an isomorphism of the algebra* $U_q(2)$.

Proof. The commutation relations (2.1) follow immediately, except

$$[\Gamma(E_{12}), \Gamma(E_{21})] = [\Gamma(E_{11}) - \Gamma(E_{22})] = [E_{11} - E_{22} - 2N], \tag{3.20}$$

which follows from (2.15) and the identity (3.18) for $a = E_{11} - E_{22} - N$, $b = N$, $c = 1$. □

4. Explicit BW states for $U(3)$. For $U(3)$ the BW states defined by (3.1) take the form

$$\left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22} \\ & & m_{11} \end{matrix}\right\rangle_{BW} = \sum_m E_m((m);z) \left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{13} & m_{23} \\ & & m \end{matrix}\right\rangle, \tag{4.1}$$

where the coefficients are the matrix elements

$$E_m((m);z) = \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{13} & m_{23} \\ & & m \end{matrix} \right| \exp(z_1E_{13} + z_2E_{23}) \left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22} \\ & & m_{11} \end{matrix}\right\rangle, \tag{4.2}$$

where (m) denotes the array (m_{ij}). We now verify the form (3.6) of these states by explicit calculation, proceeding by determining first the matrix elements in (4.2). This we do by writing the exponential in (4.2) as a product of two exponential factors (using the addition theorem for exponentials), each of which is expanded as an infinite series; however, only one term of each series contributes due to orthonormality of the Gel'fand–Weyl basis. We therefore require the matrix elements of E_{13} and E_{23}, to arbitrary powers, in the Gel'fand–Weyl basis, and then express (4.2) as a sum. We find that this can be written in terms of a hypergeometric ${}_6F_5$ function. Next we use an identity due to Whipple that expresses the ${}_6F_5$ function in terms of a ${}_3F_2$ function, which in turn can be expressed as a Clebsch–Gordan coefficient. This calculation leads us directly to the final form (3.6) that we seek; the fact that the final answer *must* be

of this form in fact implies the existence of Whipple's identity, as well as symmetries of the hypergeometric functions, as noted in the Introduction.

For $n = 3$, (3.6) reads

$$\left| \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{12} & m_{22} & \\ m_{11} & & \end{matrix} \right\rangle_{BW} = (-1)^{m_{22}-m_{23}} \left[\frac{(m_{13}-m_{33}+1)!(m_{23}-m_{33})!}{(m_{12}-m_{33}+1)!(m_{22}-m_{33})!} \right]^{\frac{1}{2}}$$

(4.3)

$$\times \sum_m \frac{z_1^{m-m_{11}}(-z_2)^{w+m_{11}-m}}{\sqrt{(m-m_{11})!(w+m_{11}-m)!}} C^{j_1 j_2 j}_{m_1 m_2 m_1+m_2} \left| \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13} & m_{23} & \\ m & & \end{matrix} \right\rangle,$$

where

$$\begin{aligned} w &= m_{13} + m_{23} - m_{12} - m_{22}, \\ j_1 &= \tfrac{1}{2}(m_{13} - m_{23}), \quad j_2 = \tfrac{1}{2} w, \quad j = \tfrac{1}{2}(m_{12} - m_{22}), \\ m_1 &= m - \tfrac{1}{2}(m_{13} + m_{23}), \qquad m_2 = m_{11} - m + \tfrac{1}{2} w. \end{aligned} \tag{4.4}$$

Here we have also substituted from (3.7) for the explicit K-factor, namely,

$$K\begin{pmatrix} [\mathbf{m}_3] \\ [\mathbf{m}_2] \end{pmatrix}^2 = \prod_{i=1}^{2} \frac{(p_{i3} - p_{33} - 1)!}{(p_{i2} - p_{33})!}, \tag{4.5}$$

and have included the phase factor $(-1)^\phi$ (see [14, eq. (2.16)]), where

$$\begin{aligned} \phi &= \phi([\mu]) - \phi([0, -w]) - \phi([\mathbf{m}_2]) \\ &= \phi([m_{13}, m_{23}]) - \phi([0, -w]) - \phi([m_{12}, m_{22}]) \\ &= m_{22} - m_{23}. \end{aligned} \tag{4.6}$$

We now provide details of the calculations, together with properties of the hypergeometric functions used in order to obtain precisely the form (4.3). Apart from its intrinsic interest, this calculation serves as a guide for establishing similar results for the quantum group case.

In order to evaluate the matrix elements (4.2) we first require the matrix elements of E_{13} and E_{23} in the Gel'fand–Weyl basis. Explicit formulas for matrix elements in $U(n)$, first found by Gel'fand and Zetlin [27], are given by Baird and Biedenharn [28], and for the particular case $n = 3$ and for E_{23} these reduce to

(4.7)

$$\left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{12} & m_{22} & \\ m_{11} & & \end{matrix} \right| E_{23}$$

$$= \left[\frac{(m_{12}-m_{11})(m_{13}-m_{12}+1)(m_{12}-m_{23})(m_{12}-m_{33}+1)}{(m_{12}-m_{22}+1)(m_{12}-m_{22})} \right]^{\frac{1}{2}} \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{12}-1 & m_{22} & \\ m_{11} & & \end{matrix} \right|$$

$$+ \left[\frac{(m_{11}-m_{22}+1)(m_{13}-m_{22}+2)(m_{23}-m_{22}+1)(m_{22}-m_{33})}{(m_{12}-m_{22}+2)(m_{12}-m_{22}+1)} \right]^{\frac{1}{2}} \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{12} & m_{22}-1 & \\ m_{11} & & \end{matrix} \right|.$$

By repeated application of E_{23} we obtain the general form

$$(4.8)\qquad \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} \quad m_{22} & \\ & m_{11} & \end{matrix} \right| (E_{23})^n = \sum_{r=0}^{n} A_r^n(m) \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12}-r \quad m_{22}+r-n & \\ & m_{11} & \end{matrix} \right|,$$

where (m) is the array $\left(\begin{smallmatrix} m_{12} & m_{22} \\ & m_{11} \end{smallmatrix}\right)$, and the coefficients A_r^n are to be determined. This can be done by deriving a recurrence relation for A_r^n using (4.7). We find

(4.9)

$$\begin{aligned} &A_r^{n+1} \\ &= A_{r-1}^n \left[\frac{(m_{12}-m_{11}-r+1)(m_{13}-m_{12}+r)(m_{12}-m_{23}-r+1)(m_{12}-m_{33}-r+2)}{(m_{12}-m_{22}+n-2r+3)(m_{12}-m_{22}+n-2r+2)}\right]^{\frac{1}{2}} \\ &\quad + A_r^n \left[\frac{(m_{11}-m_{22}+n-r+1)(m_{13}-m_{22}+n-r+2)}{(m_{12}-m_{22}+n-2r+2)}\right. \\ &\qquad \left. \times \frac{(m_{23}-m_{22}+n-r+1)(m_{22}-m_{33}-n+r)}{(m_{12}-m_{22}+n-2r+1)}\right]^{\frac{1}{2}} \end{aligned}$$

with $A_0^0 = 1$. Let us merely state the solution of these recurrence relations, as verified by direct substitution:

(4.10)

$$\begin{aligned} A_r^n &= \frac{n!(m_{12}-m_{22}-r)!}{r!(n-r)!(m_{12}-m_{22}+n-r+1)!} \\ &\quad \times \left[\frac{(m_{11}-m_{22}+n-r)!(m_{23}-m_{22}+n-r)!(m_{22}-m_{33})!}{(m_{11}-m_{22})!(m_{23}-m_{22})!(m_{22}-m_{33}-n+r)!}\right. \\ &\qquad \times \frac{(m_{12}-m_{11})!(m_{12}-m_{23})!(m_{12}-m_{33}+1)!(m_{13}-m_{12}+r)!}{(m_{12}-m_{11}-r)!(m_{12}-m_{23}-r)!(m_{12}-m_{33}-r+1)!(m_{13}-m_{12})!} \\ &\qquad \left. \times \frac{(m_{13}-m_{22}+n-r+1)!(m_{12}-m_{22}+1)(m_{12}-m_{22}+n-2r+1)}{(m_{13}-m_{22}+1)!}\right]^{\frac{1}{2}}. \end{aligned}$$

We also need to evaluate the matrix elements of $(E_{13})^\ell$, for arbitrary integers ℓ, but in this case we require the action of $(E_{13})^\ell$ only on those states for which $m_{12} = m_{13}, m_{22} = m_{23}$. Since $E_{13} = [E_{12}, E_{23}]$, the matrix elements of E_{13} are easily found

from those of E_{23}, given in (4.7), and those of E_{12} in (3.10), to give

(4.11)

$$\left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13} & m_{23} \\ & m_{11} \end{matrix} \right| E_{13} = \sqrt{\frac{(m_{13}-m_{33}+1)(m_{11}-m_{23})}{(m_{13}-m_{23}+1)}} \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13}-1 & m_{23} \\ & m_{11}-1 \end{matrix} \right| - \sqrt{\frac{(m_{23}-m_{33})(m_{13}-m_{11}+1)}{(m_{13}-m_{23}+1)}} \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13} & m_{23}-1 \\ & m_{11}-1 \end{matrix} \right| .$$

From this follows the general form:

$$\left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13} & m_{23} \\ & m_{11} \end{matrix} \right| (E_{13})^{\ell} = \sum_{s=0}^{\ell} B_s^{\ell}(m) \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13}-s & m_{23}-\ell+s \\ & m_{11}-\ell \end{matrix} \right| , \tag{4.12}$$

where (m) is the array $\left(\begin{smallmatrix} m_{13} & m_{23} \\ & m_{11} \end{smallmatrix}\right)$ for some coefficients B_s^{ℓ}. As before, we derive a recurrence relation for B_s^{ℓ}, using (4.11):

(4.13)

$$\begin{aligned} &B_s^{\ell+1} \\ &= B_{s-1}^{\ell} \left[\frac{s(m_{13}-m_{23}-s+1)(m_{13}-m_{33}-s+2)(m_{11}-m_{23}-s+1)}{(m_{13}-m_{23}+\ell-2s+3)(m_{13}-m_{23}+\ell-2s+2)} \right]^{\frac{1}{2}} \\ &- B_s^{\ell} \left[\frac{(m_{13}-m_{23}+\ell-s+2)(\ell-s+1)(m_{23}-m_{33}-\ell+s)(m_{13}-m_{11}+\ell-s+1)}{(m_{13}-m_{23}+\ell-2s+2)(m_{13}-m_{23}+\ell-2s+1)} \right]^{\frac{1}{2}} \end{aligned}$$

with $B_0^0 = 1$. The solution, which again is verified by direct substitution, is

(4.14)

$$\begin{aligned} B_s^{\ell} &= (-1)^{\ell+s}\ell! \\ &\times \left[\frac{(m_{13}-m_{33}+1)!(m_{11}-m_{23})!(m_{13}-m_{23}-s)!(m_{13}-m_{11}+\ell-s)!}{s!(\ell-s)!(m_{13}-m_{23}+\ell+1-s)!(m_{13}-m_{33}+1-s)!(m_{11}-m_{23}-s)!} \right. \\ &\times \left. \frac{(m_{23}-m_{33})!(m_{13}-m_{23}+\ell-2s+1)}{(m_{13}-m_{11})!(m_{23}-m_{33}-\ell+s)!} \right]^{\frac{1}{2}} . \end{aligned}$$

We can now evaluate $E_m((m);z)$ by expanding the exponentials, giving

(4.15)

$$E_m((m);z) = \sum_{\ell,n} \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13} & m_{23} \\ & m \end{matrix} \right| \frac{z_1^{\ell} z_2^{n}}{\ell!\, n!} (E_{13})^{\ell} (E_{23})^{n} \left| \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{12} & m_{22} \\ & m_{11} \end{matrix} \right\rangle$$

$$= \sum_{\ell,n=0}^{\infty} \sum_{s=0}^{\ell} \sum_{r=0}^{n} \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ m_{13}-s-r, & m_{23}-\ell+s+r-n & \\ & m-\ell & \end{matrix} \right| \frac{(z_1)^{\ell}(z_2)^{n}}{\ell!\, n!} B_s^{\ell} \begin{pmatrix} m_{13} & m_{23} \\ & m \end{pmatrix}$$

$$\times A_r^n \begin{pmatrix} m_{13}-s, m_{23}-\ell+s \\ m-\ell \end{pmatrix} \left| \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22} \\ & & m_{11} \end{matrix} \right\rangle .$$

Since the Gel'fand–Weyl states are orthonormal, we can eliminate three of the four summations, namely, the sums over ℓ, n, r. We must have

$$\begin{aligned} \ell &= m - m_{11}, \\ r &= m_{13} - m_{12} - s, \\ n &= w - m + m_{11}, \end{aligned} \tag{4.16}$$

where $w = m_{13} + m_{23} - m_{12} - m_{22}$. The matrix element now takes the form

(4.17)

$$\begin{aligned} &E_m((m); z) \\ &\quad = \sum_s \frac{z_1^{m-m_{11}}\, z_2^{w-m+m_{11}}}{(m-m_{11})!(w-m+m_{11})!} B_s^{m-m_{11}} \\ &\qquad \cdot \begin{pmatrix} m_{13} & m_{23} \\ & m \end{pmatrix} A^{w-m+m_{11}}_{m_{13}-m_{12}-s} \begin{pmatrix} m_{13}-s & m_{23}+s-m+m_{11} \\ & m_{11} \end{pmatrix}. \end{aligned}$$

Next, we collect all the factors as given in (4.10) and (4.14) in order to obtain the resulting matrix elements

(4.18)

$$\begin{aligned} E_m((m); z) &= z_1^{m-m_{11}}\, z_2^{w-m+m_{11}} (-)^{m-m_{11}} \\ &\times \left[\frac{(m_{13}-m_{33}+1)!(m_{23}-m_{33})!(m_{11}-m_{22})!(m_{23}-m_{22})!}{(m_{22}-m_{33})!(m_{12}-m_{33}+1)!(m_{12}-m_{11})!(m_{12}-m_{23})!(m_{13}-m_{12})!} \right. \\ &\quad \left. \times \frac{(m_{12}-m_{22}+1)}{(m_{13}-m)!(m-m_{23})!(m_{13}-m_{22}+1)!} \right]^{\frac{1}{2}} \\ &\times \frac{(m_{13}-m_{23})!(m_{12}+m-m_{23}-m_{11})!(m_{13}-m_{11})!}{(m-m_{11})!(m_{13}+m-m_{11}-m_{23})!(m_{11}+m_{23}-m_{22}-m)!} \sum_s R(s), \end{aligned}$$

where $R(s)$ is a product of all those terms containing s-dependent factors, and may be written as

$$R(s) = \frac{(-)^s (1+a/2)_s (a)_s (b)_s (c)_s (d)_s (e)_s}{s!(a/2)_s (a-b+1)_s (a-c+1)_s (a-d+1)_s (a-e+1)_s}, \tag{4.19}$$

where

$$\begin{aligned} a &= m_{11} + m_{23} - m_{13} - m - 1, \qquad b = m_{11} - m, \\ c &= m_{12} - m_{13}, \qquad d = m_{23} - m, \\ e &= m_{22} - m_{13} - 1, \end{aligned} \tag{4.20}$$

and where it can be seen that all square roots have combined to give a rational function of the basis state labels.

The function obtained by summing the terms $R(s)$ over s can be identified as a generalized hypergeometic function with argument -1:

(4.21)
$$\sum_s R(s) = {}_6F_5\left(\begin{matrix} 1+a/2 & a & b & c & d & e \\ a/2 & a-b+1 & a-c+1 & a-d+1 & a-e+1 & \end{matrix}\ ;-1\right).$$

This function can be expressed in a simpler form by means of a special case of Whipple's transformation, which itself transforms a terminating well-poised ${}_7F_6$ into a Saalschützian ${}_4F_3$ (see Bailey [29, §4.3]). This transformation is derived by Bailey (§4.4, eq. (2)), and reads

(4.22)
$$\begin{aligned} &{}_6F_5\left(\begin{matrix} 1+a/2 & a & b & c & d & e \\ a/2 & a-b+1 & a-c+1 & a-d+1 & a-e+1 & \end{matrix}\ ;-1\right) \\ &\quad = (-1)^c \frac{(-a-1)_{-c}}{(a-e+1)_{-c}}\, {}_3F_2\left(\begin{matrix} 1+a-b-d & c & e \\ 1+a-b & 1+a-d & \end{matrix}\right), \end{aligned}$$

where we have used the fact that $-c = m_{13} - m_{12}$ is a positive integer.

We now obtain from (4.17),

(4.23)
$$\begin{aligned} &E_m((m);z) \\ &= \frac{z_1^{m-m_{11}}\, z_2^{w-m+m_{11}}}{\sqrt{(m-m_{11})!(w-m+m_{11})!}}\,(-)^{m-m_{11}+m_{12}-m_{13}}(m_{13}-m_{23})!(m_{13}-m_{11})! \\ &\times \left[\frac{(m_{13}-m_{33}+1)!(m_{23}-m_{33})!(m_{11}-m_{22})!(m_{23}-m_{22})!(m_{12}-m_{22}+1)}{(m_{12}-m_{33}+1)!(m_{22}-m_{33})!(m_{13}-m_{22}+1)!(m_{13}-m)!(m_{12}-m_{11})!(m-m_{23})!}\right. \\ &\quad \left.\times \frac{1}{(m_{12}-m_{23})!(m_{13}-m_{12})!(m-m_{11})!(w+m_{11}-m)!}\right]^{\frac{1}{2}} \\ &\times {}_3F_2\left(\begin{matrix} m-m_{13} & m_{22}-m_{13}-1 & m_{12}-m_{13} \\ & m_{23}-m_{13} & m_{11}-m_{13} \end{matrix}\right). \end{aligned}$$

We expect this expression, from consideration of the general form (3.6), to be proportional to the Clebsch–Gordan coefficient that couples the following states:

(4.24)
$$\left|\begin{matrix} 0 & -w \\ m_{11}-m & \end{matrix}\right\rangle, \qquad \left|\begin{matrix} m_{13} & m_{23} \\ m & \end{matrix}\right\rangle \to \left|\begin{matrix} m_{12} & m_{22} \\ m_{11} & \end{matrix}\right\rangle.$$

The Clebsch–Gordan coefficient that performs this coupling is $C^{j_1 j_2 j}_{m_1 m_2 m_1+m_2}$, where

(4.25)
$$\begin{aligned} &j_1 = \tfrac{1}{2}(m_{13}-m_{23}), \quad j_2 = \tfrac{1}{2}w, \quad j = \tfrac{1}{2}(m_{12}-m_{22}), \\ &m_1 = m - \tfrac{1}{2}(m_{13}+m_{23}), \qquad m_2 = m_{11} - m + \tfrac{1}{2}w. \end{aligned}$$

A standard expression for the Clebsch–Gordan coefficient is given by the van der Waerden form, which can be expressed as an ${}_3F_2$ function as follows (see, for example,

[30, p. 429]):

(4.26)
$$
\begin{aligned}
&C^{j_1 j_2 j}_{m_1 m_2 m_1+m_2} \\
&= \left[\frac{(m_{12}-m_{22}+1)(m_{12}-m_{33})!(m_{23}-m_{22})!(m-m_{23})!(m-m_{11})!}{(m_{13}-m_{22}+1)!(m_{13}-m)!(w+m_{11}-m)!}\right. \\
&\qquad \left.\times \frac{(m_{11}-m_{22})!(m_{12}-m_{11})!}{(m_{13}-m_{12})!}\right]^{\frac{1}{2}} \\
&\times \frac{1}{(m_{12}+m-m_{13}-m_{23})!(m_{12}+m-m_{13}-m_{11})!} \\
&\times {}_3F_2\begin{pmatrix} m_{12}-m_{13} & m-m_{13} & -w-m_{11}+m \\ m_{12}-m_{13}-m_{23}+m+1 & m_{12}-m_{11}+m-m_{13}+1 & \end{pmatrix},
\end{aligned}
$$

where we used the definitions given in (4.25). The ${}_3F_2$ function in (4.23) is not in the form (4.26); however, there is a direct transformation between the two forms, which appears in Bailey [29, p. 85] and may be written

(4.27)
$$
\begin{aligned}
{}_3F_2\begin{pmatrix} a & b & c \\ e & f & \end{pmatrix} &= \frac{(a-f)!(c-f)!(c-e)!(a-e)!}{(-e)!(-f)!(a+c-e)!(a+c-f)!} \\
&\quad \times {}_3F_2\begin{pmatrix} a & c & a+b+c-e-f+1 \\ a+c-f+1 & a+c-e+1 & \end{pmatrix},
\end{aligned}
$$

where c is a negative integer (the parameters a, b, e, f here differ from those defined in (4.20)). Now, if we substitute into (4.27) for the parameters according to

(4.28)
$$
\begin{aligned}
&a = m - m_{13}, \qquad b = m_{22} - m_{13} - 1, \\
&c = m_{12} - m_{13}, \quad e = m_{23} - m_{13}, \quad f = m_{11} - m_{13},
\end{aligned}
$$

then we transform the ${}_3F_2$ function in (4.23) into the form (4.26) for the Clebsch–Gordan coefficient, and consequently express the BW state in precisely the required form (4.3).

5. The BW realization of $U_q(3)$. It is known [31], [32] that (for generic real q) all unitary irreps of the quantum groups $U_q(n)$ are finite-dimensional and in one-to-one correspondence with those of $U(n)$. Similarly, the basis states are in one-to-one correspondence with the Gel'fand–Weyl basis states for the unitary groups, and so the same set of labels (m) can be used to label the quantum group states. In constructing q-BW states, analogous to (4.1) and (4.2) for $U(3)$, we need to determine only the appropriate q-extension of the exponential factor and its argument in (4.2). The algebra (2.4) for $U_q(3)$ is specified in terms of elements corresponding to the simple roots, namely, E_{12} and E_{23}, and we require an appropriate definition of E_{13} in order to generalize (4.2). We choose, as in [6],

$$
E_{13} = q^{-E_{22}/2}(E_{12}E_{23} - q^{-\frac{1}{2}}E_{23}E_{12}); \tag{5.1}
$$

then, as a consequence of the defining relations (2.4) and (2.6), we have

$$
[E_{13}, E_{23}] = 0 = [E_{13}, E_{12}]. \tag{5.2}
$$

(In fact these relations are equivalent to the Serre relations (2.6).) Similarly, we choose the operator E_{31} according to

$$E_{31} = q^{E_{22}/2}(E_{32}E_{21} - q^{\frac{1}{2}}E_{21}E_{32}), \tag{5.3}$$

which satisfies

$$[E_{31}, E_{21}] = 0 = [E_{31}, E_{32}]. \tag{5.4}$$

We can now define the $U_q(3)$ BW states by

$$\left| \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22} \\ & & m_{11} \end{matrix} \right\rangle_{q,BW} = \sum_m E_m^q((m); a^q)|0\rangle \otimes \left| \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{13} & m_{23} \\ & & m \end{matrix} \right\rangle, \tag{5.5}$$

where

$$E_m^q((m); a^q) = \left\langle \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{13} & m_{23} \\ & & m \end{matrix} \right| \exp_q(a_1^q E_{13}) \exp_q(a_2^q E_{23}) \left| \begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22} \\ & & m_{11} \end{matrix} \right\rangle, \tag{5.6}$$

and where the q-exponential is defined in (2.13). The product of exponentials can be combined using a q-addition formula, which is obtained with the help of the q-binomial theorem (2.24), which we may write as

$$\exp_q(a_1^q E_{13}) \exp_q(a_2^q E_{23})\,|0\rangle = \exp_q\left(q^{-N_2/2} a_1^q E_{13} + q^{N_1/2} a_2^q E_{23}\right)|0\rangle. \tag{5.7}$$

An outline of the proof of this formula is as follows. Define the operators a, b by

$$a = q^{-N_2/2} a_1^q, \qquad b = q^{N_1/2} a_2^q, \tag{5.8}$$

where N_1, N_2 are the number operators for the q-boson operators a_1^q, a_2^q, respectively. Then $ba = qab$, and we may apply the q-binomial theorem (2.24) to each of the series expansions of the q-exponentials on the left-hand side of (5.7) to obtain the right-hand side after applying the number operators to the vacuum $|0\rangle$.

We have now defined q-analogs of the BW states for $n = 3$, and we next establish the $n = 3$ version of Lemma 3.1, i.e., we calculate the realization $\Gamma(E_{ij})$ of the $U_q(3)$ generators and verify that these operators satisfy the algebra of $U_q(3)$. The results are summarized in the following lemmas.

LEMMA 5.1. *The realization $\Gamma(E_{ij})$ of the $U_q(3)$ generators acting on the basis $|(m)\rangle_{q,BW}$ is given by*

$$\begin{aligned}
\Gamma(E_{11}) &= E_{11} - N_1, \\
\Gamma(E_{22}) &= E_{22} - N_2, \\
\Gamma(E_{33}) &= E_{33} + N_1 + N_2, \\
\Gamma(E_{13}) &= \bar{a}_1^q, \\
\Gamma(E_{23}) &= \bar{a}_2^q, \\
\Gamma(E_{12}) &= q^{N_2/2} E_{12} - q^{(E_{22}+1)/2}\, a_2^q\, \bar{a}_1^q, \\
\Gamma(E_{21}) &= q^{-N_1/2} E_{21} - q^{-(E_{11}+1)/2}\, a_1^q\, \bar{a}_2^q, \\
\Gamma(E_{32}) &= q^{-E_{33}/2} a_1^q E_{12} + a_2^q[E_{22} - E_{33} - N_1 - N_2], \\
\Gamma(E_{31}) &= q^{E_{33}/2} a_2^q E_{21} + a_1^q[E_{11} - E_{33} - N_1 - N_2].
\end{aligned} \tag{5.9}$$

Proof. The expressions for $\Gamma(E_{13})$, $i = 1, 2$ are immediate since $\bar{a}_i^q \exp_q(a_i^q E_{i3}) \mid 0\rangle = \exp_q(a_i^q E_{i3}) E_{i3} \mid 0\rangle$ $(i = 1, 2)$, as indicated in (2.14), and $[E_{13}, E_{23}] = 0$. The expressions for $\Gamma(E_{ii})$, $i = 1, 2, 3$ follow in the same way as for $U_q(2)$, using

$$[E_{11} - N_1, a_1^q E_{13}] = 0 = [E_{11} - N_1, a_2^q E_{23}]$$

and

$$[E_{33} + N_1 + N_2, a_1^q E_{13}] = 0 = [E_{33} + N_1 + N_2, a_2^q E_{23}].$$

Next, we require the formula

$$E_{23}^n E_{12} = q^{n/2} E_{12} E_{23}^n - [n] q^{(E_{22}+1)/2} E_{13} E_{23}^{n-1}, \tag{5.10}$$

which is proved by induction on n. From it follows

$$\exp_q(a_2^q E_{23}) E_{12} = E_{12} q^{N_2/2} \exp_q(a_2^q E_{23}) q^{-N_2/2} - q^{(E_{22}+1)/2} E_{13} a_2^q \exp_q(a_2^q E_{23}).$$

We multiply this operator equation on the right by $\exp_q(a_1^q E_{13})$, replace $E_{13} \exp_2(a_2^q E_{23})$ by $\bar{a}_1^q \exp_q(a_2^q E_{23})$, and let the result act on the vacuum to give the required expression for $\Gamma(E_{12})$. Similarly, beginning with the formula

$$E_{13}^n E_{21} = q^{-n/2} E_{21} E_{13}^n - q^{-(E_{11}+1)/2} [n] E_{23} E_{13}^{n-1}, \tag{5.11}$$

we obtain $\Gamma(E_{21})$. In order to calculate $\Gamma(E_{32})$, we first prove

$$E_{13}^n E_{32} = q^{n/2} E_{32} E_{13}^n + q^{-E_{33}/2} [n] E_{12} E_{13}^{n-1} \tag{5.12}$$

by induction on n. Hence

$$\exp_q(a_1^q E_{13}) E_{32} = E_{32} q^{N_1/2} \exp_q(a_1^q E_{13}) q^{-N_1/2} + q^{-E_{33}/2} E_{12} a_1^q \exp_q(a_1^q E_{13}).$$

Next we use

$$\exp_q(a_2^q E_{23}) E_{32} = (E_{32} + a_2^q [E_{22} - E_{33} - N_2]) \exp_q(a_2^q E_{23}),$$

which follows from

$$E_{23}^n E_{32} = E_{32} E_{23}^n + [E_{22} - E_{33} - n + 1][n] E_{23}^{n-1}. \tag{5.13}$$

By combining these formulas, we get $\Gamma(E_{32})$ as required, and $\Gamma(E_{31})$ is obtained by using (5.3). □

LEMMA 5.2. *The map* $\Gamma : g \to U_q(g)$, *where* g *is a generator of* $U_q(3)$, *given by* (5.9), *is an isomorphism of the algebra* $U_q(3)$.

Proof. We first check that the definition (5.1) of E_{13} is preserved under the mapping Γ, i.e., we verify

$$\Gamma(E_{13}) = q^{-\Gamma(E_{22})/2} \left(\Gamma(E_{12}) \Gamma(E_{23}) - q^{-\frac{1}{2}} \Gamma(E_{23}) \Gamma(E_{12}) \right), \tag{5.14}$$

and this follows directly upon using the definition (2.9) for the q-boson operator a_2^q. It is also immediate that

$$[\Gamma(E_{13}), \Gamma(E_{23})] = 0 = [\Gamma(E_{13}), \Gamma(E_{12})]. \tag{5.15}$$

We also find

$$
\begin{aligned}
[\Gamma(E_{12}),\Gamma(E_{21})] &= q^{-(N_1-N_2)/2}[E_{12},E_{21}] - q^{(E_{22}-E_{11})/2}\,[a_1^q\bar{a}_2^q, a_2^q\bar{a}_1^q] \\
&= [\Gamma(E_{11}) - \Gamma(E_{22})],
\end{aligned} \tag{5.16}
$$

as required, and

$$
\begin{aligned}
[\Gamma(E_{23}),\Gamma(E_{32})] &= \bar{a}_2^q a_2^q[E_{22} - E_{33} - N_1 - N_2] - a_2^q\bar{a}_2^q[E_{22} - E_{33} - N_1 - N_2 + 1] \\
&= [E_{22} - E_{33} - N_1 - 2N_2],
\end{aligned} \tag{5.17}
$$

again as required, where we used (2.15) and (3.18) with $a = E_{22} - E_{33} - N_1 - N_2, b = N_2, c = 1$. □

6. Explicit BW states for $U_q(3)$. In this section we undertake the q-analog of the calculation in §4 in order to obtain the $U_q(3)$ BW states explicitly, in a form analogous to (3.6). For this we require various identities satisfied by basic hypergeometric functions, which arise naturally within the calculation. Whereas the required matrix elements can be written in terms of ${}_6F_5$ functions for $U(3)$ (see (4.21)), we find that for $U_q(3)$ they can be expressed as a certain limit of basic hypergeometric functions ${}_8\phi_7$, which are then reduced to ${}_3\phi_2$ functions by using known identities. These ${}_3\phi_2$ functions can in turn be transformed into a standard form recognizable as q-Clebsch–Gordan coefficients, enabling us to express the q-BW states as a product of a K_q-factor and a sum over a polynomial in a_1^q, a_2^q, with a q-Clebsch–Gordan coupling, tensored with the Gel'fand–Weyl states.

The q-BW states are defined by (5.5) and (5.6), and our first task is to evaluate the matrix elements in (5.6). The matrix elements of the simple roots E_{12}, E_{23} of the $U_q(3)$ algebra are identical to those of $U(3)$, but with the parentheses in (3.10) and (4.7) replaced by the square brackets defined in (2.3); compare, for example, (3.10) and (3.13). For general n a proof has been provided by Ueno et al. [33]. Hence we can calculate the matrix elements of E_{13} from those of E_{12}, E_{23} by using (5.1), and we find that

$$
\begin{aligned}
E_{13}\left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22} \\ & & m_{11} \end{matrix}\right\rangle
&= q^{-m_{22}/2}\sqrt{\frac{[m_{13}-m_{12}][m_{12}-m_{23}+1][m_{12}-m_{33}+2][m_{11}+1-m_{22}]}{[m_{12}-m_{22}+2][m_{12}-m_{22}+1]}} \\
&\quad\times\left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12}+1 & m_{22} \\ & & m_{11}+1 \end{matrix}\right\rangle \\
&\quad - q^{-(m_{12}+1)/2}\sqrt{\frac{[m_{13}-m_{22}+1][m_{23}-m_{22}][m_{22}+1-m_{33}][m_{12}-m_{11}]}{[m_{12}-m_{22}+1][m_{12}-m_{22}]}} \\
&\quad\times\left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22}+1 \\ & & m_{11}+1 \end{matrix}\right\rangle .
\end{aligned} \tag{6.1}
$$

From the matrix elements for E_{23} and E_{13} we can calculate the coefficients $A^n_r(m)$ and $B^\ell_s(m)$, which are defined as in (4.8) and (4.12), respectively. We do this by obtaining recurrence relations analogous to (4.9) and (4.13), and for A^n_r the only change is to replace parentheses $(\ldots)$ by brackets $[\ldots]$ to obtain the solution analogous to (4.10), and the recurrence relations are satisfied upon using the identity (3.18). The coefficients B^ℓ_s involve explicit q-factors, as is apparent from the matrix elements (6.1), and satisfy

(6.2)
$$\begin{aligned} B^{\ell+1}_s &= B^\ell_{s-1} q^{-(m_{23}-\ell+s-1)/2} \\ &\quad \cdot \left(\frac{[s][m_{13}-m_{23}-s+1][m_{13}-m_{33}-s+2][m_{11}-m_{23}-s+1]}{[m_{13}-m_{23}+\ell-2s+3][m_{13}-m_{23}+\ell-2s+2]} \right)^{\frac{1}{2}} \\ &\quad - B^\ell_s q^{-(m_{13}-s+1)/2} \left(\frac{[m_{13}-m_{23}+\ell-s+2][\ell-s+1]}{[m_{13}-m_{23}+\ell-2s+2]} \right. \\ &\qquad \left. \times \frac{[m_{23}-m_{33}-\ell+s][m_{13}-m_{11}+\ell-s+1]}{[m_{13}-m_{23}+\ell-2s+1]} \right)^{\frac{1}{2}}. \end{aligned}$$

The solution is

(6.3)
$$\begin{aligned} B^\ell_s &= q^{-(\ell-s)(m_{13}-s)/2} q^{-s(m_{23}-1)/2} q^{-\ell/2} (-)^{\ell+s} [\ell]! \\ &\quad \times \left(\frac{[m_{13}-m_{33}+1]![m_{11}-m_{23}]![m_{13}-m_{23}-s]![m_{13}-m_{11}+\ell-s]!}{[s]![\ell-s]![m_{13}-m_{23}+\ell+1-s]![m_{13}-m_{33}+1-s]![m_{11}-m_{23}-s]!} \right. \\ &\qquad \left. \times \frac{[m_{23}-m_{33}]![m_{13}-m_{23}+\ell-2s+1]}{[m_{13}-m_{11}]![m_{23}-m_{33}-\ell+s]!} \right)^{\frac{1}{2}}, \end{aligned}$$

which is verified by direct substitution, using (3.18) with $a = \ell+1$, $b = m_{13}-m_{23}+\ell-s+2$, $c = -s$.

The matrix element $E^q_m((m); a^q)$ is therefore given by the q-analog of (4.18) (i.e., replace $(\ldots)$ by $[\ldots]$), together with the multiplicative q-factor

$$q^{-s(s+a)/2-(m_{13}+1)(m-m_{11})/2}, \tag{6.4}$$

which comes from the coefficient B^ℓ_s in (6.3), and where a is defined in (4.20). The terms involving the summation s, analogous to $R(s)$ as given in (4.19), are

$$R(s) = \frac{[a+2s]([a])_s([b])_s([c])_s([d])_s([e])_s(-1)^s q^{-s(a+s)/2}}{[s]![a]([1+a-b])_s([1+a-c])_s([1+a-d])_s([1+a-e])_s}, \tag{6.5}$$

where a, b, c, d, e are given by (4.20).

The function $\sum_s R(s)$ is the q-analog of the ${}_6F_5$ function appearing in (4.21), and is most conveniently expressed as the limit of an ${}_8\phi_7$ function, specifically

(6.6)
$$\begin{aligned} &\sum_s R(s) \\ &= \lim_{N\to\infty} {}_8\phi_7 \left(\begin{matrix} q^a, q^{1+a/2}, -q^{1+a/2}, q^b, q^c, q^d, q^e, q^{-N} \\ q^{a/2}, -q^{a/2}, q^{1+a-b}, q^{1+a-c}, q^{1+a-d}, q^{1+a-e}, q^{1+a+N} \end{matrix} ; q, \frac{q^{2a+2+N}}{q^{b+c+d+e}} \right) \\ &= \lim_{N\to\infty} \sum_s \frac{[a+2s]([a])_s([b])_s([c])_s([d])_s([e])_s([-N])_s}{[s]![a]([1+a-b])_s([1+a-c])_s([1+a-d])_s([1+a-e])_s([1+a+N])_s}, \end{aligned}$$

where we used the definition (2.20) of a basic hypergeometric function, and the identity

$$\frac{(q^{1+a/2};q)_s(-q^{1+a/2};q)_s}{(q^{a/2};q)_s(-q^{a/2};q)_s} = q^s\frac{[a+2s]}{[a]}. \tag{6.7}$$

The limit $N \to \infty$ in (6.6) is taken using

$$\lim_{N\to\infty}\frac{([-N])_s}{([1+a+N])_s} = (-)^s q^{-s(s+a)/2} \qquad (q>1). \tag{6.8}$$

Although $\sum_s R(s)$ can be expressed in a certain way as a ${}_7\phi_6$ function, the form (6.6) is appropriate for our purposes because we can now use a formula due to Watson [34] that reduces a terminating, very well-poised ${}_8\phi_7$ series to a terminating balanced ${}_4\phi_3$ series, and is the q-analog of Whipple's formula, which transforms a well-poised ${}_7F_6$ into a Saalschützian ${}_4F_3$. By using this formula we can reduce $\sum_s R(s)$ to a ${}_3\phi_2$ function, which is identifiable with a q-Clebsch–Gordan coefficient. Watson's formula states (see Bailey [29, p. 69] or Gasper and Rahman [26, p. 35])

(6.9)

$$\begin{aligned}
&{}_8\phi_7\left(\begin{matrix} q^a, & q^{1+a/2}, & -q^{1+a/2}, & q^b, & q^c, & q^d, & q^e, & q^{-N} \\ q^{a/2}, & -q^{a/2}, & q^{1+a-b}, & q^{1+a-c}, & q^{1+a-d}, & q^{1+a-e}, & q^{1+a+N} & \end{matrix}\ ;\ q, \frac{q^{2a+2+N}}{q^{b+c+d+e}}\right) \\
&= \frac{(q^{1+a};q)_N(q^{1+a-d-e};q)_N}{(q^{1+a-d};q)_N(q^{1+a-e};q)_N}\ {}_4\phi_3\left(\begin{matrix} q^d & q^e & q^{1+a-b-c} & q^{-N} \\ & q^{1+a-b} & q^{1+a-c} & q^{e+d-a-N} \end{matrix}\ ;\ q, q\right).
\end{aligned}$$

In Bailey's notation we have replaced a by q^a, c by q^b, d by q^c, e by q^d, f by q^e, and g by q^{-N}, where N is a positive integer. In our notation Watson's formula reads

$$\begin{aligned}
&\sum_s \frac{[a+2s]([a])_s([b])_s([c])_s([d])_s([e])_s([-N])_s}{[s]![a]([1+a-b])_s([1+a-c])_s([1+a-d])_s([1+a-e])_s([1+a+N])_s} \\
&= \frac{([1+a])_N([1+a-c-e])_N}{([1+a-c])_N([1+a-e])_N} \\
&\quad\times \sum_s \frac{([1+a-b-d])_s([c])_s([e])_s([-N])_s}{[s]!([1+a-b])_s([1+a-d])_s([c+e-a-N])_s}.
\end{aligned} \tag{6.10}$$

Watson used this formula to prove the Rogers–Ramanujan identities, by taking suitable limits of the parameters; it also implies a general summation formula due to Jackson (see Gasper and Rahman [26, §§2.6 and 2.7]).

We now let $N \to \infty$ in (6.10) and the left-hand side reduces, according to (6.6), to $\sum_s R(s)$, while the multiplicative factors on the right-hand side become

$$(-)^e q^{-ec/2}\frac{[-a-1]![a-c]!}{[a-c-e]![-a+e-1]!} \qquad (q>1) \tag{6.11}$$

(in which positivity of the factorial arguments follows from the definitions of a, c, e in (4.20) and the inequalities satisfied by the Gel'fand–Weyl labels m_{ij}). We also use the limit

$$\lim_{N\to\infty}\frac{[\alpha-N]_s}{[\beta-N]_s} = q^{(\beta-\alpha)/2} \qquad (q>1) \tag{6.12}$$

and hence find

$$\sum_s R(s) = (-)^e q^{-ec/2} \frac{[-a-1]![a-c]!}{[a-c-e]![-a+e-1]!} \cdot \sum_s \frac{([1+a-b-d])_s([c])_s([e])_s q^{-s(e+c-a)/2}}{[s]!([1+a-b])_s([1+a-d])_s}, \tag{6.13}$$

which is the desired expression relating $\sum_s R(s)$ to a ${}_3\phi_2$ function.

Upon combining all terms, the matrix element $E^q_m((m); a^q)$ is now found to be

$$\begin{aligned}
&\sum_m (a_1^q)^{m-m_{11}} (a_2^q)^{w+m_{11}-m} (-)^{m_{12}-m_{13}+m-m_{11}} \\
&\quad \cdot q^{-(m_{12}-m_{13})(m_{22}-m_{13}-1)/2} q^{-(m_{13}+1)(m-m_{11})/2} \\
&\quad \times \frac{[m_{13}-m_{22}]![m_{13}-m_{11}]!}{[m-m_{11}]![w-m+m_{11}]!} \\
&\quad \times \left(\frac{[m_{13}-m_{33}+1]![m_{23}-m_{33}]![m_{11}-m_{22}]![m_{23}-m_{22}]!}{[m_{22}-m_{33}]![m_{12}-m_{33}+1]![m_{13}-m]![m_{12}-m_{11}]![m_{12}-m_{23}]!} \right. \\
&\qquad \left. \times \frac{[m_{12}-m_{22}+1]}{[m-m_{23}]![m_{13}-m_{12}]![m_{13}-m_{22}+1]!} \right)^{\frac{1}{2}} \\
&\quad \times {}_3\phi_2 \left(\begin{matrix} q^{m-m_{13}} & q^{m_{22}-m_{13}-1} & q^{m_{12}-m_{13}} \\ & q^{m_{23}-m_{13}} & q^{m_{11}-m_{13}} \end{matrix} ; q^{-1}, q^{-1} \right).
\end{aligned} \tag{6.14}$$

We wish to identify the ${}_3\phi_2$ function with that appearing in the q-Clebsch–Gordan coefficients; a standard form for these coefficients, the q analog of the van der Waerden form (4.26), has been calculated in [11], and can be expressed in terms of ${}_3\phi_2$ functions as follows [12].

$$\begin{aligned}
&{}_qC^{j_1 j_2 j}_{m_1 m_2 m} = \delta_{m,m_1+m_2} q^{(j_1+j_2-j)(j_1+j_2+j+1)/4+(j_1 m_2 - j_2 m_1)/2} \\
&\quad \times \left(\frac{[2j+1][j_1+m_1]![j_2-m_2]![j+m]![j-m]!}{[j_1+j_2+j+1]![j_1+j_2-j]![j-j_1+j_2]![j+j_1-j_2]![j_2+m_2]![j_1-m_1]!} \right)^{\frac{1}{2}} \\
&\quad \times ([j-j_2+m_1+1])_{j_1-m_1} ([j-j_1-m_2+1])_{j_2+m_2} \\
&\quad \times {}_3\phi_2 \left(\begin{matrix} q^{-j_1-j_2+j} & q^{-j_1+m_1} & q^{j_2-m_2} \\ & q^{j-j_2+m_1+1} & q^{j-j_1-m_2+1} \end{matrix} ; q, q \right).
\end{aligned} \tag{6.15}$$

We identify the parameters j_1, j_2, j, m_1, m_2 as in (4.25), and we seek to express the coefficients in the form shown in (6.14); in order to do this we require the following transformation [13, eq. (1.30)], [26, eq. (3.2.2)]:

$${}_3\phi_2 \left(\begin{matrix} q^{-n} & \alpha & \beta \\ & \gamma & \delta \end{matrix} ; q, q \right) = \alpha^n \frac{(\delta/\alpha; q)_n}{(\delta; q)_n} {}_3\phi_2 \left(\begin{matrix} q^{-n} & \alpha & \gamma/\beta \\ & \gamma & \alpha q^{1-n}/\delta \end{matrix} ; q, \beta q/\delta \right), \tag{6.16}$$

where n is a positive integer. Putting $\alpha = q^a$, $\beta = q^b$, $\gamma = q^e$, $\delta = q^f$, we can write

this as

(6.17)

$$\begin{aligned}
&\sum_s \frac{q^{-s(a+b-n-e-f+1)/2}([a])_s([b])_s([-n])_s}{[s]!([e])_s([f])_s} \\
&\quad = q^{-an/2}\frac{([f-a])_n}{([f])_n}\sum_s \frac{q^{-s(b-f)/2}([a])_s([e-b])_s[-n])_s}{[s]!([e])_s([a+1-n-f])_s} \\
&\quad = \frac{([f-a])_n([e-a])_n}{([e])_n([f])_n}\sum_s \frac{q^{-sb/2}([a])_s([a+b+1-n-e-f])_s([-n])_s}{[s]!([a+1-n-f])_s([a+1-n-e])_s},
\end{aligned}$$

where we used the identity a second time in the last step. Hence we have

(6.18)

$$\begin{aligned}
&{}_3\phi_2\left(\begin{matrix} q^a & q^b & q^{-n} \\ & q^e & q^f \end{matrix}; q^{-1}, q^{-1}\right) \\
&\quad = \frac{([f-a])_n([e-a])_n}{([e])_n([f])_n}\,{}_3\phi_2\left(\begin{matrix} q^a & q^{a+b+1-n-e-f} & q^{-n} \\ & q^{a+1-n-e} & q^{a+1-n-f} \end{matrix}; q^{-1}, q^{-1}\right).
\end{aligned}$$

This relation is the q-analog of (4.27), with $c = -n$ and, with the parameters identified as in (4.28), enables us to express the ${}_3\phi_2$ function in (6.14) in terms of the ${}_3\phi_2$ function in (6.15), i.e., we can now explicitly identify the q-Clebsch–Gordan coefficients in the matrix elements $E^q_m((m); a^q)$. Upon collecting all factors, we find that

(6.19)

$$\begin{aligned}
\left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{12} & m_{22} \\ & & m_{11} \end{matrix}\right\rangle_{q,BW} &= (-)^{m_{22}-m_{23}} K_q\left(\begin{matrix}[\mathbf{m}_3] \\ [\mathbf{m}_2]\end{matrix}\right)\sum_m \frac{(a_1^q)^{m-m_{11}}(-a_2^q)^{w+m_{11}-m}}{\sqrt{[m-m_{11}]![w+m_{11}-m]!}} \\
&\quad \times q^{\mu/4}{}_{q^{-1}}C^{j_1 j_2 j}_{m_1 m_2 m_2+m_2}\,|0\rangle \otimes \left|\begin{matrix} m_{13} & m_{23} & m_{33} \\ & m_{13} & m_{23} \\ & & m \end{matrix}\right\rangle,
\end{aligned}$$

where the K_q-factor is given by

$$K_q\left(\begin{matrix}[\mathbf{m}_3] \\ [\mathbf{m}_2]\end{matrix}\right)^2 = \frac{[m_{13}-m_{33}+1]![m_{23}-m_{33}]!}{[m_{12}-m_{33}+1]![m_{22}-m_{33}]!} \tag{6.20}$$

and

(6.21)

$$\mu = -(m_{13}-m_{12})(m_{13}-m_{22}+1) - (3m_{13}-m_{23}+2)(m-m_{11}) + w(m_{13}-m).$$

This form for the general $U_q(3)$ basis vector is a special case of general results we have obtained [6] for the form of the BW basis vectors for $U_q(n)$. The origin of the various factors is explained in [6]; however, as discussed in the Introduction, our calculation here explicitly shows how the theory of special functions and their q-analogs is interwoven with properties of the representations of the classical Lie groups and their q-analogs, the quantum groups.

REFERENCES

[1] L. C. BIEDENHARN AND M. A. LOHE, *Quantum Groups and Basic Hypergeometric Functions*, Proceedings of the Argonne Workshop on Quantum Groups, T. Curtright, D. Fairlie, and C. Zachos, eds., World Scientific, Singapore, 1990, pp. 123–132.

[2] V. G. DRINFELD, *Quantum Groups*, Proc. Int. Congr. of Math., Berkeley, CA, 1986, pp. 798–820; *A new realisation of Yangians and quantised affine algebras*, Sov. Math. Dokl., 36 (1988), pp. 212–216.

[3] M. JIMBO, *A q-analogue of $U(gl(N+1))$, Hecke algebra, and the Yang–Baxter equation*, Lett. Math. Phys., 11 (1986), pp. 247–252.

[4] L. C. BIEDENHARN, *An Overview of Quantum Groups*, 18th International Colloquium on Group Theoretical Methods in Physics, Moscow, Russia, June 1990, Lecture Notes in Phys., 382, Springer-Verlag, New York.

[5] L. C. BIEDENHARN AND M. A. LOHE, *Induced Representations and Tensor Operators for Quantum Groups*, Proceedings of the Quantum Group Workshop, Euler International Mathematical Institute, Leningrad, October 1990, to appear.

[6] ———, *An extension of the Borel–Weil construction to the quantum group $U_q(n)$*, Comm. Math. Phys., 146 (1992), pp. 483–504.

[7] N. J. VILENKIN, *Special functions and the theory of group representations*, Amer. Math. Soc. Transl., 22 (1968).

[8] E. P. WIGNER, *Application of group theory to the special functions of mathematical physics*, Princeton University, Princeton, NJ, 1955, unpublished lecture notes.

[9] J. TALMAN, *Special Functions: A Group Theoretical Approach*, W. A. Benjamin, New York, 1968.

[10] L. C. BIEDENHARN, R. S. GUSTAFSON, M. A. LOHE, J. D. LOUCK, AND S. C. MILNE, *Special functions and group theory in theoretical physics*, in Special Functions, Group Theoretical Aspects and Applications, R. Askey, T. H. Koornwinder, and W. Schempp, eds., Riedel, New York, 1984, pp. 129–162.

[11] A. N. KIRILLOV AND N. YU RESHITIKHIN, *Representations of the algebra $U_q(sl(2))$, q-orthogonal polynomials and invariants of links*, USSR Academy of Sciences, Leningrad, preprint, 1988.

[12] L. C. BIEDENHARN AND M. A. LOHE, *Symmetries of Quantum Coupling Coefficients*, Differential Geometry, Group Representations, and Quantization, J. D. Hennig, W. Lücke, and J. Tolar, eds., Lecture Notes in Phys., 379, Springer-Verlag, New York, 1990, pp. 193–206.

[13] R. ASKEY AND J. A. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985).

[14] R. LEBLANC AND L. C. BIEDENHARN, *Implementation of the $U(n)$ tensor operator calculus in a vector Bargmann Hilbert space*, J. Phys. A. Math. Gen., 22 (1989), pp. 31–48.

[15] C. N. YANG AND M. L. GE, EDS., *Braid Group, Knot Theory and Statistical Mechanics*, World Scientific, Singapore, New Jersey, London, Hong Kong, 1989.

[16] L. C. BIEDENHARN AND J. D. LOUCK, *Angular momentum in quantum physics, theory and application*, in Encyclopedia of Mathematics and its Applications, Vol. 8, Addison-Wesley, Reading, MA, 1981.

[17] L. C. BIEDENHARN, *The quantum group $SU_q(2)$ and a q-analogue of the boson operators*, J. Phys. A. Math. Gen., 22 (1989), pp. L873–L878.

[18] A. J. MACFARLANE, *On q-analogues of the quantum harmonic oscillator and the quantum group $SU_q(2)$*, J. Phys. A. Math. Gen., 22 (1989), pp. 4581–4588.

[19] F. H. JACKSON, *Generalization of the differential operative symbol with an extended form of Boole's equation*, Messenger Math., 38 (1909), pp. 57–61; *q-Form of Taylor's Theorem*, Messenger Math., 38 (1909), pp. 62–64; *On q-definite integrals*, Quart. J. Pure Appl. Math., 41 (1910), pp. 193–203.

[20] G. E. ANDREWS, *q-Series: Their Development and Application in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra*, Conference Board of the Mathematical Sciences 66, American Mathematical Society, Providence, RI.

[21] S. C. MILNE, *A q-Analog of the Gauss Summation Theorem for Hypergeometric Series in $U(n)$*, Adv. Math., 72 (1988), pp. 59–131.

[22] T. H. KOORNWINDER, *Representations of the twisted $SU(2)$ quantum group and some q-hypergeometric orthogonal polynomials*, Nederl. Akad. Wetensch. Proc. Ser., A92 (1989), p. 97.

[23] P. FEINSILVER, *Commutators, anti-commutators and Eulerian calculus*, Rocky Mountain J. Math., 12 (1982), pp. 171–183.

[24] J. CIGLER, *Operatormethoden für q-Identitäten*, Monatsh. Math., 88 (1979), pp. 87–105.

[25] E. HEINE, *Über die Reihe ...*, J. Reine Angew. Math., 32 (1846), pp. 210–212; *Untersuchungen über die Reihe ...*, J. Reine Angew. Math., 34 (1847), pp. 285–328.

[26] G. GASPER AND M. RAHMAN, *Basic Hypergeometric Series*, Encyclopedia of Mathematics and its Applications, 35, Cambridge University Press, London, 1990.

[27] I. M. GEL'FAND AND M. L. ZETLIN, *Finite-dimensional representations of the group of unimodular matrices*, Dokl. Akad. Nauk SSSR, 71 (1950), pp. 825–828.

[28] G. E. BAIRD AND L. C. BIEDENHARN, *On the Representations of the Semisimple Lie Groups. II*, J. Math. Phys., 4 (1963), pp. 1449–1466.

[29] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935.

[30] L. C. BIEDENHARN AND J. D. LOUCK, *The Racah–Wigner Algebra in Quantum Theory*, Encyclopedia of Mathematics and its Applications, Vol. 9, Addison-Wesley, Reading, MA, 1981.

[31] G. LUSZTIG, *Quantum deformation of certain simple modules over enveloping algebras*, Adv. Math., 70 (1988), pp. 237–249.

[32] M. ROSSO, *Finite dimensional representations of the quantum analog of the enveloping algebra of a complex simple Lie algebra*, Comm. Math. Phys., 117 (1988), pp. 581–593.

[33] K. UENO, T. TAKEBAYASHI, AND Y. SHIBUKAWA, *Gelfand–Zetlin basis for $U_q(gl(N+1))$ modules*, Lett. Math. Phys., 18 (1989), pp. 215–221.

[34] G. N. WATSON, *A new proof of the Rogers–Ramanujan identities*, J. London Math. Soc., 4 (1929), pp. 4–9.

DEDICATION

This special issue of the *SIAM Journal on Mathematical Analysis* is dedicated to our dear friends Richard A. Askey on the occasion of his sixtieth birthday on June 4, 1993, and Frank W. J. Olver on the upcoming occasion of his seventieth birthday on December 15, 1994.

Frank Olver was the Managing Editor of this journal from its inception in 1970 until the end of 1974, and he has been on its Editorial Board ever since. He was on the Editorial Boards for the *SIAM Journal on Numerical Analysis* (1964–69), *NBS Journal of Research* (1966–78) and the Springer-Verlag series *Handbook for Automatic Computation* (1959–73). In addition, he has been an Associate Editor for *Mathematics of Computation* since 1984, and in 1992 he joined the Editorial Board for the new journal *Methods and Applications of Analysis.*

According to his son Peter, Frank's mathematical career started in World War II while he was working for the British Nautical Almanac Office, just after he received his bachelor's degree from the University of London. One day he had finished whatever he was doing and was wandering in the room. He paused behind one of his coworkers who was carefully adding up an infinite series. Frank's eyes widened when he looked at the series and he suddenly remarked "You know, that series diverges!" The other fellow momentarily turned around to say "Yes, that's right..." and then resumed his calculation. Of course, the series was an asymptotic one, and Frank's curiosity on how one could sum a divergent series led directly to his eventual mathematical career. To illustrate Frank's meticulous and rather unusual proofreading technique, Peter contributed the following anecdote. Frank was and is one of the most careful proofreaders of mathematics ever. For his book, he proofread every formula backwards (!) so as not to be lulled into skipping errors. Late at night at home you could hear him mutter unintelligible (to his kids) things like "... zero, zero, equals, equals, x, x, 3, 3, plus, plus ...". He proofread not only the galley proofs and the page proofs this way, but even when the final typeset manuscript was shipped from England to the United States, he spent a day in New York looking for a few final corrections. He once even offered Peter several dollars for any typographical error he could find in his book—more for mathematical ones. Peter has yet to collect a single dollar for this despite having asked a number of friends and colleagues!

Among Frank's most significant contributions are his example of a convergent series expansion which has *twice* itself as its own asymptotic expansion (published in Vol. 1 of this journal, pp. 533–534), rigorous exponential improvement of asymptotic expansions derived from Laplace integrals or ordinary differential equations, error bounds for a great variety of asymptotic expansions, a numerical algorithm for second-order difference equations, a new method for the evaluation of zeros of solutions of second-order differential equations and the development of a new system of computer arithmetic. His book, *Asymptotics and Special Functions*, published in 1974 and translated into Russian in 1978 (shorter version) and 1990 (full version), is considered to be the classic source on these topics. For additional comments on Frank's contributions to mathematics, see the 1990 book *Asymptotic and Computational Analysis*, edited by R. Wong, which contains the proceedings of the *International Symposium on Asymptotic and Computational Analysis* that was held in honor of Frank Olver on the occasion of his sixty-fifth birthday. Although Frank received the rank of Professor Emeritus in 1992 at the University of Maryland (in the Institute for Physical Science and Technology and the Mathematics Department), he has not retired from

his research or editorial work, and he continues to make important contributions to asymptotics, numerical analysis, and special functions.

In asymptotics, it is often not the final result but the method used to obtain the result that is the most important. This is why one will not find many well-known theorems in books on asymptotics, but instead will find methods such as Laplace's method, the principle of stationary phase, the method of steepest descent, saddle-point method, and so forth in almost every book on the subject. What makes Frank's work stand out from others is that Frank not only can come up with powerful methods but he can also formulate the final result into precise and general theorems which can be applied directly to a wide variety of problems. This is certainly evidenced in his earlier work on the construction of globally valid uniform asymptotic solutions to ordinary differential equations, and in his recent work on the exponentially-improved asymptotic solutions of ordinary differential equations. Frank is so thorough that when he finishes his investigation, there may be very little left for someone else to continue.

Dick Askey has been on the Editorial Board of this journal since its first issue in 1970. He is the Szegö Professor at the University of Wisconsin in Madison and an Honorary Fellow of the Indian Academy of Sciences. He was a Guggenheim Fellow (1969–70) and Vice President of the American Mathematical Society (1986–87). During the spring of 1992 he gave the Turán lectures in Budapest. Dick has constantly been on the look-out for areas of research in science generally (not just in mathematics) where special functions might play a significant role. As a result he has interacted with physicists, statisticians, engineers, and others in important ways so that work on special functions has become widely visible.

His research during the 1960s and early 1970s concentrated primarily on harmonic analysis and classical orthogonal polynomials, related positivity questions and inequalities. Perhaps the best testimony to the power and importance of this work lies in the fact that one of the inequalities in a joint paper with one of us (G. G.) (see *American J. Math.*, 98 (1976), p. 713, Theorem 3) plays a central role in Louis de Branges' original proof of the celebrated Bieberbach Conjecture.

In 1974, Dick presented ten lectures at a C.B.M.S. Regional Conference Lecture Series at the Virginia Polytechnic Institute which were published in his book *Orthogonal Polynomials and Special Functions.* This book helped lead to a renewed interest in orthogonal polynomials, a subject area which was more or less dormant during the 1950s and 1960s, but has been one of the liveliest areas in classical analysis ever since. Starting in the late 1960s Dick managed to get together a large group of graduate and postdoctoral students at the University of Wisconsin in Madison, many of whom ended up as internationally recognized specialists in various areas of special functions and orthogonal polynomials. In the late 1970s his work expanded to include those special functions allied with combinatorics and number theory. Recent developments in quantum groups have shown that the same functions arise there. For example, these studies led to the embedding of the "$6-j$ symbols" of physics into the theory of orthogonal polynomials. He unearthed and extended beautiful results of L. J. Rogers; this led to new families of orthogonal polynomials known as the Askey–Wilson polynomials and the "sieved" polynomials. He has also vigorously participated in the renaissance of interest in Ramanujan. In particular, he has been a leader in applying some of Ramanujan's integrals to work in hypergeometric and basic hypergeometric series.

Although Dick proudly and jokingly classifies himself as one of the last breed of 18th-century mathematicians, he is, in fact, very much a 21st-century mathematician.

He helped to keep classical analysis alive and interacting with modern mathematics. He has pursued excellence in every aspect of mathematics, including teaching, libraries and history, and has given never ending encouragement and support to younger colleagues, including those in crises (political, economic, personal, and scientific). Many of us feel that he is a bridge between the great classical analysts such as Hardy, Littlewood, Ramanujan, Pólya, and Szegö, just to name a few, and future mathematics and related areas.

Dick is also well known for his unique mathematical vocabulary which includes "marvellous" (way before Billy Crystal) and "preposterous" so that very few of his letters or recommendations and referee reports can remain truly anonymous. His observation that "a ten minute nap in a colloquium is equivalent to one hour of a good night's sleep" has been put to practice and validated by each of us.

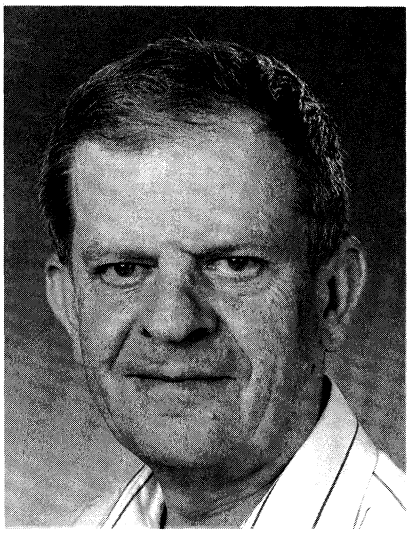
We wish the very best to Dick and Frank in their future personal and professional endeavors.

George Andrews
George Gasper
Mourad Ismail
Paul Nevai

N.B. Due to restrictions beyond our control, it was impossible to include in this issue all of the papers dedicated to Dick and Frank that were accepted for publication; the papers listed below will appear in future issues of this journal.

1. Mizan Rahman and Sergei K. Suslov, "Barnes and Ramanujan-type integrals on the q-linear lattice."

2. Bo Gao, Donald J. Newman, and V. V. Popov, "Convex approximation by rational functions."



SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 243–255, March 1994

A HYPERGEOMETRIC ANALYSIS OF THE GENUS SERIES FOR A CLASS OF 2-CELL EMBEDDINGS IN ORIENTABLE SURFACES*

G. E. ANDREWS†, D. M. JACKSON‡, AND T. I. VISENTIN§

Abstract. The genus series for dipoles is used to determine an explicit expression for the number of dipoles of given genus with a given number of edges. The approach uses a hypergeometric argument, which may assist with other classes of maps such as vertex-regular maps. Such maps are of importance in combinatorial theory and appear to have application to two-dimensional quantum gravity.

Key words. hypergeometric function, 2-cell embedding, rooted map, integral representation, genus series

AMS subject classifications. 33C20, 57N37, 05A15

1. Introduction.

A *dipole* is a map on two vertices of the same degree, with loops allowed. It is a generalization of a map consisting of a multiple edge, whose vertices necessarily have the same degree. In a loose sense, dipoles are the next class of maps in order of complexity beyond the one-vertex maps (called *monopoles*), and dipoles with no loops have been an object of study in topological graph theory [6]. A *k-pole* is a map (loops and multiple edges allowed) on k vertices of the same degree. The set of all k-poles is the set of *vertex-regular* maps. The study of poles may therefore lead indirectly to information about important subclasses of regular maps, namely, quadrangulations and triangulations. The former occur, for example, in the ϕ^4-model of two-dimensional quantum gravity [7].

The *genus series* for a class of maps is the generating series for the number of such maps with respect to genus and the number of edges. The genus series for dipoles (technically, they are rooted) has been given by Jackson [9], using integral representations of sums of characters of the symmetric group. If $\nu = (\nu_1, \nu_2, \ldots)$ is a partition of $2n$, let p_ν denote $p_{\nu_1} p_{\nu_2} \ldots$, where $p_k = \lambda_1^k + \cdots + \lambda_N^k$ and N is a positive integer. Let $V(\lambda)$ be the Vandermonde determinant in $\lambda_1, \ldots, \lambda_N$, and let $[x^n] f(x)$ denote the coefficient of x^n in the formal power series f. Let $m_g(n), d_g(n)$ be the numbers of monopoles and dipoles, respectively, of genus g on n edges. Then

$$m_g(n) = [u^{n+1-2g}] A_{[2n]}(u), \tag{1}$$

$$d_g(n) = \frac{1}{n} [u^{n-2g}] \left(A_{[n^2]}(u) - A_{[n]}^2(u) \right), \tag{2}$$

where

$$A_\nu(N) = \frac{\int_{\mathcal{V}_N} e^{-\frac{1}{2}\operatorname{trace} \mathbf{M}^2} \prod_{k \geq 1} \left(\operatorname{trace} \mathbf{M}^k\right)^{a_k} d\mathbf{M}}{\int_{\mathcal{V}_N} e^{-\frac{1}{2}\operatorname{trace} \mathbf{M}^2} d\mathbf{M}}$$

* Received by the editors April 15, 1992; accepted for publication (in revised form) April 13, 1993.
† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania. The work of this author was supported by National Science Foundation grant DMS 8702695-03.

‡ Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada. The work of this author was supported by a Natural Sciences and Engineering Research Council of Canada grant.

§ Department of Mathematics, University of Winnipeg, Winnipeg, Manitoba, Canada. The work of this author was supported by a Natural Sciences and Engineering Research Council of Canada grant.

$$= \frac{\sqrt{2}^{N^2-N} 2^n}{\sqrt{\pi}^N \prod_{j=1}^N j!} \int_{\mathbb{R}^N} V^2(\lambda) e^{-p_2} p_\nu d\lambda,$$

and a_j is the number of occurrences of j in the partition ν. Here $\mathcal{V}_N$ is the vector space, of dimension N^2 over $\mathbb{R}$, of all $N \times N$ Hermitian complex matrices and $d\mathbf{M}$ denotes Lesbesgue measure. The integrands with respect to $d\mathbf{M}$ are invariant under adjoint action of the unitary group, and diagonalization under this action gives the transformed integral in $d\lambda$. The constant that appears is related to the volume of the unitary group.

In determining $d_g(u)$, it is necessary first to carry out the integration to exhibit $A_\nu(N)$ as a polynomial in N, and then replace N formally by the indeterminate u. This is shown in the cases of monopoles [8] and dipoles [9], in the following two results.

THEOREM 1.1. *The number of monopoles of genus g on n edges is*

$$m_g(n) = \frac{(2n)!}{2^n n!} [u^{n+1-2g}] \sum_{k=1}^{n+1} \binom{u}{k} \binom{n}{k-1} 2^{k-1}.$$

THEOREM 1.2. *The number of dipoles of genus g with n edges is*

$$d_g(n) = (n-1)! \, [u^{n-2g}] \sum_{j=0}^{\lfloor \frac{1}{2}(n-1) \rfloor} \binom{2j}{j} \sum_{r=0}^{n-2j-1} \binom{n-2j-1}{r} \binom{u+j+r}{n}$$
$$\cdot \sum_{k=0}^{\lfloor \frac{1}{2}(n-2j-1) \rfloor} \frac{1}{4^k} \binom{2k}{k} \binom{n}{2k}.$$

As seen in these results, this procedure typically leads to a representation of $A_\nu(u)$ with respect to bases for the ring of formal power series other than the standard (monomial) basis, and this may be inconvenient for subsequent work.

The approach through integral representations can be applied to other classes of maps, and it is of interest to determine whether any more detailed information can be obtained about the exact number of maps. We carry out this task for dipoles in particular and, in doing so, demonstrate the type of hypergeometric arguments that we believe to be of general value in this context. The key appears to lie in representing the genus series with respect to a basis in which its degree and its parity are manifestly apparent. We call such a form *degree respecting* and *parity respecting*. We show how series can be resolved with respect to a natural basis of odd series that is attuned to hypergeometric series. It can be shown by a combinatorial argument that the representation of the genus series for monopoles and dipoles given in Theorems 1.1 and 1.2 and is degree respecting but *not* parity respecting.

The necessary background of results is given below, but the reader is referred to [10] and [11] for basic definitions about maps since these are not given here. Throughout, we use the usual hypergeometric convention that the Pochhammer symbol is $(x)_k = x(x+1)(x+2)\cdots(x+k-1)$.

Because of the embedding theorem (see, for example, [10]), the question addressed here can be expressed in terms of permutations alone. In this context, we seek the number of permutations in $2n$ symbols, with k cycles, which are expressible as the product of a prescribed permutation with two n-cycles and a fixed point free involution. The reader who is unfamiliar with maps may prefer to think of the question in these terms. We prove the following theorem about dipoles.

THEOREM 1.3. *The number of dipoles of genus $g > 0$ with $2m$ edges and $2m+1$ edges is, respectively,*

$$\text{(i)}\quad \frac{2^{\lfloor \frac{1}{2}g \rfloor}}{(2g+1)!}\binom{2m-1}{m}^2\binom{m}{g+1}\left(m-g+\tfrac{1}{2}\right)_{(\lfloor \frac{1}{2}g \rfloor)} r_g(m),$$

$$\text{(ii)}\quad \frac{2^{\lfloor \frac{1}{2}(g-1) \rfloor}}{(2g+1)!}\binom{2m-1}{m}^2\binom{m}{g}(2m+1)\left(m-g+\tfrac{3}{2}\right)_{(\lfloor \frac{1}{2}(g-1) \rfloor)} s_g(m),$$

where $r_g(x)$ and $s_g(x)$ are polynomials of degrees at most $\lfloor \frac{1}{2}(3g-1) \rfloor$ and $\lfloor \frac{3}{2}g \rfloor$, respectively, in x. Moreover, the polynomials $r_g(x)$ and $s_g(x)$ are given explicitly in (11) and (12).

2. Parity respecting form for the series for odd dipoles. We begin by presenting the genus series for dipoles in a form that is both degree respecting *and* exhibits its parity.

Let E, Δ, I be the successor operator, the forward difference operator, and the identity operator defined on the ring of polynomials by $Ef(x) = f(x+1)$, $\Delta f(x) = (E-I)f(x) = f(x+1) - f(x)$, and $If(x) = f(x)$. Since E and I commute, then for $M \geq 0$,

$$\Delta^M f(x) = \sum_{j=0}^{M} (-1)^j \binom{M}{j} f(x+M-j). \tag{3}$$

The following lemma concerns the representation of a polynomial of odd degree with respect to a convenient basis of upper factorial polynomials of odd degree. We extend it slightly later. It is precisely the central difference operator expansion given by Steffensen [12, p. 13, eq. (19)] in the case of odd polynomials.

LEMMA 2.1. *Let $f(x)$ be an odd polynomial of degree at most $2n-1$. Then*

$$f(x) = \sum_{j=0}^{n-1} \beta_j \frac{(x-j)_{2j+1}}{(2j+1)!}, \quad \text{where } \beta_j = \Delta^{2j+1} f(x)\big|_{x=-j-1}.$$

The next result is used later, and illustrates the use of this device to derive the representation of the sort described above. It is equivalent to a result appearing in [1] and [2] in connection with one of the Bailey-type sums.

LEMMA 2.2.

$$\binom{u+a}{n} - \binom{-u+a}{n} = \sum_{j=0}^{\lfloor \frac{1}{2}(n-1) \rfloor} \binom{a-j}{n-2j-1} \frac{2a-n+1}{a-j} \frac{(u-j)_{2j+1}}{(2j+1)!}.$$

Proof. From Lemma 2.1,

$$\begin{aligned}
\beta_m &= \sum_{k=0}^{2m+1} (-1)^k \binom{2m+1}{k} \left\{ \binom{m-k+a}{n} - \binom{-m+k+a}{n} \right\} \\
&= \binom{m+a}{n} {}_2F_1\left[\begin{matrix} -2m-1 & -m-a+n \\ & -m-a \end{matrix} \,\middle|\, 1 \right] \\
&\quad + \binom{-m+a}{n} {}_2F_1\left[\begin{matrix} -2m-1 & -m+a+1 \\ & -m+a-n+1 \end{matrix} \,\middle|\, 1 \right] \\
&= \frac{2a-n+1}{a-n} \binom{a-m}{n-2m-1}
\end{aligned}$$

by the Chu–Vandermonde theorem, giving the result. □

COROLLARY 2.3.

$$\sum_{r=0}^{2m-2j}\binom{2m-2j}{r}\binom{u+j+r}{2m+1}=\sum_{i=0}^{m}\frac{1}{(2i+1)!}4^{i-j}\binom{m-j}{i-j}(u-i)_{2i+1}.$$

Proof. Let $S(m,j)$ denote the left-hand side. Then

$$S=\binom{2m-2j}{m-j}\binom{u+m}{2m+1}+\sum_{r=0}^{m-j-1}\binom{2m-2j}{r}\left\{\binom{u+j+r}{2m+1}-\binom{-u+j+r}{2m+1}\right\},$$

so S is an odd polynomial in u. Then, by doubling S and using Lemma 2.2,

$$\begin{aligned}2S&=\sum_{r=0}^{2m-2j}\binom{2m-2j}{r}\sum_{\lambda=0}^{m}\binom{j+r-\lambda}{2m-2\lambda}\frac{2j+2r-2m}{j+r-\lambda}\frac{(u-\lambda)_{2\lambda+1}}{(2\lambda+1)!}\\&=\sum_{\lambda=0}^{m}\frac{(u-\lambda)_{2\lambda+1}}{(2\lambda+1)!}\sum_{r=0}^{2m-2j}\binom{2m-2j}{r}\binom{j+r-\lambda}{2m-2\lambda}\frac{2j+2r-2m}{j+r-\lambda},\end{aligned}$$

so

$$\begin{aligned}S&=\sum_{\lambda=0}^{m}\frac{(u-\lambda)_{2\lambda+1}}{(2\lambda+1)!}\binom{j-\lambda}{2m-2\lambda}\frac{j-m}{j-\lambda}\\&\quad\cdot{}_3F_2\left[\begin{matrix}-2m+2j & j-\lambda & j-m+1\\ & j+\lambda-2m+1 & j-m\end{matrix}\,\middle|\,-1\right]\\&=\sum_{\lambda=0}^{m}\frac{(u-\lambda)_{2\lambda+1}}{(2\lambda+1)!}\binom{j-\lambda}{2m-2\lambda}\frac{j-m}{j-\lambda}\\&\quad\cdot\lim_{\substack{a\to-2m+2j\\ c\to\infty}}{}_5F_4\left[\begin{matrix}a & 1+\frac{1}{2}a & b & c & d\\ & \frac{1}{2}a & 1+a-b & 1+a-c & 1+a-d\end{matrix}\,\middle|\,1\right],\end{aligned}$$

where $b=j-\lambda, d=-m+j+\frac{1}{2}$, and the result follows by Dougall's theorem. □

A referee noted that Corollary 2.3 follows from the identity

$${}_2F_1\left[\begin{matrix}-2k & -x\\ & -N\end{matrix}\,\middle|\,2\right]={}_3F_2\left[\begin{matrix}-k & x-N & -x\\ & -\frac{1}{2}N & -\frac{1}{2}(N-1)\end{matrix}\,\middle|\,1\right],$$

connecting Krawtchouk polynomials of even degree with Hahn polynomials. This entails applying the linear transformation formula to our ${}_2F_1(-1)$ to obtain a ${}_2F_1(\frac{1}{2})$, reversing the resulting series to obtain a ${}_2F_1(2)$, and observing by standard arguments that the above identity holds for all real N, rather than just integral N. Corollary 3.1 follows in the same way from the analogous formula for Krawtchouk polynomials of odd degree.

It is useful to observe that Steffensen's expansion is a natural method for expanding an odd polynomial into a series involving rising factorials such that the oddness is evident term-by-term, and has the merit of being applicable generally to the above type of problems.

The next result is given in Bailey (see [3, p. 12, eq. (1)], with $a=-n$).

$$(4)\qquad {}_3F_2\left[\begin{matrix}-n & b & c\\ & e & f\end{matrix}\,\middle|\,1\right]=\frac{(e-b)_n}{(e)_n}{}_3F_2\left[\begin{matrix}-n & b & f-c\\ & 1-n+b-e & f\end{matrix}\,\middle|\,1\right].$$

Thus, from (4),

$$\lim_{e\to -n} {}_3F_2\left[\begin{matrix} -n & b & c \\ & e & f \end{matrix}\,\middle|\, 1\right] = \frac{(b+1)_n}{n!}\,{}_3F_2\left[\begin{matrix} -n & b & f-c \\ & 1+b & f \end{matrix}\,\middle|\, 1\right] = \sum_{j=0}^{n} \frac{(b)_j (c)_j}{j!(f)_j}.$$

We may now give a parity and degree respecting expression for the dipole generating series.

THEOREM 2.4. *Let $n = 2m+1$. Then the number of dipoles of genus g on n edges is*

$$d_g(n) = \binom{2m}{m} n! \left[u^{n-2g}\right] \sum_{i=0}^{m} \frac{(u-i)_{2i+1}}{(2i+1)!} 4^{i-m} \binom{m}{i} \cdot \sum_{k=0}^{m-i} (-1)^k \frac{1}{2k+1} \binom{m-i}{k} \binom{2m-k}{m}.$$

Proof. From Theorem 1.2 and Corollary 2.3,

$$d_g(n) = (n-1)! \left[u^{n-2g}\right] \sum_{i=0}^{m} \frac{(u-i)_{2i+1}}{(2i+1)!} \sum_{k=0}^{m} \sum_{j=0}^{m-k} 4^{i-j-k} \binom{m-j}{i-j} \binom{2j}{j} \binom{2k}{k} \binom{n}{2k}.$$

Now

$$\begin{aligned} &\sum_{j=0}^{m-k} 4^{i-j-k} \binom{m-j}{i-j} \binom{2j}{j} \\ &= 4^{i-k} \binom{m}{i} \lim_{e\to -m+k} {}_3F_2\left[\begin{matrix} \frac{1}{2} & -m+k & -i \\ & e & -m \end{matrix}\,\middle|\, 1\right] \qquad (5) \\ &= 4^{i-m} \binom{m}{i} \frac{(2m-2k+1)!}{(m-k)!^2}\, {}_3F_2\left[\begin{matrix} -m+k & \frac{1}{2} & -m+i \\ & \frac{3}{2} & -m \end{matrix}\,\middle|\, 1\right]. \end{aligned}$$

Substituting this into the sum over k we have

$$\begin{aligned} &\sum_{k=0}^{m} \sum_{j=0}^{m-k} 4^{i-j-k} \binom{m-j}{i-j} \binom{2j}{j} \binom{2k}{k} \binom{n}{2k} \\ &\quad = 4^{i-m} \frac{1}{n} \sum_{k=0}^{m} \frac{(2m+1)!}{k!^2 (m-k)!^2}\, {}_3F_2\left[\begin{matrix} -m+k & \frac{1}{2} & -m+i \\ & \frac{3}{2} & -m \end{matrix}\,\middle|\, 1\right] \\ &\quad = \frac{1}{n} 4^{i-m} \binom{m}{i} (2m+1)! \sum_{r=0}^{m-i} \frac{(-m+i)_r (\frac{1}{2})_r (-1)^r}{r!(\frac{3}{2})_r (-m)_r m!(m-r)!} \\ &\qquad \cdot \sum_{k=0}^{m} \frac{m!(m-r)!}{k!^2 (m-k)!(m-k-r)!} \\ &\quad = 4^{i-m} \binom{m}{i} \frac{(2m)!}{m!^2} \sum_{r=0}^{m-1} (-1)^r \binom{m-i}{r} \frac{1}{2r+1} \binom{2m-r}{m}, \end{aligned}$$

since the sum over k is equal to $\binom{2m-r}{m}$, and this gives the result. □

3. Parity respecting form for the series for even dipoles. We follow the approach of the previous section to give a parity respecting form for the genus series for dipoles on an even number of edges. To do this, we need the following corollary, which corresponds to Corollary 2.3.

COROLLARY 3.1.

$$\sum_{r=0}^{2m-2j-1}\binom{2m-2j-1}{r}\binom{u+j+r}{2m}=\sum_{i=j}^{m-1}\frac{4^{i-j}u}{(2i+1)!(i+1)}\binom{m-j-1}{i-j}(u-i)_{2i+1}.$$

Proof. Replace m by $m+j$ and then i by $i+j$ in the enunciation to obtain the *restatement*

$$\sum_{r=0}^{2m-1}\binom{2m-1}{r}\binom{u+j+r}{2m+2j}=\sum_{i=0}^{m-1}\frac{4^{i}u}{(2i+2j+1)!(i+j+1)}\binom{m-1}{i}(u-i-j)_{2i+2j+1},$$

or, equivalently,

$$\binom{u+j}{2m+2j}{}_2F_1\left[\begin{matrix}-2m+1 & u+j+1\\ & u-2m-j+1\end{matrix}\,\middle|\,-1\right]=\frac{u}{j+1}\binom{u+j}{2j+1}{}_3F_2\left[\begin{matrix}-m+1 & 1+j-u & u+j+1\\ & j+\frac{3}{2} & j+2\end{matrix}\,\middle|\,1\right]. \tag{6}$$

The proof is complete once this is established.

But

$$ {}_2F_1\left[\begin{matrix}a & b\\ & \kappa-b\end{matrix}\,\middle|\,-1\right]=\frac{\Gamma(\kappa-b)\Gamma(\frac{1}{2}\kappa)}{\Gamma(\kappa)\Gamma(\frac{1}{2}\kappa-b)}{}_3F_2\left[\begin{matrix}b & \frac{1}{2}(\kappa-a) & \frac{1}{2}(\kappa-a)+\frac{1}{2}\\ & \kappa-a & \frac{1}{2}(\kappa+1)\end{matrix}\,\middle|\,1\right] \tag{7}$$

from [3, p. 33, eq. (3)]

$$=\frac{\Gamma(\kappa-b)\Gamma(\frac{1}{2}\kappa)}{\Gamma(\kappa)\Gamma(\frac{1}{2}\kappa-b)}\frac{\Gamma(\kappa-a)\Gamma(\frac{1}{2}\kappa-b)}{\Gamma(\kappa-a-b)\Gamma(\frac{1}{2}\kappa)}\cdot{}_3F_2\left[\begin{matrix}b & \frac{1}{2}(a+1) & \frac{1}{2}a\\ & \frac{1}{2}\kappa & \frac{1}{2}(\kappa+1)\end{matrix}\,\middle|\,1\right] \tag{8}$$

from [3, p. 98, ex. 7]

$$=\frac{\Gamma(\kappa-b)\Gamma(\kappa-a)}{\Gamma(\kappa)\Gamma(\kappa-a-b)}{}_3F_2\left[\begin{matrix}b & \frac{1}{2}(a+1) & \frac{1}{2}a\\ & \frac{1}{2}\kappa & \frac{1}{2}(\kappa+1)\end{matrix}\,\middle|\,1\right]. \tag{9}$$

Equation (9) appears in [15]. Then with $a=1-2m, b=u+j+1, \kappa=2u-2m+2$ (so $\kappa-a=2u+1$),

$${}_2F_1\left[\begin{matrix}1-2m & u+j+1\\ & u-2m-j+1\end{matrix}\,\middle|\,-1\right]=\frac{(2u-2m+2)_{2m-1}}{(u-2m-j+1)_{2m-1}}{}_3F_2\left[\begin{matrix}u+j+1 & -m+1 & -m+\frac{1}{2}\\ & u-m+1 & u-m+\frac{3}{2}\end{matrix}\,\middle|\,1\right].$$

Thus, from (6), we must establish that

$$\binom{u+j}{2m+2j}\frac{(2u-2m+2)_{2m-1}}{(u-2m-j+1)_{2m-1}}{}_3F_2\left[\begin{matrix}u+j+1 & -m+1 & -m+\frac{1}{2}\\ & u-m+1 & u-m+\frac{3}{2}\end{matrix}\,\middle|\,1\right]=\frac{u}{j+1}\binom{u+j}{2j+1}{}_3F_2\left[\begin{matrix}-m+1 & 1+j-u & u+j+1\\ & j+\frac{3}{2} & j+2\end{matrix}\,\middle|\,1\right].$$

We now use Sheppard's [14] transformation. This is a generalization of the Pfaff–Saalschutz formula. It is not explicitly stated in [3], but is given without attribution in [4, p. 168].

$$
\begin{aligned}
{}_3F_2\left[\begin{matrix} -n & b & c \\ & e & f \end{matrix}\,\middle|\,1\right] &= \frac{(e-b)_n(f-b)_n}{(e)_n(f)_n} \\
&\cdot {}_3F_2\left[\begin{matrix} -n & b & -n+b+c+1-e-f \\ & -n+b-f+1 & -n+b-e+1 \end{matrix}\,\middle|\,1\right],
\end{aligned} \tag{10}
$$

so

$$
\begin{aligned}
&{}_3F_2\left[\begin{matrix} -m+1 & u+j+1 & 1+j-u \\ & j+\frac{3}{2} & j+2 \end{matrix}\,\middle|\,1\right] \\
&\quad = \frac{(j+\frac{1}{2}-u-j)_{m-1}(j+1-u-j)_{m-1}}{(j+\frac{3}{2})_{m-1}(j+2)_{m-1}} \\
&\quad \cdot {}_3F_2\left[\begin{matrix} -m+1 & u+j+1 & -m+\frac{1}{2} \\ & u-m+1 & u-m+\frac{3}{2} \end{matrix}\,\middle|\,1\right].
\end{aligned}
$$

The identity to be proved is therefore equivalent to

$$
\binom{u+j}{2m+2j}\frac{(2u-2m+2)_{2m-1}}{(u-2m-j+1)_{2m-1}} = \frac{u}{j+1}\binom{u+j}{2j+1}\frac{(-2u+1)_{2m-2}}{(2j+3)_{2m+2}}.
$$

But both sides are equal to

$$
\frac{(u+j)!(2u)!}{(2m+2j)!(u-j-1)!(2u-2m+1)!},
$$

and this therefore establishes the result. □

We may now give a parity and degree respecting expression for the dipole generating series.

THEOREM 3.2. *Let $n = 2m$. Then the number of dipoles of genus g on n edges is*

$$
\begin{aligned}
d_g(n) = \binom{2m}{m} n!\,[u^{n-2g-1}] &\sum_{i=0}^{m-1} \frac{4^{i-m}(u-i)_{2i+1}}{(2i+1)!(i+1)}\binom{m-1}{i} \\
&\cdot \sum_{k=0}^{m-i-1} (-1)^k \frac{1}{2k+1}\binom{m-i-1}{k}\binom{2m-k-1}{m}.
\end{aligned}
$$

Proof. From Theorem 1.2 and from Corollary 3.1,

$$
\begin{aligned}
d_g(n) = (n-1)![u^{n-2g}] &\sum_{j=0}^{m-1}\binom{2j}{j} u \sum_{i=0}^{m-1} \frac{4^{i-j}}{(2i+1)!(i+1)}\binom{m-j-1}{i-j}(u-i)_{2i+1} \\
&\cdot \sum_{k=0}^{m-j-1} \frac{1}{4^k}\binom{2k}{k}\binom{2m}{2k}.
\end{aligned}
$$

From (5), with slight modification,

$$
\sum_{j=0}^{m-k-1} 4^{i-j-k}\binom{m-j-1}{i-j}\binom{2j}{j} = A_{k,n},
$$

where

$$A_{k,n} = 4^{i-m+1}\binom{m-1}{i}\frac{(2m-2k-1)!}{(m-k-1)!^2}\,{}_3F_2\left[\begin{matrix} -m+k+1 & \frac{1}{2} & -m+i+1 \\ & \frac{3}{2} & -m+1 \end{matrix}\,\middle|\, 1\right],$$

so

$$d_g(n) = (2m-1)![u^{n-2g}]\sum_{0\le i\le m-1}\frac{u(u-i)_{2i+1}}{(2i+1)!(i+1)}\sum_{k=0}^{m-1}\binom{2k}{k}\binom{2m}{2k}\binom{m-1}{i}A_{k,n}.$$

But

$$\begin{aligned}
&(2m-1)!\sum_{k=0}^{m-1}\binom{2k}{k}\binom{2m}{2k}\binom{m-1}{i}A_{k,m} \\
&= \frac{4^{i-m+1}}{n}\binom{m-1}{i}\sum_{k=0}^{m}\frac{(2m)!^2}{k!^2(m-k-1)!^2(2m-2k)} \\
&\quad\cdot\sum_{h\ge 0}\frac{(-1)^h}{2h+1}\binom{m-k-1}{h}\frac{(-m+i+1)_h}{(-m+1)_h} \\
&= \frac{4^{i-m+1}}{2n}\binom{m-1}{i}\sum_{h\ge 0}\frac{(2m)!^2(-1)^h(-m+i+1)_h}{h!(2h+1)(-m+1)_h m!(m-h-1)!}
\end{aligned}$$

since

$$\sum_{k\ge 0}\frac{m!(m-h-1)!}{k!^2(m-k)!(m-k-h-1)!} = \binom{2m-h-1}{m-h-1}.$$

The result follows. □

4. The sphere and the torus. As an example of the use of Theorems 2.4 and 3.2, we consider dipoles on the sphere and the torus. In the odd case, when $g = 0$, there is a single term to consider and this gives

$$d_0(2m+1) = \binom{2m}{m}^2.$$

When $g = 1$, we note that $(u-m)_{2m+1} = u^{2m+1} - \frac{1}{6}m(m+1)(2m+1)u^{2m-1} + \cdots$, so

$$\begin{aligned}
d_1(2m+1) &= -\tfrac{1}{6}m(m+1)(2m+1)\binom{2m}{m}^2 \\
&\quad + \tfrac{1}{2}(2m+1)m^2\binom{2m}{m}\sum_{r=0}^{1}\binom{1}{r}\frac{(-1)^r}{2r+1}\binom{2m-r}{m} \\
&= \tfrac{1}{3}\binom{2m-1}{m}^2 m(2m+1)(3m-2).
\end{aligned}$$

In the even case

$$d_0(2m) = \binom{2m}{m}(2m)!\frac{1}{(2m-1)!m}\frac{1}{4}\binom{2m-1}{m} = \binom{2m-1}{m}^2,$$

and

$$d_1(2m) = \binom{2m}{m}(2m)!\left\{\frac{-\sum_{j=1}^{m-1} j^2}{(2m-1)!}\frac{1}{4m}\binom{2m-1}{m}\right.$$
$$\left. + \frac{1}{(2m-3)!}\frac{1}{4^2}\sum_{h=0}^{1}\binom{1}{h}\frac{(-1)^h}{2h+1}\binom{2m-h-1}{m}\right\}$$
$$= \tfrac{1}{6}\binom{2m-1}{m}^2 m(m-1)(3m-1).$$

The number of terms to be considered in the summation increases with genus.

5. The number of dipoles. We are in a position to prove Theorem 1.3 and to obtain full structural information about the polynomials $r_g(m)$ and $s_g(m)$. The following result is needed (the proof is sketched in [13, pp. 234–236]).

PROPOSITION 5.1. *Let* $(u-j)_{2j+1} = \sum_{i=0}^{j}\mu_i(j)u^{2i+1}$. *Then, for* $0 \le k \le j$,

$$\mu_{j-k}(j) = \sum_{\lambda=1}^{k}\alpha_{k,\lambda}\binom{2j+2}{2k+\lambda} = (2j-2k+2)_{2k+1}\hat{\mu}_k(j),$$

where $\alpha_{k,\lambda}$ *is independent of* j *and* $\hat{\mu}_k(j)$ *is a polynomial in* j *of degree* $k-1$.

Proof. Let $\mu_i(j)$ be defined by $(u-j)_{2j+1} = \sum_{i=0}^{j} u^{2i+1}\mu_i(j)$. It is readily seen that the $\mu_i(j)$ are determined by $\mu_j(j) = 1, \mu_0(j) = (-1)^j j!^2$, and for $0 < i < j$, $\mu_i(j) = \mu_{i-1}(j-1) - j^2\mu_i(j-1)$. In Riordan's notation [13, p. 213], $x^{[n]} = x(x+\frac{1}{2}n-1)(x+\frac{1}{2}n-2)\cdots(x+\frac{1}{2}n-n+1)$, so $x^{[2j+2]} = x(x+j)(x+j-1)\cdots(x-j)$ and $x^{[2j+2]} = \sum_{k=0}^{2j+2} t(2j+2,k)x^k$, when $\sum_{i=0}^{j} u^{2i+2}\mu_i(j) = u(u-j)_{2j+1} = u^{[2j+2]} = \sum_{k=0}^{2j+2} t(2j+2,k)u^k$, and so $\mu_i(j) = t(2j+2, 2i+2)$. Now [13, pp. 234–236] $4^k t(n, n-2k) = \sum_{j=1}^{k}\alpha_{k,j}\binom{n}{2k+j}$, where the $\alpha_{k,j}$ are rational numbers. Therefore, $\mu_{j-k}(j) = t(2j+2, 2j-2k+2) = 4^{-k}\sum_{\lambda=1}^{k}\alpha_{k,\lambda}\binom{2j+2}{2k+\lambda}$. Thus $\mu_{j-k}(j)$ is a polynomial in j of degree $3k$ and has $(2j-2k+2)_{2k+1}$ as a factor, so $\mu_{j-k}(j) = (2j-2k+2)_{2k+1}\hat{\mu}_k(j)$, where $\hat{\mu}_k(j)$ is a polynomial in j of degree at most $k-1$. □

The $\mu_i(j)$ are more or less the *divided central differences of zero*. The odd and even cases seem to be different, and we treat them separately.

Proof of Theorem 1.3(ii). From Theorem 2.4 with $g > 0$ and $n = 2m+1$,

$$d_g(n) = \binom{2m}{m}n!\,[u^{n-2g}]\sum_{i=0}^{m}\frac{1}{(2i+1)!}\sum_{\lambda=0}^{i}\mu_\lambda(i)u^{2\lambda+1}4^{i-m}\binom{m}{i}$$
$$\cdot\sum_{k=0}^{m-i}\frac{(-1)^k}{2k+1}\binom{m-i}{k}\binom{2m-k}{m}$$
$$= \binom{2m}{m}n!\sum_{i=0}^{g}\frac{\mu_{m-g}(m-i)}{(2m-2i+1)!}4^{-i}\binom{m}{i}\sum_{k=0}^{i}\frac{(-1)^k}{2k+1}\binom{i}{k}\binom{2m-k}{m}$$
$$= 4\binom{2m-1}{m}^2\binom{m}{g}(2m+1)$$
$$\cdot\sum_{0\le k\le i\le g}\frac{(-1)^k\mu_{m-g}(m-i)g!4^{-i}}{k!(i-k)!(2k+1)}\,\frac{(2m-2i+2)_{2i-k-1}m!(m-g)!}{(m-i)!(m-k)!}$$

$$= 4\binom{2m-1}{m}^2\binom{m}{g}(2m+1)(m-g+\tfrac{3}{2})_{\lfloor\frac{1}{2}(g-1)\rfloor}$$
$$\cdot \sum_{0\le k\le i\le g} \frac{(-1)^k g! 4^{-i}}{k!(i-k)!(2k+1)} B_{k,i,g}(m),$$

where

$$B_{k,i,g}(m) = \frac{\hat{\mu}_{g-i}(m-i)(2m-2i+2)_{2i-k-1}(m+1-i)_i}{(m-g+1)_{g-k}(m-g+\frac{3}{2})_{\lfloor\frac{1}{2}(g-1)\rfloor}}(2m-2g+2)_{2g-2i+1}.$$

If $B_{k,i,g}(m)$ is a polynomial in m, then its degree is at most $\lfloor\frac{3}{2}g\rfloor$, and consequently,

$$\frac{2^{\lfloor\frac{1}{2}(g-1)\rfloor}}{(2g+1)!}s_g(m) = \sum_{0\le k\le i\le g} \frac{(-1)^k g! 4^{1-i}}{k!(i-k)!(2k+1)} B_{k,i,g}(m) \tag{11}$$

would produce $s_g(m)$ as a polynomial of degree at most $\lfloor\frac{3}{2}g\rfloor$. All that now remains is to prove the assertion that $B_{k,i,g}(m)$ is indeed a polynomial. Since $(2a)_k = 2^k(a)_{\lfloor\frac{1}{2}(k+1)\rfloor}(a+\frac{1}{2})_{\lfloor\frac{1}{2}k\rfloor}$, we have

$$B_{k,i,g}(m) = \hat{\mu}_{g-i}(m-i)(m-i+1)2^{2g-k}(m+1-i)_i$$
$$\cdot\frac{(m-g+1)_{\lfloor\frac{1}{2}(2g-k)\rfloor}}{(m-g+1)_{g-k}}\frac{(m-g+\frac{3}{2})_{\lfloor\frac{1}{2}(2g-k-1)\rfloor}}{(m-g+\frac{3}{2})_{\lfloor\frac{1}{2}(g-1)\rfloor}}.$$

Now $k \le g$, so $(m-g+\frac{3}{2})_{\lfloor\frac{1}{2}(2g-k-1)\rfloor}/(m-g+\frac{3}{2})_{\lfloor\frac{1}{2}(g-1)\rfloor}$ is a polynomial in m, and since $\lfloor\frac{1}{2}(2g-k)\rfloor \ge g-k$, $(m-g+1)_{\lfloor\frac{1}{2}(2g-k)\rfloor}/(m-g+1)_{g-k}$ is a polynomial in m. This establishes the assertion, and thereby completes the proof. □

We note that (11) gives an explicit expression for the polynomial $s_g(m)$.

Proof of Theorem 1.3(i). From Theorem 3.2 with $g > 0$ and $n = 2m$,

$$d_g(n) = \binom{2m}{m} n! \left[u^{n-2g-1}\right] \sum_{i=0}^{m-1} \frac{(u-i)_{2i+1}}{(2i+1)!(i+1)} 4^{i-m}\binom{m-1}{i}$$
$$\cdot \sum_{k=0}^{m-i-1} \frac{(-1)^k}{2k+1}\binom{m-i-1}{k}\binom{2m-k-1}{m}$$
$$= 2\binom{2m-1}{m} n! \sum_{i=m-g-1}^{m-1} \frac{\mu_{m-g-1}(i)}{(2i+1)!(i+1)} 4^{i-m}\binom{m-1}{i}$$
$$\cdot \sum_{k=0}^{m-i-1} \frac{(-1)^k}{2k+1}\binom{m-i-1}{k}\binom{2m-k-1}{m}$$
$$= 2\binom{2m-1}{m} n! \sum_{i=0}^{g} \frac{\mu_{m-g-1}(m-i-1)}{(2m-2i-1)!(m-i)} 4^{-1-i}\binom{m-1}{i}$$
$$\cdot \sum_{k=0}^{i} \frac{(-1)^k}{2k+1}\binom{i}{k}\binom{2m-k-1}{m}$$
$$= \binom{2m-1}{m}^2\binom{m}{g+1}(g+1)!$$

$$\cdot \sum_{0\le k\le i\le g} \frac{(-1)^k \mu_{m-g-1}(m-i-1)4^{-i}}{k!(i-k)!(2k+1)} \frac{(2m-2i)_{2i-k}(m-k)_k}{(m-g)_{g-i+1}}$$
$$= \frac{2^{\lfloor \frac{1}{2}g\rfloor}}{(2g+1)!}\binom{2m-1}{m}^2\binom{m}{g+1}\left(m-g+\frac{1}{2}\right)_{\lfloor \frac{1}{2}g\rfloor} r_g(m),$$

where

$$r_g(m) = 2^{-\lfloor \frac{1}{2}g\rfloor}(2g+1)!(g+1)! \sum_{0\le k\le i\le g} \frac{(-1)^k 2^{2g-k-2i+1}}{k!(i-k)!(2k+1)}$$
$$\cdot \left\{ \frac{\hat{\mu}_{g-i}(m-i-1)(m-g)_{\lfloor \frac{1}{2}(2g-k+1)\rfloor}(m-g+\frac{1}{2})_{\lfloor \frac{1}{2}(2g-k)\rfloor}(m-k)_k}{(m-g)_{g-i}(m-g+\frac{1}{2})_{\lfloor \frac{1}{2}g\rfloor}} \right\}. \tag{12}$$

Now $r_g(m)$ is a polynomial in m because $g-i \le \lfloor \frac{1}{2}(2g-k+1)\rfloor$, since $0 \le k \le i$ and also $\lfloor \frac{1}{2}(2g-k)\rfloor \le \lfloor \frac{1}{2}g\rfloor$. Therefore, by inspection, $r_g(m)$ is a polynomial of degree at most $\lfloor \frac{1}{2}(3g-1)\rfloor$. □

We note that (12) gives an explicit expression for the polynomial $r_g(m)$.

6. Parity respecting forms for monopoles. We conclude by returning to monopoles and show that they can be treated in a similar way. It is readily seen by symbolic computation that the polynomial

$$\sum_{k=1}^{n+1}\binom{u}{k}\binom{n}{k-1}2^{k-1}$$

has parity $n+1$ mod 2 and is, by Theorem 1.1, up to a numerical factor, the genus series for monopoles. We transform this polynomial to a basis that exhibits the correct parity, and then determine explicit expressions for the number of monopoles.

LEMMA 6.1.

(i) $$\sum_{k=0}^{2m+1}\binom{u}{k+1}\binom{2m+1}{k}2^k = \sum_{j=0}^{m}\frac{4^j}{j+1}\binom{m}{j}\frac{(u-j)_{2j+1}}{(2j+1)!},$$

(ii) $$\sum_{k=0}^{2m}\binom{u}{k+1}\binom{2m}{k}2^k = \sum_{j=0}^{m}4^j\binom{m}{j}\frac{(u-j)_{2j+1}}{(2j+1)!}.$$

Proof. (i) The left-hand side is

$$2^{2m+1}\binom{u}{2m+2}{}_2F_1\left[\begin{matrix}-2m-1 & -2m-2\\ & u-2m-1\end{matrix}\,\middle|\,\frac{1}{2}\right]$$
$$= \binom{u}{2m+2}{}_2F_1\left[\begin{matrix}-2m-1 & u+1\\ & u-2m-1\end{matrix}\,\middle|\,-1\right]$$
$$= \binom{u}{2m+2}\frac{(2u-2m)_{2m+1}}{(u-2m-1)_{2m+1}}{}_3F_2\left[\begin{matrix}u+1 & -m & -m-\frac{1}{2}\\ & u-m & u-m+\frac{1}{2}\end{matrix}\,\middle|\,1\right] \quad \text{by (9)}$$
$$= \binom{u}{2m+2}\frac{(2u)!(u-2m-2)!}{(2u-2m-1)!(u-1)!}\cdot\frac{(-m-1)_m(-m-\frac{1}{2})_m}{(u-m)_m(u-m+\frac{1}{2})_m}$$
$$\cdot {}_3F_2\left[\begin{matrix}-m & u+1 & -u+1\\ & 2 & \frac{3}{2}\end{matrix}\,\middle|\,1\right] \text{ by (10)}$$
$$= u^2\,{}_3F_2\left[\begin{matrix}-m & u+1 & -u+1\\ & 2 & \frac{3}{2}\end{matrix}\,\middle|\,1\right].$$

and this is equivalent to the result.

(ii) The proof is similar. □

Explicit expressions for the numbers of monopoles are given in the next result.

THEOREM 6.2. *Let* $C_n = \frac{1}{n+1}\binom{2n}{n}$, *the nth Catalan number. Then, for* $g > 0$,

$$\text{(i)}\quad m_g(2m+1) = C_{2m+1}\frac{(2m+2)!}{(2m-g+1)!}\sum_{j=0}^{g} 4^{-j}\binom{m}{j}\hat{\mu}_{g-j}(m-j)$$
$$= C_{2m+1}(2m-2g+2)_{2g+1}\sigma_g(m),$$

$$\text{(ii)}\quad m_g(2m) = \frac{(4m)!}{(2m)!(2m-2g+1)!}\sum_{j=0}^{g} 4^{-j}2(m-j+1)\binom{m}{j}\hat{\mu}_{g-j}(m-j)$$
$$= C_{2m}(2m-2g+2)_{2g}\rho_g(m),$$

where $\sigma_g(m)$ *is a polynomial in* m *of degree* $g-1$ *and* $\rho_g(m)$ *is a polynomial in* m *of degree* g.

Proof. (i) From Theorem 1.1 and Lemma 6.1,

$$m_g(2m+1) = C_{2m+1}\frac{(2m+2)!}{2^{2m+1}}\left[u^{2m+1-2g}\right]\sum_{j=0}^{m}\frac{4^j}{j+1}\binom{m}{j}\frac{(u-j)_{2j+1}}{(2j+1)!}$$
$$= C_{2m+1}\frac{(2m+2)!}{4^m}$$
$$\cdot\left[u^{2m+1-2g}\right]\sum_{j=0}^{m}\frac{4^j}{(2j+2)!}\binom{m}{j}\sum_{i=0}^{j}\mu_i(j)u^{2i+1}.$$

From Proposition 5.1,

$$= C_{2m+1}\frac{(2m+2)!}{4^m}\sum_{j=0}^{m}\frac{4^j}{(2j+2)!}\binom{m}{j}\mu_{m-g}(j)$$
$$= C_{2m+1}(2m-2g+2)_{2g+1}\sum_{j=0}^{g}4^{-j}\binom{m}{j}\hat{\mu}_{g-j}(m-j),$$

and this gives the result.

(ii) From Theorem 1.1 and Lemma 6.1,

$$m_g(2m) = \frac{(4m)!}{4^m(2m)!}\left[u^{2m+1-2g}\right]\sum_{j=0}^{m}\frac{4^j}{(2j+1)!}\binom{m}{j}\sum_{i=0}^{j}\mu_i(j)u^{2i+1}.$$

From Proposition 5.1,

$$= \frac{(4m)!}{4^m(2m)!}\sum_{j=0}^{m}\frac{4^j}{(2j+1)!}\binom{m}{j}\mu_{m-g}(j)$$
$$= \frac{(4m)!}{4^m(2m)!}\sum_{j=0}^{g}4^{m-j}\binom{m}{j}\frac{2(m-j+1)}{(2m-2g+1)!}\hat{\mu}_{g-j}(m-j)$$
$$= \frac{(4m)!}{(2m)!(2m-2g+1)!}\sum_{j=0}^{g}4^{-j}\binom{m}{j}2(m-j+1)\hat{\mu}_{g-j}(m-j)$$
$$= C_{2m}(2m-2g+2)_{2g}\rho_g(m),$$

and this gives the result. □

REFERENCES

[1] G. E. ANDREWS AND W. H. BURGE, *Determinant identities*, Pacific J. Math., 158 (1993), pp. 1–14.

[2] G. E. ANDREWS, *Plane partitions* V: *The T.S.S.C.P.P. conjecture*, J. Combinatorial Theory (A), to appear.

[3] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935; reprinted by Hafner, New York, 1964.

[4] A. ERDÉLYI AND M. WEBER, *On the finite difference analog of Rodrigues's formula*, Amer. Math. Monthly, 59 (1952), pp. 163–168.

[5] J. L. GROSS AND T. W. TUCKER, San Antonio AMS Meeting, 1987, private communication.

[6] J. L. GROSS, D. P. ROBBINS, AND T. W. TUCKER, *Genus distributions for bouquets of circles*, J. Combin. Theory ser. B, 47 (1989), pp. 292–306.

[7] C. ITZYKSON AND J-M. DROUFFE, *Statistical Field Theory*, Vol. 2, Cambridge University Press, Cambridge, 1989.

[8] D. M. JACKSON, *Counting cycles in permutations by group characters, with an application to a topological problem*, Trans. Amer. Math. Soc., 299 (1987), pp. 785–801.

[9] ———, *On the integral representation for the genus series for 2-cell embeddings*, Trans. Amer. Math. Soc., to appear.

[10] D. M. JACKSON AND T. I. VISENTIN, *A character theoretic approach to embeddings of rooted maps in an orientable surface of given genus*, Trans. Amer. Math. Soc., 322 (1990), pp. 343–363.

[11] ———, *Character theory and rooted maps in an orientable surface of given genus: face coloured maps*, Trans. Amer. Math. Soc., 322 (1990), pp. 365–376.

[12] J. F. STEFFENSON, *Interpolation*, Chelsea, New York, 1950.

[13] J. RIORDAN, *Combinatorial Identities*, John Wiley, New York, 1968.

[14] W. F. SHEPPARD, *Summation of the coefficients of some terminating hypergeometric series*, Proc. London Math. Soc. (2), 10 (1912), pp. 469–478.

[15] F. J. W. WHIPPLE, *On series allied to the hypergeometric series with argument* -1, Proc. London Math. Soc. (2), 30 (1930), pp. 81–94.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 256–273, March 1994

002

ON CONNECTION COEFFICIENTS FOR q-DIFFERENCE SYSTEMS OF A-TYPE JACKSON INTEGRALS*

KAZUHIKO AOMOTO†

Abstract. General Jackson integrals are formulated. Two different kinds of special Jackson integrals are defined. The explicit relation formulae among them are obtained by the use of theta rational functions.

Key words. connection coefficient, q-difference equation, Jackson integral

1. Introduction. The well-known Ramanujan's $\psi_{1,1}$-sum formula shows that the Jackson integral

$$\int_{[0,\xi\infty]_q} t^{\alpha-1}\frac{(t)_\infty}{(q^\beta t)_\infty}d_qt = (1-q)\sum_{n=-\infty}^{+\infty} q^{n\alpha}\xi^\alpha\frac{(\xi q^n)_\infty}{(q^{\beta+n}\xi)_\infty} \tag{1.1}$$

is equal to the value

$$\left(\frac{\xi}{q}\right)^{\alpha-2}\frac{\theta(q^{\beta+1})\theta(q^{\alpha+\beta-2}\xi)}{\theta(q^{\alpha+\beta-1})\theta(q^\beta\xi)}\int_0^1 t^{\alpha-1}\frac{(t)_\infty}{(q^\beta t)_\infty}d_qt, \tag{1.2}$$

where $\theta(x)$ denotes the Jacobi elliptic theta function $(x)_\infty(q/x)_\infty(q)_\infty$. The equality (1.1) = (1.2) can be explained by saying that in a sense the countable set $[0,\xi]_q$ is homologous to scalar times of the q-interval $[0,1]_q$ with regard to the functions $\Phi = t^\alpha(t)_\infty/(q^\beta t)_\infty$, where the scalar factor

$$\left(\frac{\xi}{q}\right)^{\alpha-2}\frac{\theta(q^{\beta+1})\theta(q^{\alpha+\beta-2}\xi)}{\theta(q^{\alpha+\beta-1})\theta(q^\beta\xi)}$$

is a pseudoconstant, i.e., a constant from the viewpoint of q-difference in the parameters α,β,ξ.

Let $\Phi(t)$ be a q-analog multiplicative function on an n-dimensional algebraic torus $\bar{X}\simeq(\mathbf{C}^*)^n$,

$$\begin{aligned}\Phi(t) &= t_1^{\alpha_1}\cdots t_n^{\alpha_n}\prod_{j=1}^n\prod_{k=1}^m\frac{(t_j/x_k)_\infty}{(t_jq^{\beta_k}/x_k)_\infty},\\ &\prod_{1\le i<j\le n}\frac{(q_j^{\gamma'_{i,j}t_j}/t_i)_\infty}{(q_j^{\gamma_{i,j}t_j}/t_i)_\infty}\quad\text{for } t=(t_1,\dots,t_n)\in\overline{X},\end{aligned} \tag{1.3}$$

where we fix $(x_1,\dots,x_m)\in(\mathbf{C}^*)^m$ and $(\alpha_1,\dots,\alpha_n,\beta_1,\dots,\beta_m,\gamma_{i,j},\gamma'_{i,j})\in\mathbf{C}^{m+n^2}$. $(x)_\infty$ denotes the infinite product $\prod_{\nu=0}^\infty(1-xq^\nu)$. $(x)_k$ denotes $(x)_\infty/(xq^k)_\infty$ for $k\in\mathbf{Z}$.

In this note we consider a general Jackson integral for Φ and give the connection formula for it as a linear combination of a finite number of the fundamental ones giving special asymptotics for $\alpha_j\to+\infty$ (see Theorem in §4). We also give the holonomic

*Received by the editors March 12, 1992; accepted for publication (in revised form) June 11, 1993.

†Department of Mathematics, Nagoya University, Nagoya 461-01, Japan.

q-difference equations in the parameters $u_j = q^{\alpha_j}$ and x_k, which are satisfied by all Jackson integrals for Φ (see (3.6) and (3.7)).

A holonomic q-difference system has finite-dimensional solutions over the field of pseudoconstants (see [A3]). Given a basis of solutions for the system there arises a problem of expressing a particular solution as a linear combination of them. This problem is generally called a *connection problem.*

We denote by X the subgroup of $\bar{X}$ isomorphic to $\mathbf{Z}^n$, consisting of the elements $q^\chi = (q^{\nu_1}, \dots, q^{\nu_n}) \in \bar{X}$ for $\chi = \sum_{j=1}^n \nu_j \chi_j, \nu_j \in \mathbf{Z}$, $\{\chi_j\}_{1 \le j \le n}$ being a canonical basis of X. In the sequel we put

$$\gamma'_{j,i} = 1 - \gamma_{i,j} \quad \text{and} \quad \gamma_{j,i} = 1 - \gamma'_{i,j} \quad \text{for } i < j \tag{1.4}$$

so that we have $\gamma'_{i,j} + \gamma_{j,i} = 1$ for any pair $i \neq j$.

We do not postulate any further essential condition for $\gamma_{i,j}$. In the case where $\gamma_{i,j} = \gamma$, $\gamma'_{i,j} = \gamma'$, and where $\gamma - \gamma'$ are positive integers, and if Φ reduces to a polynomial, relevant integrals have been investigated in [A6], [K1], and [K2]. But in that case there is no connection problem in our sense.

For an arbitrary permutation σ of the n figures $\{1, 2, \dots, n\}$, i.e., $\sigma \in \mathrm{S}_n$ (the symmetric group of nth degree) we put the operations on the function $\Phi(t) = \Phi(t|\alpha, \gamma)$ depending on t, α, and γ:

$$\sigma\Phi(t|\alpha,\gamma) = \Phi(\sigma^{-1}(t|\alpha,\gamma)), \tag{1.5}$$

where $\sigma^{-1}(t|\alpha,\gamma) = (\tilde{t}|\tilde{\alpha},\tilde{\gamma})$ is defined as

$$\sigma^{-1}(t|\alpha,\gamma) = (\tilde{t}|\tilde{\alpha},\tilde{\gamma}) \tag{1.6}$$

for

$$\tilde{t}_j = t_{\sigma(j)},$$
$$\tilde{\alpha}_j = (\sigma\alpha)_j = \alpha_{\sigma(j)} + \sum_{\substack{k<j \\ \sigma(j)<\sigma(k)}} (\gamma'_{\sigma(j),\sigma(k)} - \gamma_{\sigma(j),\sigma(k)}) - \sum_{\substack{k>j \\ \sigma(j)>\sigma(k)}} (\gamma'_{\sigma(k),\sigma(j)} - \gamma_{\sigma(k),\sigma(j)}).$$

For a canonical system of generators $\{\sigma_r\}_{1 \le r \le n}$ of S_n, with the relations $\sigma_r^2 = 1$ and $\sigma_r\sigma_{r+1}\sigma_r = \sigma_{r+1}\sigma_r\sigma_{r+1}$, $\sigma_r(t|\alpha,\gamma) = (\tilde{t}|\tilde{\alpha},\tilde{\gamma})$ given by $\tilde{t}_j = t_j, \tilde{\alpha}_j = \alpha_j$ for $j \neq r$, $r+1$, and $\tilde{t}_r = t_{r+1}, \tilde{t}_{r+1} = t_r, \tilde{\alpha}_r = \alpha_{r+1} + \gamma_{r,r+1} - \gamma'_{r,r+1}$, $\alpha_{r+1} = \alpha_r + \gamma'_{r,r+1} - \gamma_{r,r+1}$, respectively. Remark that $\tilde{\alpha}_r + \tilde{\alpha}_{r+1} = \alpha_r + \alpha_{r+1}$.

The following associativity holds:

$$\sigma'\sigma\Phi(t|\alpha,\gamma) = \sigma'(\sigma\Phi)(t|\alpha,\gamma) \tag{1.7}$$

for any two elements $\sigma, \sigma' \in \mathrm{S}_n$, i.e., $\sigma^{-1}\sigma'^{-1}(t|\alpha,\gamma) = \sigma^{-1}(\sigma'^{-1}(t|\alpha,\gamma))$. Indeed, to prove (1.7) it is sufficient to show it for $\sigma' = \sigma_r$ only. In this special case $((\sigma^{-1}\sigma_r)(t))_h$ equals $\alpha_{\sigma(h)} + \sum_{\substack{k<h \\ \sigma(h)<\sigma(k)}} (\gamma'_{\sigma(h),\sigma(k)} - \gamma_{\sigma(h),\sigma(k)}) - \sum_{\substack{k>h \\ \sigma(h)>\sigma(k)}} (\gamma'_{\sigma(k),\sigma(h)} - \gamma_{\sigma(k),\sigma(h)})$ if $\sigma(h) \neq r, r+1$, and equals

$$\alpha_{r+1} + (\gamma_{r,r+1} - \gamma'_{r,r+1}) + \sum_{\substack{k<h \\ \sigma(h)>\sigma(k)}} (\gamma'_{\sigma(h),\sigma(k)} - \gamma_{\sigma(h),\sigma(k)}) - \sum_{\substack{k>h \\ \sigma(h)>\sigma(k)}} (\gamma'_{\sigma(k),\sigma(h)} - \gamma_{\sigma(k),\sigma(h)})$$

or

$$\alpha_r + (\gamma_{r,r+1} - \gamma'_{r,r+1}) + \sum_{\substack{k<h \\ \sigma(h)>\sigma(k)}} (\gamma'_{\sigma(h),\sigma(k)} - \gamma_{\sigma(h),\sigma(k)}) - \sum_{\substack{k>h \\ \sigma(h)>\sigma(k)}} (\gamma'_{\sigma(k),\sigma(h)} - \gamma_{\sigma(k),\sigma(h)}),$$

according to whether $\sigma(h) = r$ or $r+1$. Hence (1.7) holds for $\sigma' = \sigma_r$. The functions

$$U_\sigma(t) = \sigma\Phi(t|\alpha,\gamma)/\Phi(t|\alpha,\gamma) \quad \text{for } \sigma \in \mathrm{S}_n, \tag{1.8}$$

are pseudoconstants, i.e.,

$$Q^\chi U_\sigma = U_\sigma(t) \quad \text{for any } \chi \in X, \tag{1.9}$$

where $Q^\chi f(t) = Q_1^{\nu_1} \cdots Q_n^{\nu_n} f(t)$ denotes the shift operator $f(q^\chi t)$ by the element χ. Q_j denotes the partial q-difference operator on the jth coordinate t_j. In particular,

$$U_{\sigma_r}(t) = \left(\frac{t_{r+1}}{t_r}\right)^{\gamma_{r,r+1}-\gamma'_{r,r+1}} \frac{\theta\left(q^{\gamma_{r,r+1}} t_{r+1}/t_r\right)}{\theta\left(q^{\gamma'_{r,r+1}} t_{r+1}/t_r\right)}. \tag{1.10}$$

More generally, $U_\sigma(t)$ can be expressed as

$$U_\sigma(t) = \prod_{\substack{i<j \\ \sigma^{-1}(i)>\sigma^{-1}(j)}} \left(\frac{t_j}{t_i}\right)^{\gamma_{i,j}-\gamma'_{i,j}} \frac{\theta\left(q^{\gamma_{i,j}} t_j/t_i\right)}{\theta\left(q^{\gamma'_{i,j}} t_j/t_i\right)}, \tag{1.11}$$

To prove (1.11) we first remark that $\{U_\sigma(t)\}_{\sigma\in \mathrm{S}_n}$ satisfy the cocycle condition

$$U_{\sigma\sigma'}(t) = U_\sigma(t) \cdot \sigma U_{\sigma'}(t), \tag{1.12}$$

because

$$U_{\sigma\sigma'}(t) = \frac{\sigma\sigma'\Phi(t|\alpha,\gamma)}{\Phi(t|\alpha,\gamma)} = \frac{\sigma\sigma'\Phi(t|\alpha,\gamma)}{\sigma(t|\alpha,\gamma)} \cdot \frac{\sigma\Phi(t|\alpha,\gamma)}{\Phi(t|\alpha,\gamma)} = \sigma U_\sigma(t) \cdot U_{\sigma'}(t).$$

We denote the right-hand side of (1.11) by $\tilde{U}_\sigma(t)$. We must prove that $U_\sigma(t)$ coincides with $\tilde{U}_\sigma(t)$. When $\sigma = \sigma_r$ it is obvious from (1.10). We prove (1.11) by induction on the length of the reduced product expression of σ by the generators σ_r. So we have only to verify (1.12) for $\tilde{U}_\sigma$, $\tilde{U}'_\sigma(t)$ with $\sigma' = \sigma_r$. In this case, first suppose that $\sigma^{-1}(h) = r$ and $\sigma^{-1}(k) = r+1$ for $h < k$. Then

$$\tilde{U}_{\sigma'\sigma_r}(t) = \prod_{\substack{1\le i<j\le n \\ \sigma^{-1}(i)>\sigma^{-1}(j) \\ i\neq h \text{ or } j\neq k}} \left(\frac{t_j}{t_i}\right)^{\gamma_{i,j}-\gamma'_{i,j}} \frac{\theta\left(q^{\gamma_{i,j}} t_j/t_i\right)}{\theta\left(q^{\gamma'_{i,j}} t_j/t_i\right)} \cdot \left(\frac{t_k}{t_h}\right)^{\gamma_{h,k}-\gamma'_{h,k}} \frac{\theta\left(q^{\gamma_{h,k}} t_k/t_h\right)}{\theta\left(q^{\gamma'_{h,k}} t_k/t_h\right)}, \tag{1.13}$$

which is obviously equal to $\tilde{U}_{\sigma'}(t) \cdot \sigma'\tilde{U}_{\sigma_r}(t)$. The case where $h > k$ can be verified similarly. If σ' has a shorter length than σ, by induction hypothesis, $\tilde{U}'_\sigma(t)$ and $\sigma'\tilde{U}_{\sigma_r}(t)$ coincide with $U_{\sigma'}(t)$ and $\sigma'\tilde{U}_{\sigma_r}(t)$, respectively. Hence $\tilde{U}_\sigma(t)$ coincides with $U_\sigma(t)$.

2. Jackson integrals and α-stable cycles. We now want to define α-stable (or α-unstable) cycles giving special asymptotics of Jackson integrals for $\alpha_j \to \pm\infty$.

We assume the following condition, (C).

(C) For an arbitrary sequence of $(r+1)$ different figures $i_0, i_1, \dots, i_r$, the sum $c_{i_0,i_1} + \cdots + c_{i_r,i_0} \notin \mathbf{Z}$, where $c_{i,j}$ denote $\gamma_{i,j}$ or $\gamma'_{i,j}$ for $i,j \geq 1$ and $c_{0,j} = 1 + \log_q x_k$ or $-\beta_k + \log_q x_k$, respectively (we then put $c_{j,0} = -\log_q x_k$ or $1 + \beta_k - \log_q x_k$).

This condition implies that Φ has only simple poles of normal crossings. It is essential in our subsequent argument.

We denote by $[0, \xi\infty]_q$ the countable subset of $\bar{X}$, the X-orbit of a point $\xi \in \bar{X}$: $[0, \xi\infty]_q = \{q^\chi \cdot \xi | \chi \in X\}$. Then the Jackson integral of a function f on $\bar{X}$ over $[0, \xi\infty]_q$ is by definition equal to

$$\int_{[0,\xi\infty]_q} f(t)\tilde{\omega} = (1-q)^n \sum_{\chi \in X} f(q^\chi \cdot \xi) \quad \text{for } \tilde{\omega} = \frac{d_q t_1}{t_1} \wedge \cdots \wedge \frac{d_q t_n}{t_n}, \tag{2.1}$$

provided it exists (see [A2], [A3], and [G1] for various versions of Jackson integrals).

We are now considering the integral (2.1) for $f(t) = \Phi(t)$:

$$J = \int_{[0,\xi\infty]_q} \Phi(t)\tilde{\omega}. \tag{2.2}$$

Since $U_\sigma(t)$ is a pseudoconstant, as an immediate consequence of (2.2), we get the following.

LEMMA 1.

$$\int_{[0,\xi\infty]_q} \Phi\tilde{\omega} = U_\sigma(\xi)^{-1} \int_{[0,\xi\infty]_q} \sigma\Phi\tilde{\omega}, \tag{2.3}$$

provided both sides are convergent.

In fact, the right-hand side of (2.3) equals $U_\sigma(\xi)^{-1} \sum_{\chi \in X} \sigma\Phi(\xi q^\chi) = U_\sigma(\xi)^{-1} \sum_{\chi \in X} U_\sigma(\xi q^\chi)\Phi(\xi q^\chi) = \sum_{\chi \in X} \Phi(\xi q^\chi)$ since $U_\sigma(\xi q^\chi) = U_\sigma(\xi)$.

We are going to define two kinds of special cycles Y_Γ and Y^*_Γ, which are countable subsets of $[0, \xi\infty]_q$ by particular choices of ξ associated with certain forests Γ in graph theoretical sense.

DEFINITION 1. We consider a graph Γ with the following properties. We denote by $V(\Gamma)$ and $E(\Gamma)$ the sets of vertices of Γ and edges of Γ, respectively. (i) The vertices of Γ consist of the variables $t_j, 1 \leq j \leq n$, and parameters $x_k, 1 \leq k \leq m$. They are all labelled, i.e., Γ is a labelled graph. (ii) Γ has n edges. (iii) Γ has neither loops nor proper circuits, i.e., Γ is a forest that is not necessarily connected. (iv) Each connected component T of Γ has only one vertex in $\{x_1, \dots, x_m\}$, i.e., $V(T) \cap \{x_1, \dots, x_m\}$ consists of only one element. We call each of the vertices in $V(\Gamma) \cap \{x_1, \dots, x_m\}$ a root of Γ. We have called these graphs "admissible" in [A2].

It is known, thanks to the matrix tree theorem (see [M3]), that the number of such forests Γ is equal to $\kappa = m(m+n)^{n-1}$.

Each connected component of T of Γ is a tree and has a natural distance function between two vertices of T. For a vertex t_j of Γ, we denote by $t_{p(j)}, 0 \leq p(j) \leq n$, the unique vertex of Γ neighboring t_j and having a shorter distance by one from a root in the same connected component of T of Γ such that $t_j \in T$. We call $t_{p(j)}$ the predecessor of t_j. t_0 denotes a root itself.

If $p(j) < j$ holds for all vertices t_j, we call Γ, Y_Γ, or Y^*_Γ standard.

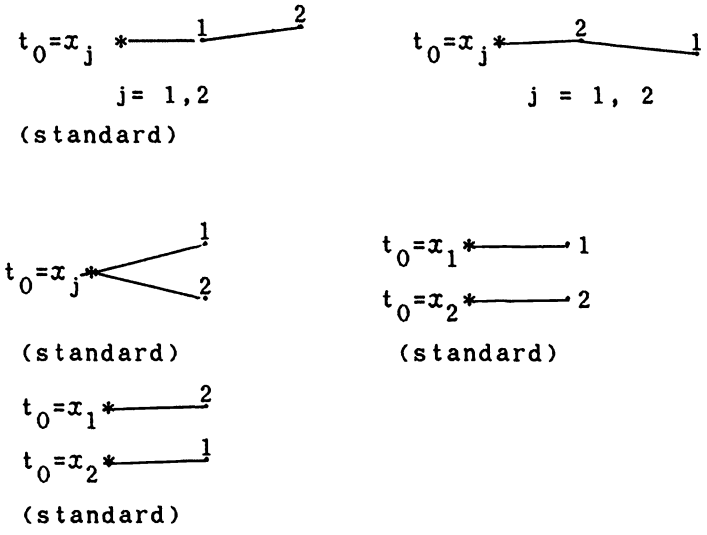


FIG. 1

For example, when $m = 2$ and $n = 2$, κ equals 8. A complete list is given in Fig. 1.

DEFINITION 2. Given an admissible graph Γ, we define a countable subset Y_Γ (or Y_Γ^*) in $\bar{X}$ consisting of the points $t \in \bar{X}$ satisfying the following properties:

(i) $t_j/t_{p(j)}$
$= q^{\gamma_{j,p(j)}}, q^{\gamma_{j,p(j)}+1}, q^{\gamma_{j,p(j)}+2}, \ldots$ (or $q^{-\gamma_{p(j),j}}, q^{-\gamma_{p(j),j}-1}, q^{-\gamma_{p(j),j}-2}, \ldots$);

(ii) $t_0 = qx_k \in V(\Gamma) \cap \{x_1, \ldots, x_m\}$ (or $t_0 = q^{-\beta_k}x_k$) for some $k, 1 \leq k \leq m$.

Likewise we denote by $\eta = (\eta_1, \ldots, \eta_n)$ the point of $\bar{X}$ satisfying the following properties:

(i)' $\eta_j/\eta_{p(j)} = q^{\gamma_{j,p(j)}}$ (or $q^{-\gamma_{p(j),j}}$).

(ii)' $\eta_0 = qx_k$ (or $q^{-\beta_k}x_k$) for some k.

We call this point η the base point of Γ, Y_Γ, and Y_Γ^*.

An admissible graph Γ denoted by Γ_η and the corresponding countable sets Y_Γ and Y_Γ^* are uniquely determined by its base point η. We may denote Y_Γ and Y_Γ^* by Y_η and Y_η^*, respectively, or more simply by $\langle\eta\rangle$. $\langle\eta\rangle$ is the subset of $[0, \eta\infty]_q$. We call the set $\langle\eta\rangle = Y_\eta$ (or Y_η^*) *α-stable* (or *α-unstable*) *cycle*. There are exactly $m(m+n)^{n-1}$ α-stable cycles (or α-unstable cycles).

Now we want to define a regularization of the cycles $[0, \xi\infty]_q$. Assume that $\xi_{j_1}/\xi_{p(j_1)} = q^{-\gamma_{p(j_1),j_1}}, \ldots, \xi_{j_r}/\xi_{p(j_r)} = q^{-\gamma_{p(j_r),j_r}}$ so that $\theta(q^{\gamma_{p(j_1),j_1}}\xi_{j_1}/\xi_{p(j_1)}) = \cdots = \theta(q^{\gamma_{p(j_r),j_r}}\xi_{j_r}/\xi_{p(j_r)}) = 0$ for $j_1 > p(j_1), \ldots, j_r > p(j_r)$. From the genericity of β_k and $\gamma_{i,j}$ stated in condition (C), we can choose the local coordinates $\tau_s = t_{j_s}/t_{p(j_s)}, 1 \leq s \leq n$, of $\bar{X}$ at $t = \xi$. A function

$$f(t) = \frac{g(t)}{\theta\left(q^{\gamma_{p(j_1),j_1}}t_{j_1}/t_{p(j_1)}\right)\cdots\theta\left(q^{\gamma_{p(j_r)j_r}}t_{j_r}/t_{p(j_r)}\right)} \tag{2.4}$$

has the residue at $t = \xi$ along $\tau_1 = q^{-\gamma_{p(j_1),j_1}}, \cdots \tau_r = q^{-\gamma_{p(j_r),j_r}}$ as follows:

$$\operatorname{res}_{t=\xi} f(t) = \frac{q^{-\sum_{s=1}^{r}\gamma_{p(j_s),j_s}}g(\xi)}{\theta'(1)^r} \tag{2.5}$$

for a function $g(t)$ holomorphic at ξ, where $\theta'(1) = -(q)_\infty^3$ (see [G2] for a general definition of residues). This will be denoted by reg $f(\xi)$. For $r = 0$, reg $f(\xi)$ reduces to $f(\xi)$.

When $\langle\eta\rangle$ is standard and α-stable, the Jackson integral of Φ over $\langle\eta\rangle$ is well defined provided it is summable, because Φ is holomorphic along $\langle\eta\rangle$. We have

$$\int_{\langle\eta\rangle} \Phi\tilde{\omega} = \int_{[0,\eta\infty]_q} \Phi\tilde{\omega}, \tag{2.6}$$

seeing that Φ vanishes on the complement $[0,\eta\infty]_q - \langle\eta\rangle$. If $\langle\eta\rangle$ is not standard, then there exists a permutation $\sigma \in \mathrm{S}_n$ such that $\sigma^{-1}\langle\eta\rangle$ is standard. We formally have

$$\begin{aligned}\int_{\langle\eta\rangle} \Phi\tilde{\omega} &= \textstyle\int_{\langle\eta\rangle} U_\sigma(t)^{-1}\sigma\Phi(t|\alpha,\gamma)\tilde{\omega} \\ &= U_\sigma(\eta)^{-1}\int_{\sigma^{-1}\langle\eta\rangle} \hat{\sigma}\Phi(t|\alpha,\gamma)\tilde{\omega},\end{aligned} \tag{2.7}$$

where we denote $\hat{\sigma}\Phi = \Phi(t|\sigma^{-1}(\alpha,\gamma))$. $\sigma^{-1}\langle\eta\rangle$ is also standard for $\hat{\sigma}\Phi(t|\alpha,\gamma)$ when $\int_{\sigma^{-1}\langle\eta\rangle} \hat{\sigma}\Phi(t|\alpha,\gamma)\tilde{\omega}$ is well defined. But $U_\sigma(\eta)^{-1}$ has no meaning because the function $U_\sigma^{-1}(t)$ has poles on $\langle\eta\rangle$. By replacing $U_\sigma^{-1}(\eta)$ by the residue of $U_\sigma^{-1}(t)$ at $t = \eta$ as in (2.5), we have the regularization of (2.6):

$$\mathrm{reg}\int_{\langle\eta\rangle} \Phi\tilde{\omega} = \int_{\sigma^{-1}\langle\eta\rangle} \hat{\sigma}\Phi(t|\alpha,\gamma)\cdot[\mathrm{reg}\, U_\sigma(\eta)^{-1}]. \tag{2.8}$$

This is also equivalent to taking residues with respect to the variables $\tau_1,\dots,\tau_r$ and then doing Jackson integrals with respect to the remaining coordinates $\tau_{r+1},\dots,\tau_n$.

LEMMA 2. *Equation* (2.8) *can be represented by a contour integral avoiding poles of* Φ. *In this sense the corresponding cycle will be denoted by* reg $\langle\eta\rangle$*:*

(2.9)

$$\begin{aligned}\mathrm{reg}\int_{\langle\eta\rangle} \Phi\tilde{\omega} &= \int_{\mathrm{reg}\ \langle\eta\rangle} \Phi\tilde{\omega} \\ &= (1-q)^r \int_{[0,\tilde{\eta}\infty]_q} \frac{d_q\tau_{r+1}\wedge\cdots\wedge d_q\tau_n}{\tau_{r+1}\cdots\tau_n} \sum_{\nu_j\in\mathbf{Z}}{}', \quad \mathrm{res}_{\nu_1,\dots,\nu_r}\left(\frac{\Phi}{t_1\cdots t_n}\right) \\ &\quad / \ [\mathrm{reg}\, U_\sigma(\eta)^{-1}],\end{aligned}$$

where we take the residues at $\tau_1 = q^{-\gamma_{p(j_1),j_1}-\nu_1},\dots,\tau_r = q^{-\gamma_{p(j_r),j_r}-\nu_r}$. $\tilde{\eta}$ *denotes the point* $(\eta_{j_{r+1}}/\eta_{p(j_{r+1})},\dots,\eta_{j_n}/\eta_{p(j_n)}) \in (\mathbf{C}^*)^{n-r}$.

In fact this can be proved by successive applications of the following lemma on a one-dimensional Jackson integral.

LEMMA 3.

$$\begin{aligned} x\ \mathrm{res}_{\xi=x^{-1}} &\int_{[0,\xi\infty]_q} \frac{\theta(txq^\lambda)}{\theta(tx)}\varphi(t)d_qt \\ &= (1-q)\sum_{n=-\infty}^{\infty} \mathrm{res}_{t=x^{-1}q^{-n}}\left[\frac{\theta(txq^\lambda)}{\theta(tx)}\varphi(t)\right] \\ &= x^{-\lambda}\frac{\theta(q^\lambda)}{\theta'(1)}\int_{[0,x^{-1}\infty]_q} t^{-\lambda}\varphi(t)d_qt.\end{aligned} \tag{2.10}$$

In particular, when $\langle\eta\rangle$ is a standard α-unstable cycle, r equals n and we can take as σ the permutation

$$\sigma_0 = \begin{pmatrix} 1,2,\dots,n \\ n,n-1,\dots,1 \end{pmatrix}.$$

The terminologies of α-stable and α-unstable cycles may be justified by the following proposition (see [A2] and [A3]).

PROPOSITION 1. *We put $\alpha_j = \alpha'_j + N\omega_j$ for α'_j and ω_j fixed such that $\omega_j \in \mathbf{Z}^+$. In case of a stable cycle $\langle\eta\rangle$, there exists a $\sigma \in \mathrm{S}_n$ such that $\sigma^{-1}\langle\eta\rangle$ is standard. If $N \in \mathbf{Z}^+$ tends to $+\infty$, then* (2.8) *or* (2.9) *has an asymptotic expansion*

$$\begin{aligned} \operatorname{reg} \int_{\langle\eta\rangle} \Phi\tilde{\omega} &= (1-q)^n \cdot \sigma\Phi(\eta|\alpha,\gamma)\left(1+O\left(\frac{1}{N}\right)\right) \\ &= \eta_1^{\alpha_1}\cdots\eta_n^{\alpha_n} R_+(\eta)\left(1+O\left(\frac{1}{N}\right)\right), \end{aligned} \tag{2.11}$$

where $R_+(\eta)(\neq 0)$ does not depend on α. In particular, if $\langle\eta\rangle$ is itself standard, then

$$\int_{\langle\eta\rangle} \Phi\tilde{\omega} = (1-q)^n \Phi(\eta|\alpha,\gamma)\left(1+O\left(\frac{1}{N}\right)\right). \tag{2.12}$$

Similarly, in case of an α-unstable cycle $\langle\eta\rangle$, for $N \to -\infty$,

$$\operatorname{reg} \int_{\langle\eta\rangle} \Phi\tilde{\omega} = \eta_1^{\alpha_1}\cdots\eta_n^{\alpha_n} R_-(\eta)\left(1+O\left(\frac{1}{N}\right)\right), \tag{2.13}$$

where $R_-(\eta)(\neq 0)$ does not depend on α.

In other words α-stable (or α-unstable) cycles correspond to the simplest asymptotic expansions for $\alpha_j \to +\infty$ (or $-\infty$).

It has been proved in [A2] *and* [A3] *that the α-stable (or α-unstable) cycles give a dual basis of the de Rham cohomology $H^n(\overline{X},\Phi,\nabla_q)$ associated with the Jackson integrals* (2.2). *The dimension of $H^n(\overline{X},\Phi,\nabla_q)$ is equal to $m(m+n)^{n-1}$.*

So an arbitrary cycle $[0,\xi\infty]_q$ can be described as a linear combination of the $m(m+n)^{n-1}\alpha$-stable cycles reg Y_Γ (or α-unstable cycles reg Y^*_Γ) over the field of pseudoconstants in the parameters $u_h, \beta_i, \gamma_{j,k}$, and ξ_ℓ :

$$\int_{[0,\xi\infty]_q} \Phi(t)\tilde{\omega} = \sum c_\Gamma \int_{\operatorname{reg}(Y_\Gamma)} \Phi(t)\tilde{\omega} \tag{2.14}$$

$$= \sum c^*_\Gamma \int_{\operatorname{reg}(Y^*_\Gamma)} \Phi(t)\tilde{\omega} \tag{2.15}$$

for some pseudoconstants c_Γ and c^*_Γ.

We evaluate c_Γ and c^*_Γ explicitly in §4.

3. Holonomic q-difference equations. The notion of holonomic q-difference equations has been investigated in [A1], [A2], [A3], and [S]. In [F] it has been discussed in relation to so-called R-matrices and quantum KZ equations. Here we want to explicitly give the holonomic system satisfied by (2.2) in the parameters u_j and x_k.

We denote by $\tilde{Q}_j (1 \le j \le n)$, and $\tilde{Q}_{n+j} (1 \le j \le m)$ the q-shift operators for functions of $(u,x) = (u_1,\dots,u_n;x_1,\dots,x_m)$ induced by the displacements $u_j \to u_j q$ and $x_k \to x_k q$, respectively. Then we have $\tilde{Q}_j \Phi(t) = u_j \Phi(t)$ and

$$\tilde{Q}_{n+k}\Phi(t) = \prod_{j=1}^{n} \frac{(1-q^{-1}t_j/x_k)}{(1-q^{\beta_k-1}t_j/x_k)}\Phi(t).$$

Now

$$Q^{\chi}\Phi(t) = b_{\chi}(t)\Phi(t), \qquad \chi = \sum_{j=1}^{n} \nu_j \chi_j \in X, \tag{3.1}$$

where $b_{\chi}(t)$ is represented as $u^{\chi} b_{\chi}^{+}(t)/b_{\chi}^{-}(t)$ for $u^{\chi} = u_1^{\nu_1}\cdots u_n^{\nu_n}$, such that $b_{\chi}^{+}(t)$ and $b_{\chi}^{-}(t)$ are Laurent polynomials in $t, \prod_{k=1}^{m}\prod_{\nu_j \ge 0} (t_j q^{\beta_k}/x_k)_{\nu_j} \cdot \prod_{\nu_j<0}(t_j q^{\nu_j}/x_k)_{-\nu_j}$ and $\prod_{k=1}^{m}\prod_{\nu_j\ge 0}(t_j/x_k)_{\nu_j}\cdot\prod_{\nu_j<0}(t_j q^{\beta_k+\nu_j}/x_k)_{-\nu_j}$, respectively. We take as φ the Laurent polynomial $Q^{-\chi}b_{\chi}^{-}(t)$ in t:

$$\varphi(t) = \prod_{k=1}^{m}\prod_{\nu_j\ge 0}\left(t_j q^{-\nu_j}/x_k\right)_{\nu_j} \prod_{\nu_j<0}\left(t_j q^{\beta_k}/x_k\right)_{-\nu_j}. \tag{3.2}$$

Then we get $Q^{\chi}(\Phi(t)\varphi(t)) = b_{\chi}(t)\Phi(t)\cdot Q^{\chi}\varphi(t) = b_{\chi}^{+}(t)\Phi(t)$. Since

$$\int_{[0,\xi\infty]_q} \Phi(t)\varphi(t)\tilde{\omega} = \int_{[0,\xi\infty]_q} Q^{\chi}(\Phi(t)\varphi(t))\tilde{\omega}, \tag{3.3}$$

we have

$$\int_{[0,\xi\infty]_q} \Phi(t)Q^{-\chi}b_{\chi}^{-}(t)\tilde{\omega} - \int_{[0,\xi\infty]_q} \Phi(t)u^{\chi}b_{\chi}^{+}(t)\tilde{\omega} = 0. \tag{3.4}$$

For an arbitrary Laurent polynomial $f(t)$ in t, the equality $f(\tilde{Q}_1,\dots,\tilde{Q}_n)\Phi(t) = f(t_1,\dots,t_n)\Phi(t)$ holds. Hence by the definition of Jackson integrals, (3.4) is equivalent to the equations

$$Q^{-\chi}b_{\chi}^{-}(\tilde{Q})J - u^{\chi}b_{\chi}^{+}(\tilde{Q})J = 0 \tag{3.5}$$

for any $\chi \in X$. Since the set $\{\pm\chi_j, 1 \le j \le n\}$ consists of primitive corner vectors spanning rational polyhedral cones of the fan in the theory of torus embeddings (see [O] for the definition) associated with equation (3.5), (3.5) is equivalent to the following system of q-difference equations:

$$\left(Q_j^{-1}b_{\chi_j}^{-}\right)(\tilde{Q})J - u_j b_{\chi_j}^{+}(\tilde{Q})J = 0, \qquad 1 \le j \le n. \tag{3.6}$$

(See [A3] for more details.) In the same way we have

$$\left\{\prod_{j=1}^{n}\left(1-q^{\beta_k-1}x_k^{-1}\tilde{Q}_j\right)\tilde{Q}_{k+n} - \prod_{j=1}^{n}\left(1-q^{-1}x_k\tilde{Q}_j\right)\right\}J = 0. \tag{3.7}$$

One can prove that the system q-difference equations (3.6) and (3.7) in the parameters u_j and x_k are *holonomic* in the sense that they have only finite-dimensional solutions over the field of pseudoconstants in u_j and x_k. In fact its dimension is equal to $m(m+n)^{n-1}$, which coincides with the dimension of $H^n(\overline{X},\Phi,\nabla_q)$ (see Theorem 2 in [A3]).

4. Main result. We denote by $\Psi_n(\xi,\eta|\alpha)$ the function in the parameters ξ_j,η_j, u_j, and β_k, which is expressed by

$$\Psi_n(\xi,\eta|\alpha)=(1-q)^n\prod_{j=1}^{n}\left\{\left(\frac{\xi_j}{\eta_j}\right)^{\alpha_j}\frac{(q)_\infty^3\theta\left(q^{\alpha_j+\cdots+\alpha_n+1}\xi_j\eta_{j-1}/(\xi_{j-1}\eta_j)\right)}{\theta\left(q^{\alpha_j+\cdots+\alpha_n+1}\right)\theta(q\xi_j\eta_{j-1}/(\xi_{j-1}\eta_j))}\right\},\tag{4.1}$$

where we put $\xi_0=\eta_0=1$. This can be seen to be a pseudoconstant.

We can now state our main theorem.

THEOREM. *For a generic $\xi\in\overline{X}$, we have*

$$\int_{[0,\xi\infty]_q}\Phi\tilde{\omega}=\sum_{\langle\eta\rangle}([0,\xi\infty]_q:\operatorname{reg}\langle\eta\rangle)_\Phi\int_{\operatorname{reg}\langle\eta\rangle}\Phi\tilde{\omega}\tag{4.2}$$

for certain pseudoconstants $([0,\xi\infty]_q:\operatorname{reg}\langle\eta\rangle)_\Phi$, where $\langle\eta\rangle$ ranges over the set of all α-unstable cycles. If $\langle\eta\rangle$ is standard, then $([0,\xi\infty]_q:\langle\eta\rangle)_\Phi$ equals the pseudoconstant $\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)$ defined by the sum

$$\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)=\sum_{\sigma\in S_n}\sigma\Psi_n(\xi,\eta|\alpha)\cdot U_\sigma^{-1}(\xi)U_\sigma(\eta).\tag{4.3}$$

In case of a nonstandard α-unstable cycle $\langle\eta\rangle$, there exists a $\rho\in S_n$ such that $\rho^{-1}\langle\eta\rangle$ is standard and α-unstable. We then have

$$\begin{aligned}([0,\xi\infty]_q:\operatorname{reg}\langle\eta\rangle)_\Phi&=(\rho^{-1}[0,\xi\infty]_q:\rho^{-1}\langle\eta\rangle)_{\hat{\rho}\Phi}U_\rho(\xi)^{-1}\operatorname{reg}[U_\rho(\eta)]\\&=\rho\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)U_\rho(\xi)^{-1}\operatorname{reg}[U_\rho(\eta)],\end{aligned}\tag{4.4}$$

where

$$\int_{\operatorname{reg}\langle\eta\rangle}\Phi\tilde{\omega}=\int_{(\sigma_0\rho)^{-1}\langle\eta\rangle}\hat{\sigma}_0\hat{\rho}\Phi\tilde{\omega}/[\operatorname{reg}U_{\sigma_0\rho}(\eta)^{-1}]\tag{4.5}$$

by definition. $\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)$ is quasi-symmetric, i.e.,

$$\sigma\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)=\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)U_\sigma(\xi)U_\sigma^{-1}(\eta).\tag{4.6}$$

$\operatorname{reg}\langle\eta\rangle$ is not unique and depends on the choice of an element $\rho\in S_n$ such that $\rho^{-1}\langle\eta\rangle$ is standard. We may choose as ρ the unique element having a minimal expression in terms of the generators $\sigma_1,\sigma_2,\dots,\sigma_n$.

$\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)$ can also be written as the quotient of theta polynomials:

$$\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)=\frac{H(\xi,\eta|\alpha,\gamma)}{G(\xi,\eta|\alpha,\gamma)},\tag{4.7}$$

where

$$\begin{aligned}G(\xi,\eta|\alpha,\gamma)=&\prod_{j=1}^{n}\left(\frac{\xi_j}{\eta_j}\right)^{j-1-a_j}\cdot\prod_{0\le k<j\le n}\theta\left(q\xi_j\eta_k/(\xi_k\eta_j)\right)\\&\cdot\prod_{1\le k<j\le n}\left\{\theta\left(q^{\gamma_{k,j}}\xi_j/\xi_k\right)\theta\left(q^{\gamma'_{k,j}}\eta_j/\eta_k\right)\right\}\end{aligned}\tag{4.8}$$

satisfies the quasi skew-symmetric property : $\sigma G(\xi,\eta|\alpha,\gamma) = \operatorname{sgn}\sigma \cdot G(\xi,\eta|\alpha,\gamma)U_\sigma(\xi)^{-1}$ $U_\sigma(\eta)$ for $\sigma \in \mathrm{S}_n$. $H(\xi,\eta|\alpha,\gamma)$ is skew-symmetric:

$$\sigma H(\xi,\eta|\alpha,\gamma) = \operatorname{sgn}\sigma \cdot H(\xi,\eta|\alpha,\gamma). \tag{4.9}$$

From (4.3), $H(\xi,\eta|\alpha,\gamma)$ can be expressed as an alternating sum

$$H(\xi,\eta|\alpha,\gamma) = \sum_{\sigma\in\mathrm{S}_n} \sigma E(\xi,\eta|\alpha,\gamma)\operatorname{sgn}\sigma, \tag{4.10}$$

$$\begin{aligned} E(\xi,\eta|\alpha,\gamma) = (1-q)^n \prod_{j=1}^{n} & \left\{ \frac{\theta\left(q^{\alpha_j+\cdots+\alpha_n+1}\xi_j\eta_{j-1}/(\xi_{j-1}\eta_j)\right)}{\theta\left(q^{\alpha_j+\cdots+\alpha_n+1}\right)} \left(\frac{\xi_j}{\eta_j}\right)^{j-1} \right\} \\ & \cdot \prod_{\substack{0\le k<j\le n \\ j-k\ge 2}} \theta(q\xi_j\eta_k/(\xi_k\eta_j)) \\ & \cdot \prod_{1\le k<j\le n} \left\{ \theta\left(q^{\gamma_{k,j}}\xi_j/\xi_k\right)\theta\left(q^{\gamma'_{k,j}}\xi_k/\xi_j\right)\right\}. \end{aligned} \tag{4.11}$$

LEMMA 4. *As a function of ξ and η in $\overline{X}$, $H(\xi,\eta|\alpha,\gamma)$ is divided out by the product*

$$\prod_{j=1}^{n}\left(\frac{\xi_j}{\eta_j}\right)^{j-1} \prod_{0\le k<j\le n} \theta\left(q\frac{\xi_j\eta_k}{\xi_k\eta_j}\right)$$

in the ring of θ-polynomials on $\bar{X}\times\bar{X}$.

Proof. In fact, as a function of (ξ,η) on $\bar{X}\times\bar{X}$, $G(\xi,\eta) = G(\xi,\eta|\alpha,\gamma)$, and $H(\xi,\eta) = H(\xi,\eta|\alpha,\gamma)$ have the quasi-periodic properties for the q-shift operators $Q_r(\xi), Q_r(\eta)$ induced by the shifts $\xi_r \to \xi_r q$ and $\eta_r \to \eta_r q$, respectively. □

$$\begin{aligned} [Q_r(\xi)G(\xi,\eta)/G(\xi,\eta)] &= [Q_r(\xi)H(\xi,\eta)]/H(\xi,\eta) \\ &= q^{-n+r-\alpha_r+A_r} \cdot \prod_{\substack{k=1 \\ k\ne r}}^{n} \left(\xi_k^2\eta_r/(\xi_r^2\eta_k)\right), \end{aligned} \tag{4.12}$$

$$\begin{aligned} [Q_r(\eta)G(\xi,\eta)]/G(\xi,\eta) &= [Q_r(\eta)H(\xi,\eta)]/H(\xi,\eta) \\ &= q^{-n-r+2+\alpha_r+A'_r} \cdot \prod_{\substack{j=1 \\ j\ne r}}^{n} (\xi_r\eta_j^2/(\xi_j\eta_r^2)) \end{aligned} \tag{4.13}$$

for $A_r = -\sum_{1\le k<r}\gamma_{k,r} + \sum_{r<k\le n}\gamma_{r,k}$ and $A'_r = -\sum_{1\le k<r}\gamma_{r,k} + \sum_{r<k\le n}\gamma_{k,r}$, in view of (1.4). Furthermore, $H(\xi,\eta|\alpha,\gamma)$ vanishes if $\xi_2/\eta_2 = \xi_1/\eta_1$. This can be seen as follows. First we see that $-\sigma_1\sigma' E(\xi,\eta|\alpha,\gamma) + \sigma' E(\xi,\eta|\alpha,\gamma)$ vanishes for any $\sigma' \in \mathrm{S}_n$. Indeed, if $\sigma'(r) = 1$ and $\sigma'(s) = 2$ for some r, s such that $|r-s| > 1$, then $\sigma_1\sigma' E(\xi,\eta|\alpha,\gamma)$ and $\sigma' E(\xi,\eta|\alpha,\gamma)$ both vanish because the factor $\theta(q(\xi_2\eta_1/\xi_1\eta_2))$ or $\theta(q(\xi_1\eta_2/\xi_2\eta_1))$ appears in both. Otherwise we may suppose that $\sigma'(r) = 1$ and $\sigma'(r+1) = 2$. We put $\sigma'^{-1}(\alpha) = \tilde{\alpha}$. Then $\tilde{\alpha}_r = \alpha_1 + \sum_{\substack{1\le k<r \\ \sigma'(k)>r+1}} (\gamma'_{1,\sigma'(k)} - \gamma_{1,\sigma'(k)})$, $\tilde{\alpha}_{r+1} =$

$\alpha_2 + \sum_{\substack{1\le k<r \\ \sigma'(k)>r+1}} (\gamma'_{1,\sigma'(k)} - \gamma_{1,\sigma'(k)})$ and $\tilde{\xi}_r = \xi_1, \tilde{\xi}_{r+1} = \xi_2, \tilde{\eta}_r = \eta_1, \tilde{\eta}_{r+1} = \eta_2$, respectively.

Hence,

(4.14)

$$
\begin{aligned}
&-\sigma_1\sigma' E(\xi,\eta|\alpha,\gamma) + \sigma' E(\xi,\eta|\alpha,\gamma) \\
&= (1-q)^n (q)_\infty^{3n} \prod_{j=1}^{n} \left(\frac{\tilde{\xi}_j}{\tilde{\eta}_j}\right)^{j-1} \cdot \left(\frac{\xi_1\xi_2}{\eta_1\eta_2}\right)^{r-1} \\
&\cdot \prod_{\substack{j=1 \\ j\neq r,r+1,r+2}}^{n} \frac{\theta(q^{\tilde{\alpha}_j+\cdots+\tilde{\alpha}_n+1}\tilde{\xi}_j\tilde{\eta}_{j-1}/(\tilde{\xi}_{j-1}\tilde{\eta}_j))}{\theta(q^{\tilde{\alpha}_j+\cdots+\tilde{\alpha}_n+1})} \prod_{\substack{0\le k<j\le n \\ 2\le j-k, \\ (k,j)\neq(r-1,r+1) \text{ or } (r,r+2)}} \theta\left(q\frac{\tilde{\xi}_k\tilde{\eta}_j}{\tilde{\eta}_k\tilde{\xi}_j}\right) \\
&\cdot \left\{ \frac{\xi_2}{\eta_2} \frac{\theta(q^{\tilde{\alpha}_r+\cdots+\tilde{\alpha}_n+1}\xi_1\tilde{\eta}_{r-1}/(\tilde{\xi}_{r-1}\eta_1))\theta(q^{\tilde{\alpha}_{r+1}+\cdots+\tilde{\alpha}_n+1}\xi_2\eta_1/(\xi_1\eta_2))}{\theta(q^{\tilde{\alpha}_r+\cdots+\tilde{\alpha}_n+1})\,\theta(q^{\tilde{\alpha}_{r+1}+\cdots+\tilde{\alpha}_n+1})} \right. \\
&\cdot \frac{\theta(q^{\tilde{\alpha}_{r+2}+\cdots+\tilde{\alpha}_n+1}\tilde{\xi}_{r+2}\eta_2/(\xi_2\tilde{\eta}_{r+2}))}{\theta(q^{\tilde{\alpha}_{r+2}+\cdots+\tilde{\alpha}_n+1})} \theta\left(q\frac{\tilde{\xi}_{r-1}\eta_2}{\xi_2\tilde{\eta}_{r-1}}\right)\theta\left(q\frac{\xi_1\tilde{\eta}_{r+2}}{\eta_1\tilde{\xi}_{r+2}}\right) \\
&- \frac{\xi_1}{\eta_1} \frac{\theta(q^{\tilde{\alpha}_r+\cdots+\tilde{\alpha}_n+1}\xi_2\tilde{\eta}_{r-1}/(\tilde{\xi}_{r-1}\eta_2))\theta(q^{\tilde{\alpha}_{r+2}+\cdots+\tilde{\alpha}_n+1}\xi_{r+2}\eta_1/(\xi_1\eta_{r+2}))}{\theta(q^{\tilde{\alpha}_r+\cdots+\tilde{\alpha}_n+1})\theta(q^{\tilde{\alpha}_{r+2}+\cdots+\tilde{\alpha}_n+1})} \\
&\left. \cdot \frac{\theta(q^{\tilde{\alpha}_r+\tilde{\alpha}_{r+2}+\cdots+\tilde{\alpha}_n+1+\gamma'_{1,2}-\gamma_{1,2}}\xi_1\eta_2/(\xi_2\eta_1))}{\theta(q^{\tilde{\alpha}_r+\tilde{\alpha}_{r+2}+\cdots+\tilde{\alpha}_n+1+\gamma'_{1,2}-\gamma_{1,2}})} \theta\left(q\frac{\tilde{\xi}_{r-1}\eta_1}{\tilde{\eta}_{r-1}\xi_1}\right)\theta\left(q\frac{\xi_2\tilde{\eta}_{r+2}}{\eta_2\tilde{\xi}_{r+2}}\right) \right\}
\end{aligned}
$$

since $\sigma_1(\tilde{\alpha})_r = \tilde{\alpha}_{r+1} + \gamma_{1,2} - \gamma'_{1,2}, \sigma_1(\tilde{\alpha})_{r+1} = \tilde{\alpha}_r + \gamma'_{1,2} - \gamma_{1,2}$, and $\sigma_1(\tilde{\alpha})_j = \tilde{\alpha}_j$ otherwise. Hence (4.14) and $H(\xi,\eta|\alpha,\gamma)$ vanish for $\xi_1/\eta_1 = \xi_2/\eta_2$. Because $H(\xi,\eta|\alpha,\gamma)$ is symmetric in $(\xi,\eta|\alpha,\gamma)$, it also vanishes if $\xi_j/\eta_j = \xi_k/\eta_k$ for every $j \neq k$ and must be divided out by the factor $\theta(q\xi_j\eta_k/(\eta_j\xi_k))$. Hence the lemma follows.

Remark. Assume that η is standard. If Γ_η (see Definition 2) is a tree and if the vertices of Γ_η are totally ordered, then $U_\sigma(\eta)$ vanishes for all σ different from the identity. Hence $\tilde{\Psi}_n(\xi,\eta|\alpha,\gamma)$ reduces to $\Psi_n(\xi,\eta|\alpha,\gamma)$ itself. On the other hand, if every edge of Γ_η has a root in its ends, then all the terms $\sigma\Psi_\eta(\xi,\eta|\alpha), \sigma \in \mathrm{S}_n$, appear.

To obtain a similar formula to (4.2) with regards to the α-stable cycles Y_Γ in place of Y^*_Γ, we also consider the multiplicative function

$$
\Phi^*(t) = \prod_{j=1}^{n} t_j^{\alpha^*_j} \prod_{j=1}^{n}\prod_{k=1}^{m} \frac{(q^{1-\beta_k}t_jx_k)_\infty}{(qt_jx_k)\infty} \prod_{1\le i<j\le n} \frac{(q^{\gamma'_{j,i}}t_j/t_1)_\infty}{q^{\gamma_{j,i}}t_j/t_i)_\infty} \tag{4.15}
$$

for $\alpha^*_j = -\alpha_j - \beta_1 - \cdots - \beta_m + \sum_{1\le i<j}(\gamma'_{i,j} - \gamma_{i,j}) + \sum_{j<k\le n}(-\gamma'_{j,k} + \gamma_{j,k})$ and $\gamma^*_{i,j} = \gamma_{j,i}$. Then $\Phi(t)$ and $\Phi^*(t^{-1})$ satisfy the same q-difference equations:

$$
Q^\chi\Phi(t)/\Phi(t) = (Q^{-\chi}\Phi^*(t^{-1}))/\Phi^*(t^{-1}), \quad \text{for } \chi \in X. \tag{4.16}
$$

In fact the relation

$$
\Phi(t) = U_0(t)\Phi^*(t^{-1}) \tag{4.17}
$$

holds for the pseudoconstant

$$U_0(t) = \prod_{j=1}^{n}\prod_{k=1}^{m} t_j^{-\beta_k}\frac{\theta(t_j/x_k)}{\theta(q^{\beta_k}t_j/x_k)}\cdot\prod_{1\le i<j\le n}\left(\frac{t_j}{t_i}\right)^{\gamma'_{i,j}-\gamma_{i,j}}\frac{\theta(q^{\gamma'_{i,j}}t_j/t_i)}{\theta(q^{\gamma_{i,j}}t_j/t_i)}. \tag{4.18}$$

(Remark that $\gamma'_{r,s}+\gamma_{s,r}=1$ for $r\neq s$.) Hence, as in Lemma 1, we have the following.

LEMMA 5.

$$\int_{[0,\xi\infty]_q}\Phi(t)\tilde{\omega} = U_0(\xi)\int_{[0,\xi^{-1}\infty]_q}\Phi^*(t)\tilde{\omega}. \tag{4.19}$$

$\langle\eta\rangle$ is an α-stable (or unstable) cycle for Φ according to whether $\langle\eta^{-1}\rangle$ is an α-unstable (or stable) cycle for Φ^*. Hence the following.

PROPOSITION 2. *For an arbitrary α-stable cycle $\langle\eta\rangle$ for Φ,*

$$([0,\xi\infty]_q : \operatorname{reg}\langle\eta\rangle)_\Phi = ([0,\eta^{-1}\infty]_q : \operatorname{reg}\langle\eta^{-1}\rangle)_{\Phi^*}U_0(\xi)U_0(\eta)^{-1}. \tag{4.20}$$

In particular, if $\langle\eta\rangle$ is standard, then $\langle\eta^{-1}\rangle$ becomes a standard α-unstable cycle and we have

$$([0,\xi\infty]_q : \langle\eta\rangle)_\Phi = \tilde{\Psi}_n(\xi^{-1},\eta^{-1}|\alpha^*,\gamma^*)U_0(\xi)U_0(\eta)^{-1}. \tag{4.21}$$

By using these formulae and the theorem, one can also compute connection coefficients for all α-stable cycles.

5. Proof of main result. For $n=1$, Φ reduces to

$$\Phi = t_1^{\alpha_1}\prod_{k=1}^{m}\frac{(t_1/x_k)_\infty}{(t_1q^{\beta_k}/x_k)_\infty}. \tag{5.1}$$

In this case the following lemma, due to Mimachi, will play a central role in the sequel (for similar calculus see [M1] and [M2]).

LEMMA 6.

$$\begin{aligned}\tilde{\Psi}_1(\xi_1,\eta_1|\alpha) &= \Psi_1(\xi,\eta_1|\alpha)\\ &= (1-q)\left(\frac{\xi_1}{\eta_1}\right)^{\alpha_1}\frac{(q)_\infty^3\theta(q^{\alpha_1+1}\xi_1/\eta_1)}{\theta(q^{\alpha_1+1})\theta(q\xi_1/\eta_1)},\end{aligned} \tag{5.2}$$

where η_1 ranges over the set $\{x_1q^{-\beta_1},\dots,x_mq^{-\beta_m}\}$.

Proof. Using Lemma 3, we can express J as

$$J = -\frac{(1-q)(q)_\infty^3}{\theta(q^{\alpha_1+1})}\sum_{\ell=-\infty}^{+\infty}\operatorname{res}_{t_1=\xi_1q^\ell}\left\{\frac{\theta(t_1q^{\alpha_1+1}/\xi_1)}{\theta(t_1/\xi_1)t_1}\prod_{j=1}^{m}\frac{(t_1/x_j)_\infty}{(t_1q^{\beta_j}/x_j)_\infty}\right\}. \tag{5.3}$$

The right-hand side equals

$$\frac{(1-q)(q)_\infty^3}{\theta(q^{\alpha_1+1})}\sum_{j=1}^{n}\left(\frac{\xi_1q^{\beta_j}}{x_j}\right)^{\alpha_1}\frac{\theta(q^{\alpha_1+\beta_j+1}\xi_1/x_j)}{\theta(q^{\beta_j+1}\xi_1/x_j)}\int_{\mathcal{Z}^*_{q^{-\beta_j}x_j}}\Phi\tilde{\omega}, \tag{5.4}$$

where the integral part is, by definition, equal to

$$\sum_{\ell=0}^{+\infty} \operatorname{res}_{t_1=q^{-\beta_j-\ell_{x_j}}} (\Phi/t_1) \quad \text{i.e., reg } \langle \eta_1 \rangle = Z^*_{q^{-\beta_j} x_j}.$$

This means Lemma 6. □

Now we are going to prove the theorem.

Proof of the theorem. By the change of variables $t_1 = \tau_1, t_j = \tau_1 \tau_j$ for $j \geq 2$, (2.2) is rewritten as

$$\begin{aligned} J = & \int_{[0,(\xi_2/\xi_1,\dots,\xi_n/\xi_1)\infty]_q} \frac{d_q\tau_2}{\tau_2} \wedge \cdots \wedge \frac{d_q\tau_n}{\tau_n} \tau_2^{\alpha_2} \cdots \tau_n^{\alpha_n} \\ & \cdot \prod_{2\leq j\leq n} \frac{(q^{\gamma_{1,j}}\tau_j)_\infty}{(q^{\gamma'_{1,j}}\tau_j)_\infty} \prod_{2\leq i<j\leq n} \frac{(q^{\gamma'_{i,j}}\tau_j/\tau_i)_\infty}{(q^{\gamma_{i,j}}\tau_j/\tau_i)_\infty} \int_{[0,\xi_1\infty]_q} \Phi_1 \frac{d_q t_1}{t_1}, \end{aligned} \tag{5.5}$$

where Φ_1 denotes

$$t_1^{\alpha_1+\cdots+\alpha_n} \cdot \prod_{k=1}^{m} \frac{(t_1/x_k)_\infty}{(t_1 q^{\beta_k}/x_k)_\infty} \cdot \prod_{\substack{2\leq j\leq n \\ 1\leq k\leq m}} \frac{(t_1\tau_j/x_k)_\infty}{(t_1\tau_j q^{\beta_k}/x_k)_\infty}. \tag{5.6}$$

We fix $\tau_2, \dots, \tau_n$ for the moment and integrate (5.5) with respect to t_1 over the one-dimensional set $[0, \xi_1\infty]_q$. As a result of Lemma 1,

(5.7)

$$\begin{aligned} & \int_{[0,\xi_1\infty]_q} \Phi_1 \frac{d_q t_1}{t_1} \\ & = (1-q)(q)_\infty^3 \left\{ \sum_{k=1}^{m} \left[\left(\frac{\xi_1}{\eta_1}\right)^{\alpha_1+\cdots+\alpha_n} \frac{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\xi_1/\eta_1\right)}{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\right)\theta\left(q\frac{\xi_1}{\eta_1}\right)} \right.\right. \\ & \qquad \left.\left. \cdot \int_{\text{reg } <\eta_1>} \Phi \frac{d_q t_1}{t_1} \right]_{\eta_1 = x_k q^{-\beta_k}} \right. \\ & \quad + \sum_{j=2}^{n} \sum_{k=1}^{m} \left[\left(\frac{\xi_1}{\eta_1}\tau_j\right)^{\alpha_1+\cdots+\alpha_n} \right. \\ & \qquad \left.\left. \cdot \frac{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\xi_1\tau_j/\eta_1\right)}{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\right)\theta(q\frac{\xi_1}{\eta_1}\tau_j)} \int_{\text{reg } \langle \tau_j^{-1}\eta_1 \rangle} \Phi_1 \frac{d_q t_1}{t_1} \right]_{\eta_1 = x_k q^{-\beta_k}} - \beta_k \right\}. \end{aligned}$$

Since the function (5.2) of ξ_1 and η_1 is pseudoconstant and since τ_j ranges over the set $\xi_j q^{\nu_j}/\xi_1$ for $\nu_j = 0, \pm 1, \pm 2, \dots$, none of the coefficients in the right-hand side depend on either τ_j, i.e., τ_j may be replaced by ξ_j/ξ_1, respectively. By another change of variables, $t_1 = t_1$ and $t_j = t_1\tau_j$ for $j \geq 2, J$ can be expressed as

$$J = \sum_{r=1}^{n} J_r, \tag{5.8}$$

where

$$J_r = (1-q)^n (q)_\infty^{3n} \sum_{k=1}^{m} \left\{ \left[\left(\frac{\xi_r}{\eta_r} \right)^{\alpha_1+\cdots+\alpha_n} \frac{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\xi_r/\eta_r\right)}{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\right)\theta\left(q\frac{\xi_r}{\eta_r}\right)} \right.\right. \tag{5.9}$$
$$\left.\left. \int_{\mathrm{reg}\ [0,\xi\eta_r/\xi_r\infty]_q} \Phi\tilde{\omega} \right]_{\eta_r = x_k q^{-\beta_k}} \right\}, \qquad r \geq 2.$$

We denote by $W_r = \sigma_{r-1}\cdots\sigma_1 \cdot W_1$, where W_1 is the subgroup of S_n consisting of elements leaving the number 1 fixed, and by $\xi^{(r)}$ the $(n-1)$ tuple $(\xi_1,\ldots,\xi_{r-1},\xi_{r+1},\ldots,\xi_n)$ for $\xi \in \overline{X}$. $\tilde{\omega}_r$ will denote the $(n-1)$-difference form

$$(-1)^{r-1} \cdot \frac{d_q t_1}{t_1} \wedge \cdots \wedge \frac{d_q t_{r-1}}{t_{r-1}} \wedge \frac{d_q t_{r+1}}{t_{r+1}} \wedge \cdots \wedge \frac{d_q t_n}{t_n}.$$

Now we start from evaluating J_1. $t_1 = \eta_1 q^{-\nu_1}, \nu_1 \in \mathbf{Z}$, being fixed, by an induction hypothesis we have

(5.10)
$$\int_{[0,\xi^{(1)}\eta_1/\xi_1\infty]_q} \Phi(t)\tilde{\omega}_1 == \sum_{\langle\eta^{(1)}\rangle} \left([0,\xi^{(1)}\eta_1/\xi_1\infty]_q : \mathrm{reg}\ \langle\eta^{(1)}\rangle \right)_\Phi \int_{\mathrm{reg}\ \langle\eta^{(1)}\rangle} \Phi\tilde{\omega}_1,$$

where $\langle\eta^{(1)}\rangle$ ranges over the set of all α-unstable cycles for the function Φ restricted to the variables $t_2,\ldots,t_n$. As for J_r for $r \geq 2$, $t_r = \eta_r q^{-\nu_r}$ being fixed, Lemma 1 shows similarly that

$$\begin{aligned} &\int_{[0,\xi^{(r)}\eta_r/\xi_r\infty]_q} \Phi\tilde{\omega}_r \\ &\quad = U_{\sigma_{r-1}\cdots\sigma_1}(\xi)^{-1} \int_{[0,\xi^{(r)}\eta_r/\xi_r\infty]_q} \sigma_{r-1}\cdots\sigma_1\Phi(t|\alpha,\gamma)\tilde{\omega}_r \\ &\quad = U_{\sigma_{r-1}\cdots\sigma_1}(\xi)^{-1} \sum_{\langle\eta^{(r)}\rangle} ([0,\xi^{(r)}\eta_r/\xi_r\infty]_q : \mathrm{reg}\ \langle\eta^{(r)}\rangle)_{\sigma_{r-1}\cdots\sigma_1\Phi} \\ &\qquad \int_{\mathrm{reg}\ \langle\eta^{(r)}\rangle} \sigma_{r-1}\cdots\sigma_1\Phi\cdot\tilde{\omega}_r, \end{aligned} \tag{5.11}$$

where $\langle\eta^{(r)}\rangle$ ranges over the set of all α-unstable cycles for Φ restricted to the variables $t_1,\ldots,t_{r-1},t_{r+1},\ldots,t_n$. Remark that we have used the fact that $\sigma_{r-1}\cdots\sigma_1\Phi(t)$ can be expressed as

$$\begin{aligned} &\sigma_{r-1}\cdots\sigma_1\Phi(t|\alpha,\gamma) \\ &\quad = \left\{ \prod_{j=1}^{r-1} t_j^{\alpha_j+\gamma'_{j,r}-\gamma_{j,r}} \right\} t_r^{\alpha_r+\sum_{s=1}^{r-1}(\gamma_{s,r}-\gamma'_{s,r})} \\ &\qquad \cdot \left\{ \prod_{j=r+1}^{n} t_j^{\alpha_j} \right\} \prod_{j\neq r} \frac{(q^{\gamma'_{r,j}}t_j/t_r)_\infty}{(q^{\gamma_{r,j}}t_j/t_r)_\infty} \prod_{\substack{1\leq i<j\leq n \\ j\neq r, i\neq r}} \frac{(q^{\gamma'_{i,j}}t_j/t_i)_\infty}{(q^{\gamma_{i,j}}t_j/t_i)_\infty}, \end{aligned} \tag{5.12}$$

and $\theta(q^{\gamma_{r,j}}u) = \theta(q^{\gamma'_{j,r}}u^{-1})$ since $\gamma_{r,j}+\gamma'_{j,r} = 1$ for $j \neq r$. Equations (5.8), (5.10), and (5.11) imply (4.2). Assume now that $\langle\eta\rangle$ is standard. Then $\langle\eta^{(r)}\rangle$ is also standard for the function $\sigma_{r-1}\cdots\sigma_1\Phi(t)$. Hence by induction hypothesis we have

$$
\begin{aligned}
&([0,\xi^{(r)}\eta_r/\xi_r\infty]_q : \operatorname{reg}\langle\eta^{(r)}\rangle)_{\sigma_{r-1}\cdots\sigma_1\Phi} \\
&\quad = \sum_{\sigma\in W_1} \sigma_{r-1}\cdots\sigma_1\left\{\sigma\Psi_{n-1}(\xi^{(r)},\eta^{(r)}|\alpha)U_\sigma(\xi^{(r)})^{-1}U_\sigma(\eta^{(r)})\right\}.
\end{aligned}
\tag{5.13}
$$

Hence the coefficient of the integral of Φ over the cycle $\operatorname{reg}\langle\eta\rangle$ contributed in J_r equals

$$
\begin{aligned}
&(1-q)^n(q)_\infty^{3n}\left(\frac{\xi_r}{\eta_r}\right)^{\alpha_1+\cdots+\alpha_n}\frac{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\xi_r/\eta_r\right)}{\theta\left(q^{\alpha_1+\cdots+\alpha_n+1}\right)\theta(q\xi_r/\eta_r)} \\
&U_{\sigma_{r-1}\cdots\sigma_1}(\xi)^{-1}\sum_{\sigma\in W_1}\sigma_{r-1}\cdots\sigma_1\left\{\sigma\Psi_{n-1}(\xi^{(r)}\eta_r/\xi_r,\eta^{(r)}|\alpha)\right. \\
&\qquad\left.\cdot U_\sigma(\xi^{(r)}\eta_r/\xi_r)^{-1}U_\sigma(\eta^{(r)})\right\}U_{\sigma_{r-1}\cdots\sigma_1}(\eta) \\
&= (1-q)^n(q)_\infty^{3n}\sum_{\sigma\in W_r}\sigma\Psi_n(\xi,\eta|\alpha)U_\sigma(\xi)^{-1}U_\sigma(\eta),
\end{aligned}
\tag{5.14}
$$

because of (4.1) and the cocycle condition for $U_\sigma(t)$:

$$
\sigma_{r-1}\cdots\sigma_1 U_\sigma(t)\cdot U_{\sigma_{r-1}\cdots\sigma_1}(t) = U_{\sigma_{r-1}\cdots\sigma_1\sigma}(t). \tag{5.15}
$$

Hence (5.8) and (5.14) imply (4.3). The theorem has thus been proved. □

6. Examples.

(1) $n = 1$. The formula (5.2) has already been given in [M1] and [A4].

(2) $n = 2$. There are four kinds of admissible graphs:

(i) $\left\langle x_r q^{-\beta_r} * \overset{1}{\frown} 2 \right\rangle$, $\quad 1 \le r \le m$ (standard),

(ii) $\left\langle x_r q^{-\beta_r} * \overset{2}{\frown} 1 \right\rangle$, $\quad 1 \le r \le m$,

(iii) $\left\langle x_r q^{-\beta_r} * \begin{matrix} \nearrow 1 \\ \searrow 2 \end{matrix} \right\rangle$, $\quad 1 \le r \le m$ (standard),

(iv) $\left\langle \begin{matrix} x_r q^{-\beta_r} * & 1 \\ x_s q^{-\beta_s} * & 2 \end{matrix} \right\rangle$, $\quad 1 \le r, s \le m$ (standard).

We put $\gamma = \gamma_{1,2}$ and $\gamma' = \gamma'_{1,2}$, respectively. The corresponding α-stable (or α-unstable) cycles $\langle\eta\rangle$ are given as follows:

(i) $\eta_1 = qx_r, \eta_2 = q^{1-\gamma'}\eta_1$ (or $\eta_1 = x_r q^{-\beta_r}, \eta_2 = q^{-\gamma}\eta_1$);
(ii) $\eta_2 = q^{-\gamma}\eta_1, \eta_2 = qx_r$ (or $\eta_2 = q^{1-\gamma'}\eta_1, \eta_2 = x_r q^{-\beta_r}$);
(iii) $\eta_1 = \eta_2 = qx_r$ (or $\eta_1 = \eta_2 = x_r q^{-\beta_r}$);
(iv) $\eta_1 = qx_r, \eta_2 = qx_s$ (or $\eta_1 = q^{-\beta_r}x_r, \eta_2 = q^{-\beta_s}x_s$).

Moreover, we have

$$
\begin{aligned}
&\tilde{\Psi}_2(\xi,\eta|\alpha,\gamma) \\
&= (1-q)^2(q)_\infty^6 \left(\frac{\xi_1}{\eta_1}\right)^{\alpha_1}\left(\frac{\xi_2}{\eta_2}\right)^{\alpha_2} \\
&\cdot\left\{\frac{\theta\left(q^{\alpha_1+\alpha_2+1}\xi_1/\eta_1\right)\theta\left(q^{\alpha_2+1}\xi_2\eta_1/(\xi_1\eta_2)\right)}{\theta\left(q^{\alpha_1+\alpha_2+1}\right)\theta\left(q^{\alpha_2+1}\right)\theta\left(q\frac{\xi_1}{\eta_1}\right)\theta\left(q\frac{\xi_2\eta_1}{\xi_1\eta_2}\right)}\right. \\
&\qquad + \frac{\theta\left(q^{\alpha_1+\alpha_2+1}\xi_2/\eta_2\right)\theta\left(q^{\alpha_1+\gamma'-\gamma+1}\xi_1\eta_2/(\xi_2\eta_1)\right)}{\theta\left(q^{\alpha_1+\alpha_2+1}\right)\theta(q^{\alpha_1+\gamma'-\gamma+1})\theta\left(q\frac{\xi_2}{\eta_2}\right)\theta\left(q\frac{\xi_1\eta_2}{\xi_2\eta_1}\right)} \qquad (6.1) \\
&\qquad \left.\cdot\frac{\theta\left(q^{\gamma'}\xi_2/\xi_1\right)\theta\left(q^{\gamma}\eta_2/\eta_1\right)}{\theta\left(q^{\gamma}\xi_2/\xi_1\right)\theta\left(q^{\gamma'}\eta_2/\eta_1\right)}\right\} \\
&= (1-q)^2(q)_\infty^6 \left(\frac{\xi_1}{\eta_1}\right)^{\alpha_1}\left(\frac{\xi_2}{\eta_2}\right)^{\alpha_2}\left\{\frac{\theta\left(q^{1+\alpha_1}\xi_1/\eta_1\right)\theta\left(q^{1+\alpha_2}\xi_2/\eta_2\right)}{\theta\left(q^{1+\alpha_1}\right)\theta\left(q^{1+\alpha_2}\right)\theta\left(q\frac{\xi_1}{\eta_1}\right)\theta\left(q\frac{\xi_2}{\eta_2}\right)}\right. \\
&\qquad \left.-q^{\gamma}\frac{\eta_2}{\eta_1}\frac{\theta(q^{\alpha_1+\alpha_2+1}\xi_2/\eta_2)\theta(q^{\gamma'-\gamma})\theta(q^{-\alpha_1+\gamma}\xi_2/\xi_1)\theta(q^{-\alpha_1-\gamma'}\eta_1/\eta_2)}{\theta\left(q^{\alpha_1+\alpha_2+1}\right)\theta\left(q\frac{\xi_2}{\eta_2}\right)\theta\left(q^{\alpha_1+\gamma'-\gamma+1}\right)\theta(q^{\alpha_1+1})\theta\left(q^{\gamma}\frac{\xi_2}{\xi_1}\right)\theta\left(q^{\gamma'}\frac{\eta_2}{\eta_1}\right)}\right\}
\end{aligned}
$$

by the use of the three-term equation for theta functions. Indeed the three-term equation gives the following equality (see [H1]).

LEMMA 7. *For an arbitrary* $\lambda \in \mathbf{C}$,

$$
\begin{aligned}
&\frac{\theta(q^{\lambda+1})\theta(q^{\lambda+1+\gamma'-\gamma}\xi_1\eta_2/(\xi_2\eta_1))\theta(q^{\gamma}\eta_2/\eta_1)\theta(q^{\gamma'}\xi_2/\xi_1)}{\theta(q^{\lambda+1}\xi_1\eta_2/(\xi_2\eta_1))\theta(q^{\lambda+\gamma'-\gamma+1})\theta(q^{\gamma}\xi_2/\xi_1)\theta(q^{\gamma'}\eta_2/\eta_1)} - 1 \\
&\qquad (6.2) \\
&= q^{\gamma}\frac{\xi_2}{\xi_1}\frac{\theta(q^{\gamma'-\gamma})\theta(\xi_1\eta_2/(\xi_2\eta_1))\theta(q^{-\lambda+\gamma}\xi_2/\xi_1)\theta(q^{-\lambda-\gamma'}\eta_1/\eta_2)}{\theta(q^{\lambda+1}\xi_1\eta_2/(\xi_2\eta_1))\theta(q^{\lambda+\gamma'-\gamma+1})\theta(q^{\gamma}\xi_2/\xi_1)\theta(q^{\gamma'}\eta_2/\eta_1)}.
\end{aligned}
$$

Hence

(i) $([0,\xi\infty]_q : \operatorname{reg}\langle\eta\rangle)_\Phi$

$$
\begin{aligned}
&= (1-q)^2\left(\frac{\xi_1}{\eta_1}\right)^{\alpha_1}\left(\frac{\xi_2}{\eta_2}\right)^{\alpha_2}\frac{(q)_\infty^6}{\theta(q^{\alpha_1+\alpha_2+1})\theta(q^{\alpha_2+1})} \\
&\cdot\frac{\theta(q^{\alpha_1+\alpha_2+1}\xi_1/\eta_1)\theta(q^{\alpha_2+\gamma+1}\xi_2/\xi_1)}{\theta(q\xi_1/\eta_1)\theta(q^{1+\gamma}\xi_2/\xi_1)}
\end{aligned}
$$

since $\eta_2 = q^{-\gamma}\eta_1$.

(ii) $([0,\xi\infty]_q : \operatorname{reg}\langle\eta\rangle)_\Phi$

$$= (1-q)^2 \left(\frac{\xi_1}{\eta_1}\right)^{\alpha_1} \left(\frac{\xi_2}{\eta_2}\right)^{\alpha_2}$$

$$\cdot \frac{(q)_\infty^3 \theta(q^{\alpha_1+\alpha_2+1}\xi_2/\eta_2)\theta(q^{\alpha_1-\gamma+2}\xi_1/\xi_2)\theta(q^{\gamma'}\xi_2/\xi_1)\theta(q^{1+\gamma-\gamma'})}{\theta(q^{\alpha_1+\alpha_2+1})\theta(q^{\alpha_1+\gamma'-\gamma+1})\theta(q\xi_2/\eta_2)\theta(q^{\gamma}\xi_2/\xi_1)\theta(q^{2-\gamma'}\xi_1/\xi_2)},$$

by taking the residue of $\tilde{\Psi}(\xi,\eta|\alpha,\gamma)$ for $\eta_2 = q^{1-\gamma'}\eta_1$.

(iii) and (iv) $([0,\xi\infty]_q : \langle\eta\rangle)_\Phi = \tilde{\Psi}_2(\xi,\eta|\alpha,\gamma)$,

which is well defined in view of (6.1).

It is also interesting to consider the case where $q = 1$. A similar formula to our theorem may be possible in view of the result obtained in [G3].

REFERENCES

[A1] G. E. ANDREWS, *q-series: Their development and application in analysis, number theory, combinatorics, physics, and computer algebra*, CBMS Regional Conference Series in Math., 66, American Mathematical Society, Providence, RI, 1986.

[A2] K. AOMOTO, *Finiteness of a cohomology associated with certain Jackson integrals*, Tôhoku Math. J., 43 (1991), pp. 75–101.

[A3] K. AOMOTO AND Y. KATO, *A q-analogue of de Rham cohomology associated with Jackson integrals*, in Special Functions, M. Kashiwara and T. Miwa, eds., Proc. of the Hayashibara Forum, Springer-Verlag, New York, 1990.

[A4] K. AOMOTO, Y. KATO, AND K. MIMACHI, *A solution of Yang–Baxter equation as connection coefficients of a holonomic q-difference system*, International Math. Research Notices, No. 1, Duke. Math. J., 1992, pp. 7–15.

[A5] ———, *Gauss matric decomposition and a solution of Yang–Baxter equation*, preprint, 1991; J. Math. Anal., to appear.

[A6] R. ASKEY, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938–951.

[A7] ———, *Beta integrals in Ramanujan's papers, his unpublished work and further examples, Ramanujan revisited*, G. E. Andrews, ed., Proc. of the Centenary Conference, 1987, pp. 561–590.

[A8] ———, *Beta integrals and q-extensions*, Proc. of the Ramanujan Centennial International Conference, Annamalainagar, 1987, pp. 85–102.

[F] I. B. FRENKEL AND N. YU. RESHETIKHIN, *Quantum affine algebras and holonomic difference equations*, in Differential Geometric Methods in Theoretical Physics, S. Catto and A. Rocha, eds., World Scientific, New York, 1992.

[G1] G. GASPER AND M. RAHMAN, *Basic hypergeometric series*, in Encyclopedia of Mathematics and its Applications, Cambridge Univ. Press, London, 1990.

[G2] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, John Wiley & Sons, New York, 1978.

[G3] R. A. GUSTAFSON, *A generalization of Selberg integral*, Bull. Amer. Math. Soc., 22 (1990), pp. 97–105.

[H1] H. HANCOCK, *Lectures on the Theory of Elliptic Functions*, Dover, New York, 1958.

[H2] F. HARARY, *Graph Theory*, Addison Wesley, Reading, MA, 1969.

[K1] J. KANEKO, *Selberg integrals and hypergeometric functions*, in Special Differential Equations, M. Yosida and M. Namba, eds., Proc. of the Taniguchi Workshop, 1991.

[K2] K. KADELL, *A proof of Askey's conjectured q-analogue of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal., 19 (1988), pp. 969–986.

[M1] K. MIMACHI, *A proof of Ramanujan's identity by use of loop integrals*, SIAM J. Math. Anal., 19 (1988), pp. 1490–1493.

[M2] K. MIMACHI, *Connection problem in holonomic q-difference system associated with a Jackson integral of Jordan–Pochhammer type*, Nagoya Math. J., 116 (1989), pp. 149–161.

[M3] J. W. MOON, *Various proofs of Cayley's formula for counting trees*, in A Seminar on Graph Theory, F. Harary, ed., Holt, Rinehart and Winston, New York, 1967, pp. 70–78.

[O] T. ODA, *Convex bodies and algebraic geometry—an introduction to the theory of toric varieties*, Ergebnisse der Math., Springer-Verlag, New York, 1988.

[S] C. SABBAH, *Systèmes holonomes d'équations aux q-différences*, preprint, 1991.

[V] A. VARCHENKO, *Quantized Knizhnik–Zamolodchikov equations, Quantum Yang–Baxter equation, and difference equations for q-hypergeometric functions*, preprint, 1993.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 274–287, March 1994

003

GROUP THEORETICAL INTERPRETATIONS OF SPECIAL FUNCTION IDENTITIES: TWO EXAMPLES*

L. C. BIEDENHARN† AND A. K. ÇIFTÇI‡

Abstract. Two examples, taken from quantum physics, are used to illustrate how group theoretical concepts afford an intuitive understanding of relationships between certain special function identities.

Key words. special functions, symmetry groups, quantal angular momentum group $SU(2)$, group representations, Kronecker product (Wigner product law), group contraction

AMS subject classifications. 33C10, 33C45, 33C55

1. Introduction. The applications of symmetry techniques using group theory have been an important source of special functions [1], [2], and it is only to be expected that quantum physics—with its current strong emphasis on symmetry—has similarly contributed greatly to the supply of special functions [3]. Results in special function theory which stem from group theory are usually rather particular and often lack the generality of analytic results. This situation is characteristic of the group-theoretic approach to special functions [1]. Despite this particularity, results of the group theoretic approach nevertheless have a certain coherence, clarity, and intuitiveness that the analytic approach lacks.

In this contribution we will discuss two examples of the way in which group theoretic concepts can lead to a more transparent and intuitive understanding of certain special function identities and their interrelationships. Both examples are taken from quantum physics and both involve special functions arising from applications of angular momentum techniques. Our first example—the relationship of Gaunt's integral [4] to Sharp's integral [5]—is not new and is possibly well known. Gaunt's integral (see (2.10a) below) involves special functions of the quantal angular momentum group $SU(2)$, whereas Sharp's integral (see (2.19)) involves special functions of the Euclidean group $E(2)$ of the plane. The latter group is a *contraction* of the former (see §3), and it is this asymptotic relationship that we will emphasize and discuss as interrelating the two integrals.

The second example, which we believe is new, concerns integrals involving special functions of the four-dimensional rotation group $SO(4)$ (see (4.10)) and of its contracted group, the Euclidean group in three-space $E(3)$. This latter group is of evident importance in nonrelativistic quantum physics and, in fact, the results we will discuss arose from a recent investigation [6], [7] of the muon-catalyzed fusion process [8], [9].

The interrelationship between the special functions of a symmetry group and those of its contracted (asymptotic) groups is a general feature and can be applied systematically. In our concluding remarks we conjecture the extension of this procedure to quantum groups [10], [11] and their q-analog special functions.

2. First example: Gaunt's integral and Sharp's integral. Angular momentum in quantum physics involves the symmetry group $SU(2)$, whose generators are the three angular momentum operators $\{J_1, J_2, J_3\}$ obeying the commutation rules:

* Received by the editors May 4, 1992; accepted for publication February 8, 1993.
† Present address, Physics Department, The University of Texas at Austin, Austin, Texas 78712.
‡ Fizik Bölümü, Ankara University Fen Fakültesi, 06100 Tandogan, Ankara, Turkey.

$$[J_i, J_j] = i\varepsilon_{ijk}J_k, \tag{2.1}$$

where $\varepsilon_{ijk} = \pm 1$ for (positive/negative) permutations of 1, 2, 3, and is zero otherwise.

In the standard way [12] one constructs eigenbases (quantal states) for the representations—ket vectors denoted $|jm\rangle$—having sharp total angular momentum $j(2j \in \mathbb{Z}^+)$ and z-component $J_z \to m(-j \leq m \leq j, 2m \in \mathbb{Z}, j-m \in \mathbb{Z})$. The irreducible representations (irreps) are denoted by $\mathcal{D}^{(j)}(g), g \in SU(2)$, with matrix elements: $\langle jm'|\mathcal{D}^{(j)}(g)|jm\rangle \equiv \mathcal{D}^{(j)}_{m',m}(\alpha,\beta,\gamma)$, where α, β, and γ are the Euler angles of the rotation g. The rotation operator matrix elements are special functions, the Jacobi polynomials [12].

The group-theoretical approach to special functions has several characteristic general features [1], [2], [3]. For compact groups, the Peter–Weyl theorem shows that matrix elements of the irreducible representation matrices supply a *complete set of orthonormal special functions over the group manifold.* Similarly, it is characteristic of the group theoretic approach that the *product law* for the group $g_1 \circ g_2 = g_{12}$, when applied to the group representations $\mathcal{D}(g_1)\mathcal{D}(g_2) = \mathcal{D}(g_{12})$, yields an *addition theorem* for the special functions of the group.

A closely related further characteristic of the group theoretic approach—in a certain sense dual to the product law—is the *Kronecker product* [3] for matrix elements of representations involving the *same* group element. This general structure is at present well defined only for multiplicity-free groups [13] or for groups having a canonical resolution of the multiplicity [14]. For angular momentum theory this Kronecker product relation is called the *Wigner product law* [12]:

$$D^{J_1}_{M_1,M_1'}(g)D^{J_2}_{M_2,M_2'}(g) = \sum_{J_3,M_3,M_3'} C^{J_1J_2J_3}_{M_1M_2M_3}C^{J_1J_2J_3}_{M_1'M_2'M_3'}D^{J_3}_{M_3M_3'}(g). \tag{2.2}$$

In this result, the terms $C^{J_1J_2J_3}_{M_1M_2M_3}$ are the *Wigner–Clebsch–Gordan* coefficients [15], which, as group objects, affect the reduction of the direct (Kronecker) product of the irreps J_1 and J_2 into the irreps J_3. Considered as special functions, the Wigner–Clebsch–Gordan coefficients are ${}_3F_2$ generalized hypergeometric functions with the specific relation [12]:

$$C^{abc}_{\alpha\beta\gamma} = \delta_{\alpha+\beta,\tau}\left[(2c+1)(a+\alpha)!(a-\alpha)!(b+\beta)!(b-\beta)!(c+\gamma)!(c-\gamma)!\right]^{\frac{1}{2}}$$

$$\times \frac{(-1)^{a+b+\gamma+\delta_1}\Delta(abc)\,{}_3F_2\left(\begin{matrix}\varepsilon_1-\delta_1, \varepsilon_2-\delta_1, \varepsilon_3-\delta_1 \\ \delta_2-\delta_1+1, \delta_3-\delta_1+1\end{matrix}; 1\right)}{(\delta_2-\delta_1)!(\delta_3-\delta_1)!(\delta_1-\varepsilon_1)!(\delta_1-\varepsilon_2)!(\delta_1-\varepsilon_3)!}. \tag{2.3}$$

Here the parameters $(\delta_1, \delta_2, \delta_3)$ are any permutation of

$$(\alpha+b+\beta, b-\beta+c+\gamma, a+\alpha+c+\gamma), \tag{2.4}$$

except that δ_1 is required to be the smallest integer in this set; the parameters $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are any permutation of

$$(a+\alpha, b+\alpha+\tau, c+\tau). \tag{2.5}$$

The *triangle coefficient* $\Delta(abc)$ (which is a special function in its own right in angular momentum theory [12]) is given by

$$\Delta(abc) = \left[\frac{(a+b-c)!(a-b+c)!(-a+b+c)!}{(a+b+c+1)!}\right]^{\frac{1}{2}}. \tag{2.6}$$

By using the orthonormality relations, we can invert (2.2) and obtain a general integral relation for a triple product of $SU(2)$ special functions:

$$\int dg \left(D^{J_3}_{M_3 M_3'}(g)\right)^* D^{J_1}_{M_1 M_1'}(g) D^{J_2}_{M_2 M_2'}(g) = \left(\frac{8\pi_2}{2J_3+1}\right) C^{J_1 J_2 J_3}_{M_1 M_2 M_3} C^{J_1 J_2 J_3}_{M_1' M_2' M_3'}. \tag{2.7}$$

This result is a special function identity relating an integral over a product of three Jacobi functions to a product of two ${}_3F_2$ special functions.

Gaunt's integral [4] was obtained in the early days of quantum mechanics and involves a special case of this general result, (2.7). For integral angular momenta, denoted by l (orbital angular momenta), the representation functions are related to the spherical harmonics $Y_{lm}(\theta\varphi)$ by

$$Y_{lm}(\theta\varphi) \equiv \langle \theta\varphi | lm \rangle = \left[\frac{2l+1}{4\pi}\right]^{\frac{1}{2}} D^{l*}_{m,0}(\varphi\theta 0). \tag{2.8}$$

Specializing further to $m = 0$ relates the representation functions to the Legendre polynomials

$$D^{l}_{0,0}(0\theta 0) = P_l(\cos\theta). \tag{2.9}$$

Gaunt's integral is an integral over three Legendre polynomials:

$$I_{\text{Gaunt}} = \int_0^\pi \sin\theta d\theta P_{l_1}(\cos\theta) P_{l_2}(\cos\theta) P_{l_3}(\cos\theta), \tag{2.10a}$$

which, using (2.7) and (2.9), can be expressed as

$$I_{\text{Gaunt}} = \left(\frac{2}{2l_3+1}\right) \left(C^{l_1 l_2 l_3}_{000}\right)^2. \tag{2.10b}$$

A more general result is easily obtained from (2.7), using (2.8) rather than (2.9). This is the *generalized Gaunt integral*:

$$\begin{aligned} &\int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\varphi Y^*_{l_3 m_3}(\theta\varphi) Y_{l_1 m_1}(\theta\varphi) Y_{l_2 m_2}(\theta\varphi) \\ &= \left[\frac{(2l_1+1)(2l_2+1)}{4\pi(2l_3+1)}\right]^{\frac{1}{2}} C^{l_1 l_2 l_3}_{000} \cdot C^{l_1 l_2 l_3}_{M_1 M_2 M_3}. \end{aligned} \tag{2.11}$$

Now let us turn to Sharp's integral. This integral involves representation functions of the Euclidean group in two dimensions $E(2)$. This (noncompact) group has three generators: two translation operators $\{T_1, T_2\}$ and a rotation operator M acting in the (1, 2) plane. The commutation relations for these three generators are

$$[T_1, T_2] = 0, \quad [M, T_1] = T_2, \quad [M, T_2] = -T_1. \tag{2.12}$$

The group theoretic results are analogous to the previous case: one constructs the representations (which include as special functions the Bessel functions of integer

order), and uses the $E2$ analog of the Kronecker product law to construct a general threefold Bessel function integral.

The representations $\mathcal{D}(g)$ in which we are interested are labeled by one real parameter p, physically the magnitude of the (nonvanishing) momentum $(0 < p < \infty)$. The irreps $\mathcal{D}^{(p)}(g)$ have the matrix elements [5]

$$\langle m'|\mathcal{D}^{(p)}(g)|m\rangle = e^{i(m-m')\theta} J_{m-m'}(pr) e^{im'\alpha}, \qquad m, m' \in \mathbb{Z}, \tag{2.13}$$

where the basis ket-vectors $|m\rangle$ are eigenfunctions of the rotation operator $M (M \to m, m \in \mathbb{Z})$. In (2.13), the group element g is a translation, parametrized by (r, θ) using polar coordinates, and a rotation by the angle α.

The construction of the Kronecker product law for the $E(2)$ group involves a typical difficulty that occurs for many groups: the existence of multiplicity in the reduction of the Kronecker product [5], [13]. The analog of the Wigner–Clebsch–Gordan coefficients for a nonmultiplicity-free group is not well defined unless there exists a canonical resolution of the multiplicity [14]. For the Euclidean group $E(2)$ the multiplicity is at most 2, and a canonical resolution exists [5]. This resolution adjoins a discrete operation, reflection in the x-axis, yielding the extended Euclidean group $E(2)'$. Denoting the involutary operation by I, we have the additional commutation relations:

$$IM = -MI, \quad IT_1 = T_1 I, \quad IT_2 = -T_2 I. \tag{2.14}$$

The irreducible representations (irreps) are now specified by two labels: p (as before) and $\varepsilon = \pm 1$. The irreps $\mathcal{D}^{(p,\varepsilon)}(g)$ have the same matrix elements as in (2.13) for group elements in $E(2)$, and the additional relation

$$\langle m'|\mathcal{D}^{(p,\varepsilon)}(I)|m\rangle = (-1)^{m+\varepsilon}\delta_{m'}^{-m} \tag{2.15}$$

for the involution I. The extra irrep label ε distinguishes between the two occurrences of the $E(2)$ irreps labelled by p in the reduction of the Kronecker product.

The analog of the Wigner–Clebsch–Gordan (WCG) coefficients for the extended Euclidean group $E(2)'$ have the form [5]

(2.16)

$$C_{m_1 m_2 m_3}^{p_1\varepsilon_1 p_2\varepsilon_2 p_3\varepsilon_3} = \frac{\delta_{m_1+m_2+m_3}}{\sqrt{8\pi A(p_1 p_2 p_3)}} \left(e^{i(m_2\tau_3 - m_3\tau_2)} + (-1)^{\varepsilon_1+\varepsilon_2+\varepsilon_3} e^{-i(m_2\tau_3 - m_3\tau_2)} \right),$$

where $A(p_1 p_2 p_3)$ is the area of the triangle formed by the three momenta $p_i > 0$ and the τ_i are the *exterior* angles of this triangle,

$$\tau_1 + \tau_2 + \tau_3 \equiv 0 \bmod 2\pi. \tag{2.17}$$

If the three momenta do not form a triangle, the coefficient above vanishes, or equivalently, we may define the area of the triangle as infinite.

The general integral [5] for the product of three irreps of the $E(2)'$ group—specialized to have nonzero momenta—is the $E(2)'$ analog of (2.7) and has the form

$$\int dg D_{m_3,m_3'}^{(p_3\varepsilon_3)*}(g)\, D_{m_1,m_1'}^{(p_1\varepsilon_1)}(g)\, D_{m_2,m_2'}^{(p_2\varepsilon_2)}(g) = C_{m_1 m_2 m_3}^{p_1\varepsilon_1 p_2\varepsilon_2 p_3\varepsilon_3}\, C_{m_1' m_2' m_3'}^{p_1\varepsilon_1 p_2\varepsilon_2 p_3\varepsilon_3}. \tag{2.18}$$

Sharp's integral is that special case of (2.18) for which $m_i' = 0$. For this special case, the WCG coefficient ((2.16) with $m_1' = 0$) vanishes unless $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 \equiv 0 \mod 2$.

Using (2.13) and (2.16) in the general relation equation (2.18) and noting the restriction $m_1 = 0$, we obtain *Sharp's integral*:

(2.19)

$$\int_0^\infty r\,dr J_{m_1}(p_1 r) J_{m_2}(p_2 r) J_{m_3}(p_3 r) = \frac{\cos(m_1\tau_2 - m_2\tau_1)}{2\pi A(p_1 p_2 p_3)} \quad \text{if } m_1 + m_2 + m_3 = 0.$$

Sharp pointed out [5] that the integral (2.19) does not appear in the standard compilations by Watson [16] or by Bateman/Erdélyi [17].

3. The contraction relationship. The two integrals above, the generalized Gaunt integral in (2.11) and Sharp's integral in (2.19), have each been obtained by a straightforward application of standard group techniques, and are in consequence similar results differing only in the specific group used. Our objective, however, is not simply to see that the integrals are *analogous* but to show—from group-theoretic concepts—how the particular structures can be understood in such a way that one integral can be *directly* derived from the other.

To do this, let us recall that in physics it is not infrequent that a given theory is subsumed in a larger theory such that the earlier theory may be recovered as a limit. Thus, for example, nonrelativistic physics (Newtonian relativity) is contained in Einsteinian relativity (the physics of special relativity) such that Newtonian relativity is recovered in the limit that the velocity of light (c) becomes infinite. Expressed in group-theoretic terms [18], [19], the Poincaré group of Einsteinian relativistic symmetry *contracts* to the Galilean group of Newtonian relativistic symmetry in the limit where the parameter $c \to \infty$.

A similar relationship exists for the group $SU(2)$, which contracts to $E(2)$. This limit is intuitively understandable as the limit in which the two-sphere S^2 (the space of spherical harmonics) has a "large" radius so that, locally at a specific point, the neighborhood looks flat, becoming the Euclidean two-plane in the limit.

To be more precise let us consider the generators of the $SU(2)$ group: J_1, J_2, J_3, and the commutation rules (2.1):

$$[J_1, J_2] = iJ_3, \quad [J_2, J_3] = iJ_1, \quad \text{and} \quad [J_3, J_1] = iJ_2. \tag{3.1}$$

Multiply J_1 and J_2 in (3.1) by ε so as to obtain

$$[\varepsilon J_1, \varepsilon J_2] = i\varepsilon^2 J_3, \quad [\varepsilon J_2, J_3] = i\varepsilon J_1, \quad \text{and} \quad [J_3, \varepsilon J_1] = i\varepsilon J_2. \tag{3.2}$$

Let J_1 and J_2 become large, and let ε approach zero, such that εJ_i is finite:

$$\varepsilon J_1 \to T_1, \qquad \varepsilon J_2 \to T_2, \tag{3.3}$$

and redefine T_3 to be M. We find that in this contraction limit, the commutation relations become

$$[T_1, T_2] = 0, \quad [M, T_1] = iT_2, \quad [M, T_2] = -iT_1, \tag{3.4}$$

which are precisely the commutation relations of $E(2)$, (2.12).

It is important to note that the contraction process is not necessarily smooth, and the contraction limit can be, and often is, singular. As a result, the importance of the contraction process—considered as a systematic procedure—lies primarily in the *qualitative* insights it affords into the existence, and properties, of asymptotic relationships between special functions.

Let us now apply these ideas to the $SU(2) \to E(2)$ contraction limit. Applied to the representation matrices, the contraction limit is the relation [5]

$$\lim_{j\to\infty} \left(\langle m' | \mathcal{D}^{(j)}(\theta - \tfrac{\pi}{2}, \varepsilon r, \tfrac{\pi}{2} - \theta + \alpha) | m \rangle \right) = \langle m' | \mathcal{D}^{(p)}(r, \theta, \alpha | m \rangle, \tag{3.5}$$

where in the limit j runs over integral values, with $j\varepsilon = p$ (the momentum) fixed. This relation is the well-known asymptotic relation of the Bessel function as a limit of Jacobi polynomials. Accordingly, we can conclude that *the contraction limit of the integrand of the generalized Gaunt integral over three spherical harmonics yields the integrand of Sharp's integral over three Bessel functions*, noting that the orthonormality of both sets of representation matrices determines in each case a well-defined absolute normalization (which is group dependent).

The contraction limit of the two sets of WCG coefficients is a somewhat less straightforward matter because of the multiplicity problem for $E(2)$, which requires the use of the extended Euclidean group $E(2)'$. It is not difficult to adjoin the involution I to the group $SU(2)$. (Note that the group volume is doubled in going from $SU(2)$ to $SU(2)'$, affecting the normalization of the representation functions.) As expected, $E(2)'$ is the contraction limit of $SU(2)'$, but the problem is that this limit is singular. In particular, for $SU(2)'$, there exists a *parity rule*[1] in $SU(2)'$ such that $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = 0 \bmod 2$. This parity rule is *broken* for $E(2)'$ in the (singular) contraction limit.

In the specific case at hand (the generalized Gaunt integral and Sharp's integral) this problem is *avoided*, since in these two cases the integrals involve representation functions with $m_1' = m_2' = m_3' = 0$, a condition that *enforces* the parity rule even for $E(2)'$. For this special case finds the following the contraction limit of the WCG coefficients:

$$\left(C^{p_1\varepsilon_1 p_2\varepsilon_2 p_3\varepsilon_3}_{m_1 m_2 m_3} \cdot C^{p_1\varepsilon_1 p_2\varepsilon_2 p_3\varepsilon_3}_{0\ 0\ 0} \right)_{E(2)'} = \underset{\text{limit}}{\text{contraction}} \left(C^{l_1 l_2 l_3}_{m_1 m_2 m_3} \cdot C^{l_1 l_2 l_3}_{000} \right)_{SU(2)'}. \tag{3.6}$$

It follows that *the contraction limit carries the generalized Gaunt integral identity into Sharp's integral identity.* This result was first shown by Sharp [5].

4. Second example: $SO(4)$ and $E(3)$. Let us now consider the rotation group in four dimensions: $SO(4)$. This group has six generators, $\vec{L}$ and $\vec{K}$, which obey the commutation relations:

$$[L_i, L_j] = i\varepsilon_{ijk} L_k, \quad [L_i, K_j] = i\varepsilon_{ijk} K_k, \quad \text{and} \quad [K_i, K_j] = i\varepsilon_{ijk} L_k. \tag{4.1a–c}$$

The irreps (for this presentation of the group) are labeled by two invariant operators which may be taken to be $\mathcal{I}_1 \equiv \vec{L}^2 + \vec{K}^2 + 1$ and $\mathcal{I}_2 \equiv \vec{L} \cdot \vec{K}$. For the orbital irreps in which we are most interested these two invariants assume the eigenvalues $\mathcal{I}_1 \to n^2, (n = 1, 2 \ldots)$ and $\mathcal{I}_2 \to 0$. The corresponding irreps are labelled $\mathcal{D}^{(n,0)}(g)$.

[1] To see this, note that the reflection I in the x-axis combines with a rotation by π around the x-axis to yield a central reflection (parity).

It is a fortunate group theoretical "accident" that under the substitution $\vec{L} = \vec{j}_1 + \vec{j}_2, \vec{K} = \vec{j}_1 - \vec{j}_2$ the commutation relations (4.1) take the form

$$\vec{j}_i \times \vec{j}_i = i\vec{j}_i (i = 1, 2) \quad \text{with } [\vec{j}_1, \vec{j}_2] = 0. \tag{4.2}$$

Thus we see that $SO(4)$ is locally the direct product, $SU(2) \times SU(2)$, of two commuting $SU(2)$ groups. This has the important consequence that $SO(4)$ is multiplicity-free and possesses a well-defined Kronecker product law. This basic result is the key to our special function considerations [20], [21].

The general irrep of $SO(4)$, using the realization given in (4.2), may be denoted by $\mathcal{D}^{[j_1,j_2]}(g)$, where the labels j_1, j_2 specify the two $SU(2)$ invariants of the two $SU(2)$ groups. (Note that we use square brackets for these labels to distinguish this choice of invariants from $\mathcal{I}_1, \mathcal{I}_2$ shown earlier.) Here g denotes a generic group element in $SU(2) \times SU(2)$. Matrix elements of the irrep $\mathcal{D}^{[j_1 j_2]}$ are denoted by

$$\langle [j_1 j_2] j' m' | \mathcal{D}^{[j_1 j_2]}(g) | [j_1 j_2] j m \rangle \equiv D^{[j_1,j_2]}_{j'm';jm}(g), \tag{4.3}$$

where the basis ket-vectors $|[j_1 j_2] j m\rangle$ are direct product ket-vectors from $SU(2) \times SU(2)$ coupled to angular momentum j with z-component m. These $SO(4)$ irrep matrices are accordingly vector-coupled $SU(2) \times SU(2)$ rotation matrices [20], [21], [22].

The Kronecker product law for this group [20], [21] has the form

$$\begin{aligned} \mathcal{D}^{[j_1 j_2]}_{j'm';jm}(g) : \mathcal{D}^{[k_1 k_2]}_{k'\kappa';k\kappa}(g) = \sum_{l_1 l_2 l' \lambda' l \lambda} & C\left[\begin{pmatrix} j_1 & j_2 \\ j' & m' \end{pmatrix}\begin{pmatrix} k_1 & k_2 \\ k' & \kappa' \end{pmatrix}\begin{pmatrix} l_1 & l_2 \\ l' & \lambda' \end{pmatrix}\right] \\ & \times C\left[\begin{pmatrix} j_1 & j_2 \\ j & m \end{pmatrix}\begin{pmatrix} k_1 & k_2 \\ k & \kappa \end{pmatrix}\begin{pmatrix} l_1 & l_2 \\ l & \lambda \end{pmatrix}\right] \mathcal{D}^{[l_1 l_2]}_{l'\lambda';l\lambda}(g). \end{aligned} \tag{4.4}$$

In this result, (4.4), we have used the Wigner–Clebsch–Gordan coefficients for the $SO(4)$ group [20], which are given explicitly by a special function that is the sum of products of five ${}_3F_2$ functions:

$$C\left[\begin{pmatrix} j_1 & j_2 \\ j & m \end{pmatrix}\begin{pmatrix} k_1 & k_2 \\ k & \mu \end{pmatrix}\begin{pmatrix} l_1 & l_2 \\ l & \nu \end{pmatrix}\right] \equiv \sum_{\substack{\nu_1 \nu_2 \\ m_1 m_2 \\ \mu_1 \mu_2}} C^{l_1 l_2 l}_{\nu_1 \nu_2 \nu} C^{j_1 j_2 j}_{m_1 m_2 m} C^{k_1 k_2 k}_{\mu_1 \mu_2 \mu} C^{j_1 k_1 l_1}_{m_1 \mu_1 \nu_1} C^{j_2 k_2 l_2}_{m_2 \mu_2 \nu_2}. \tag{4.5}$$

The sum in (4.5) can be put in the more explicit form

$$\begin{aligned} & C\left[\begin{pmatrix} j_1 & j_2 \\ j & m \end{pmatrix}\begin{pmatrix} k_1 & k_2 \\ k & \kappa \end{pmatrix}\begin{pmatrix} l_1 & l_2 \\ l & \lambda \end{pmatrix}\right] \\ & = [(2l_1+1)(2l_2+1)(2j+1)(2k+1)]^{\frac{1}{2}} \begin{Bmatrix} j_1 & j_2 & j \\ k_1 & k_2 & k \\ l_1 & l_2 & l \end{Bmatrix} C^{jkl}_{m\kappa\lambda}, \end{aligned} \tag{4.6a}$$

where the term in brackets $\{\ldots\}$ is an angular momentum special function known as the $9-j$ symbol having the definition

(4.6b)
$$\begin{Bmatrix} a & b & c \\ d & e & f \\ h & i & j \end{Bmatrix} \equiv (-1)^{c+f-j} \frac{\Delta(dah)\Delta(bei)\Delta(jhi)}{\Delta(def)\Delta(bac)\Delta(jcf)} \times \sum_{xyz} \frac{(-1)^{x+y+z}}{x!y!z!} \frac{(2f-x)!(2a-z)!}{(2i+1+y)!(a+d+h+1-z)!} \times \frac{(d+e-f+x)!(c+j-f+x)!}{(e+f-d-x)!(c+f-j-x)!} \frac{(e+i-b+y)!(h+i-j+y)!}{(b+e-i-y)!(h+j-i-y)!} \times \frac{(b+c-a+z)!}{(a+d-h-z)!(a+c-b-z)!} \times \frac{(a+d+j-i-y-z)!}{(d+i-b-f+x+y)!(b-f-a+j+x+z)!},$$

where the triangle function $\Delta(abc)$ is given in (2.6).

Using the orthonormality of the irrep matrices, just as in §2, we find an interesting general result for the product of three $SO(4)$ special functions:

(4.7)
$$\left(\frac{(2l_1+1)(2l_2+1)}{64\pi^4}\right) \cdot \int dg D^{[l_1 l_2]*}_{l'\lambda';l\lambda}(g)\, D^{[j_1 j_2]}_{j'm';jm}(g)\, D^{[k_1 k_2]}_{\kappa'\kappa';k\kappa}(g) = C\left[\begin{pmatrix} j_1 & j_2 \\ j & m \end{pmatrix}\begin{pmatrix} k_1 & k_2 \\ k & n \end{pmatrix}\begin{pmatrix} l_1 & l_2 \\ l & \lambda \end{pmatrix}\right] C\left[\begin{pmatrix} j_1 & j_2 \\ j & m \end{pmatrix}\begin{pmatrix} k_1 & k_2 \\ k & \kappa \end{pmatrix}\begin{pmatrix} l_1 & l_2 \\ l & \lambda \end{pmatrix}\right].$$

Our interest here is in a special case of (4.7) corresponding to representation matrices which are the spherical harmonics in four dimensions. To obtain this, one specializes $\mathcal{D}^{[j_1 j_2]}(g)$ to have the $SU(2) \times SU(2)$ invariants $j_1 = j_2 \equiv j$, which implies for the realization given by (4.1) that the invariant $\mathcal{I}_1 = (2j+1)^2$—so that $n = 2j+1$—and $\mathcal{I}_2 = 0$. The corresponding irrep, $D^{[j,j]*}_{lm;00}(g)$, will be related to the spherical harmonic $Y_{nlm}(\chi,\theta,\varphi)$ in four dimensions (see (4.8a) below). The group element g will be specified by the angles (χ,θ,φ), which are the polar angles for a point on the surface of the sphere S^3, with $x^2+y^2+z^2+t^2 = 1$ and $t = \cos\chi, z = \sin\chi\cos\theta, y = \sin\chi\sin\theta\sin\varphi, x = \sin\chi\sin\theta\cos\varphi$. All this is in exact analogy with the previous discussion for the spherical harmonics $Y_{lm}(\theta,\varphi)$ in three dimensions.

The spherical harmonics $Y_{nlm}(\chi,\theta,\varphi)$ factorize [20] and [22] into the product of a Gegenbauer polynomial in χ and the usual $SO(3)$ spherical harmonic

(4.8a)
$$\left(\frac{n^2}{2\pi^2}\right)^{\frac{1}{2}} \cdot \mathcal{D}^{(\frac{n-1}{2},\frac{n-1}{2})*}_{lm;00}(g) \equiv Y_{nlm}(\chi,\theta,\varphi) = F_{nl}(\chi)Y_{lm}(\theta\varphi),$$

where

(4.8b)
$$F_{nl}(\chi) = \left(\frac{2n\Gamma(n-l)}{\pi\Gamma(n+l+1)}\right)^{\frac{1}{2}} (l!)(2i\sin\chi)^l C^{l+1}_{n-l-1}(\cos\chi).$$

The function $C^{l+1}_{n-l-1}(\cos\chi)$ in (3.8b) is the special function known as the Gegenbauer polynomial defined in the standard way [17].

For completeness, let us note that the spherical harmonics $Y_{nlm}(\chi,\theta,\varphi)$ are orthogonal and normalized by

(4.9)
$$\int_0^\pi \sin^2\chi d\chi \int_0^\pi \sin\theta d\theta \int_0^\pi d\varphi Y^*_{n'l'm'}(\chi,\theta,\varphi)Y_{nlm}(\chi,\theta,\varphi) = \delta^{n'}_n \delta^{l'}_l \delta^{m'}_m.$$

Introducing these special functions into the general integral, (4.7), we obtain

$$\int dg Y^*_{n_3 l_3 m_3}(g) Y_{n_1 l_1 m_1}(g) Y_{n_2 l_2 m_2}(g) = \left(\frac{n_1 n_2}{2\pi^2 n_3}\right)^{\frac{1}{2}} C\left[\begin{pmatrix} \frac{n_1-1}{2} & \frac{n_1-1}{2} \\ l_1 & m_1 \end{pmatrix}\begin{pmatrix} \frac{n_2-1}{2} & \frac{n_2-1}{2} \\ l_2 & m_2 \end{pmatrix}\begin{pmatrix} \frac{n_3-1}{2} & \frac{m_3-1}{2} \\ l_3 & m_3 \end{pmatrix}\right], \tag{4.10}$$

where the coefficient $C[\ldots]$ is defined in (4.5) and both g and the differential dg are defined implicitly in (4.9). Equation (4.10) is the four-dimensional analog of the generalized Gaunt integral of (2.11).

Using (4.8) this result can be put in the form of an integral over three Gegenbauer polynomials, (evaluating the integral over the $SO(3)$ spherical harmonics by (2.11)), but this explicit result is not necessary for our present purpose.

Remark. The $SO(4)$ spherical harmonic integral, (4.10), implies the restriction (see (4.6a)) that the parameters in two of the three columns of the $9-j$ symbol are *identical.* The $9-j$ symbol has the general symmetry [12] that under the exchange of two columns the symbol is multiplied by the factor: $(-1)^S$, where $S \equiv$ sum of the nine parameters in the $9-j$ symbol. Thus we see that for two identical columns, we must have $(-1)^S = +1$ in order to obtain a nonzero value. This implies that integral in (4.10) is nonzero only for $l_1 + l_2 + l_3 \equiv 0 \bmod 2$. For (4.10) this restriction on the $SO(4)$ WCG coefficient (the right-hand side of (4.10)) is exactly the same restriction as implied by the $SO(3)$ spherical harmonic integral of the left-hand side of (4.10). This integral involves the coefficient $C^{l_1 l_2 l_3}_{000}$, which similarly [12] requires $l_1+l_2+l_3 \equiv 0 \bmod 2$.

Now let us turn to the Euclidean group in three-space, $E(3)$. Rather than proceeding to develop the $E(3)$ analogs to the $SO(4)$ results directly from the $E3$ group itself, let us use the contraction method, applied to the limit $SO(4) \to E(3)$, as more intuitive and illuminating.

The $SO(4) \to E(3)$ contraction limit preserves the diagonal $SO(3)$ group, and hence uses the realization of the commutation relations in (4.1) (where $\vec{L}$ generates the diagonal $SO(3)$ subgroup). These relations are

$$[L_i, L_j] = i\varepsilon_{ijl} L_k, \quad [L_i, K_j] = i\varepsilon_{ijk} K_k, \quad \text{and} \quad [K_i, K_j] = i\varepsilon_{ijk} L_k. \tag{4.11a–c}$$

To obtain the contraction limit we multiply (4.1b,c) by ε^2 on both sides; let $\varepsilon \to 0$ and $K_i \to \infty$, such that $\varepsilon K_i \to T_i =$ finite (momentum) operator. We obtain in this limit the commutation relations

$$[L_i, L_j] = i\varepsilon_{ijl} L_k, \quad [L_i, T_j] = i\varepsilon_{ijk} T_k, \quad [T_i, T_j] = 0, \tag{4.12a–c}$$

which show that $E(3)$ has the structure of a semidirect product group $E(3) \simeq T_3 @ SO3$, with the three-dimensional translation subgroup (T_3) normal.

Consider now the $SO(4)$ irreps $\mathcal{D}^{(n,0)}(g)$ corresponding to the four-dimensional spherical harmonics (cf. (4.8a)). The contraction limit corresponds to taking the radius $\rho = (x^2+y^2+z^2+t^2)^{\frac{1}{2}}$ of a sphere in four dimensions to be "large," and considering the neighborhood about the "origin": $x = y = z = 0$ with $t = \rho$. Rotations R_{tx}, R_{ty}, R_{tz} thus become, for ρ large, translations of this neighborhood, which in the limit becomes planar. For the representation matrix (4.8a), the contraction keeps (θ, φ) fixed. The contraction takes the eigenvalue n of the invariant $\mathcal{I}_1 = \vec{L}^2 + \vec{K}^2 + 1$

to become large (corresponding to $\vec{K}$ becoming large in obtaining (4.12)). The limit $\varepsilon K_3 \to T_3 =$ finite is achieved by defining the rotation operator $e^{-i\chi K_3} \to e^{-irT_3}$, the translation operator with eigenvalue e^{-irk}, ($k = n\varepsilon =$ momentum eigenvalue of $T_3, r =$ displacement). Thus the angle χ in (4.8a) is small, given by $\chi = r\varepsilon$, with $\varepsilon \to 0$. The required asymptotic limit of (4.8a) is found [20], [22] to be

$$\lim_{n\to\infty}\left(\left(\frac{2\pi^2}{n^2}\right)^{\frac{1}{2}} Y_{nlm}(\chi\theta\varphi)\right) = i^l(4\pi)^{\frac{1}{2}} j_l(kr)Y_{lm}(\theta\varphi). \tag{4.12}$$

Put differently, the contraction limit involving the Gegenbauer polynomials is the limit

$$\lim_{n\to\infty}\left(\left(\frac{2\pi^2}{n^2}\right)^{\frac{1}{2}} F_{nl}\left(\frac{kr}{n}\right)\right) = i^l(4\pi)^{\frac{1}{2}} j_l(kr), \tag{4.13}$$

with the Gegenbauer polynomials entering through (4.8b). The special functions of $E(3)$ in (4.14) are the spherical Bessel functions, $j_l(kr)$, using the conventional notation. (It should be noted that these functions are real.)

There is one more detail before we can establish the contraction limit of the left-hand side of (4.10). This concerns the differential dg, which in (4.10) has the value

$$dg = \sin^2\chi d\chi \cdot \sin\theta d\theta d\varphi.$$

Since the angle χ is small ($\chi = \varepsilon \cdot r$ with $\varepsilon \to 0$) we see that $\sin^2\chi d\chi \to \varepsilon^3 r^2 dr$, and the limits of integration become 0 and ∞.

We now use the asymptotic limit given by (4.13) for the left-hand side of (4.10), and note that the ε^3 in dg combines with the terms $n_1n_2n_3$—introduced from using (4.13)—to give $k_1k_2k_3$, that is, $n_i\varepsilon = k_1$. This establishes that the contraction limit of the left-hand side of (4.10) is given by

$$\begin{aligned} & i^{l_1+l_2-l_3}\left(\frac{2}{\pi}\right)^{\frac{3}{2}} k_1k_2k_3\left(\int_0^\infty r^2dr\, j_{l_3}(k_3r)j_{l_1}(k_1r)j_{l_2}(k_2r)\right) \\ & \times\left(\left(\frac{(2l_1+1)(2l_2+1)}{4\pi(2l_3+1)}\right)^{\frac{1}{2}} C^{l_1l_2l_3}_{m_1m_2m_3}C^{l_1l_2l_3}_{0\,0\,0}\right), \end{aligned} \tag{4.14}$$

where the terms in the second bracket (...) use (2.11) to evaluate the ($SO(3)$) spherical harmonic integral.

Let us now consider the contraction limit for the $SO(4)$ WCG coefficients, appearing as the right-hand side of (4.10) and defined in (4.5). This limit will yield new angular special functions, and it is helpful to introduce these functions first and then show that they appear as the contraction limit of (4.5).

These new angular functions were first defined in the context of angular correlation theory [23] for nuclear radiations (for example, gamma rays). The appropriate angle functions for discussing the angular correlation of, say, *two* gamma rays are the Legendre functions of the planar angle defined by the two gamma ray directions, a rotational invariant. The new functions in question are the appropriate angular functions for correlations involving the *three* rotationally invariant angles defined by

three observed directions. Specifically, we have the definition [12], [24]

$$P_{l_1,l_2l_3}(\hat{k}_1\hat{k}_2\hat{k}_3) \equiv (4\pi)^{\frac{3}{2}}\left((2l_1+1)(2l_2+1)(\ldots)^2\right)^{-\frac{1}{2}}$$

$$\cdot i^{l_1+l_2-l_3}\sum_{\alpha,\beta,\gamma}(-1)^{\gamma}C^{l_1l_2l_3}_{\alpha\beta\gamma}Y_{l_1\alpha}(\hat{k}_1)Y_{l_2\beta}(\hat{k}_2)Y_{l_3\gamma}(\hat{k}_3), \tag{4.15}$$

where $\hat{k}_i$ are unit vectors of the directions measured with respect to a fixed, but arbitrary, coordinate frame. Despite the appearance of the (quantal) $SO(3)$ WCG coefficient in (4.16) this is a classical result, for the WCG coefficients appear here as the coefficients of a general vector algebra. For example, the P_{111} function is just the rotationally invariant, antisymmetric combination of three vectors, $\hat{k}_1 \cdot \hat{k}_2 \times \hat{k}_3$ (to within a constant).

The special functions defined by (4.16) are *real* (the factor in i is determined by time reversal invariance in quantum mechanics in order to effect this). Moreover, these functions have the symmetry property

$$P_{l_il_jl_k}(\hat{k}_i\hat{k}_j\hat{k}_k) = (-1)^{l_i+l_j+l_k}P_{l_{i'}l_{j'}l_{k'}}(\hat{k}_{i'}\hat{k}_{j'}\hat{k}_{k'}), \tag{4.16}$$

when $(i'j'k')$ is an odd permutation of (ijk). Note that the normalization is chosen such that $P_{000} = 1$.

We are now in a position to examine the contraction limit of the coefficients (4.5) appearing on the right-hand side of (4.10).

For this limit we find

$$\underset{\text{limit}}{\text{contraction}}\left(i^{l_3-l_1-l_2}\left[(2a+1)(2b+1)(2c+1)\right]^{\frac{1}{2}}\begin{Bmatrix} a & a & l_1 \\ b & b & l_2 \\ c & c & l_3 \end{Bmatrix}\right) = \begin{cases} P_{l_1l_2l_3}(\hat{a}\hat{b}\hat{c}) \\ 0 \end{cases}, \tag{4.18}$$

where the contraction limit and the two alternatives are explained as follows. First note that the $(9-j)$ coefficient appearing in (4.18) has the restriction that two columns are identical. As discussed in the remark above, this implies that $l_1+l_2+l_3 \equiv 0 \bmod 2$ in order to obtain a nonzero value for the $(9-j)$ symbol. Exactly the same restriction also applies, from (4.17), to the $P_{l_1l_2l_3}(\hat{a}\hat{b}\hat{c})$ on the right-hand side of (4.18). To understand the meaning of this restriction, recall our example where $P_{111} = (\text{constant})$ $\hat{k}_1 \cdot \hat{k}_2 \times \hat{k}_3$. Since in this example $(-1)^{l_1+l_2+l_3} = -1, P_{111}$ vanishes, which implies that the three unit vectors $\hat{k}_1, \hat{k}_2, \hat{k}_3$ are *coplanar*. This is the meaning of the restriction on the $9-j$ symbol appearing in (4.18). More precisely, the contraction limit takes the three parameters, a, b, and c large (say, $a = a'/\varepsilon, b = b'/\varepsilon, c = c'/\varepsilon$ with a', b', c' constant as $\varepsilon \to 0$) such that a, b, and c—which form an angular momentum triangle (a condition stemming from (4.5))—continue to form a finite scaled triangle in the limit, with sides a', b', c'. We express this as the statement: *the finite scaled variables a', b', c' form a triangle*, with the unit vectors $\hat{a}, \hat{b}, \hat{c}$ accordingly being coplanar. Note that the angles appearing explicitly in $P_{l_1l_2l_3}(\hat{a}\hat{b}\hat{c})$ are invariant to this scaling and hence well defined.

If the parameters a, b, c in (4.18) do not form a triangle, then both sides of (4.18) vanish.

We can now determine the contraction limit of the right-hand side of (4.10). Using (4.6a) and (4.18) we find

$$\frac{i^{l_1+l_2-l_3}}{(8\pi^2)^{\frac{1}{2}}}\left[(2l_1+1)(2l_2+1)\right]^{\frac{1}{2}} P_{l_1l_2l_3}(\hat{k}_1\hat{k}_2\hat{k}_3)C^{l_1l_2l_3}_{m_1m_2m_3}. \tag{4.19}$$

These results, (4.15), and (4.19) establish the contraction limit of the two sides, independent of (4.10). Equating the two results, and cancelling common factors, we obtain an identity for the special functions of the contracted group, $E(3)$:

$$4k_1k_2k_3\int_0^\infty r^2dr j_{l_1}(k_1r)j_{l_2}(k_2r)j_{l_3}(k_3r)\cdot(2l_3+1)^{-\frac{1}{2}}C^{l_1l_2l_3}_{0\,0\,0} = P_{l_1l_2l_3}(\hat{k}_1\hat{k}_2\hat{k}_3) \tag{4.20}$$

for $l_1+l_2+l_3\equiv 0 \bmod 2$, with k_1, k_2, and k_3 in (4.20) forming a (nondegenerate) triangle.

If the three parameters k_1, k_2, and k_3 in (4.20) form a degenerate triangle (having vanishing area), the result given above must be modified. The contraction limit, as we have noted before, can be singular, and, in fact is singular for a degenerate triangle in the present case. A more careful investigation shows that, for the degenerate case, there is a factor of $\frac{1}{2}$ appearing on the right side of (4.19). (Since the degenerate case implicitly appears in the inversion of (4.4) into (4.7) there is a subtle factor of 2 which renormalizes the limit of the right-hand side of (4.10) also.)

Combining these various cases for the final result we obtain

$$\begin{aligned} &4k_1k_2k_3\int_0^\infty r^2dr j_{l_1}(k_1r)j_{l_2}(k_2r)j_{l_3}(k_3r)\cdot(2l_3+1)^{-\frac{1}{2}}C^{l_1l_2l_3}_{0\,0\,0} \\ &\quad = \Delta\cdot P_{l_1l_2l_3}(\hat{k}_1\hat{k}_2\hat{k}_3) \end{aligned} \tag{4.21a}$$

for $l_1+l_2+l_3\equiv 0 \bmod 2$, with Δ defined by

$$\Delta = \begin{cases} 1 & \text{if } k_1k_2k_3 \text{ forms a triangle,} \\ \frac{1}{2} & \text{if } k_1k_2k_3 \text{ forms a degenerate triangle,} \\ 0 & \text{if not.} \end{cases} \tag{4.21b}$$

The identity given in (4.20) was first established by direct integration, without the use of group theory, by Jackson and Maximon [25]. *The fact that this identity is the contraction limit of the $SO(4)$ identity* (4.10)*—with the interpretation* (4.18)*, for the 9-j contraction limit—is, we believe, new.*

5. Concluding remarks. Within the last few years there has been an almost explosive growth of interest (and new results) from the discovery of a group theoretic basis for q-analog special functions, based on the concept of a *quantum group* [10], [11]. Quantum groups are not actually groups, but rather Hopf algebras, usually of infinite dimension, which are deformations of the universal enveloping algebras of classical Lie groups. Although technically mislabeled, quantum groups do share many group-like properties. The limit $q\to 1$, in which a genuine classical group is attained, is reminiscent of the quantum $\to$ classical limit, which in a sense also justifies the name.

Quantum groups, in particular, permit the construction of unitary irreps (and thereby q-analog special functions), product laws, and the construction (for a suitable classical group) of the q-analog of the Wigner–Clebsch–Gordan coefficients. Thus we see that essentially all of the necessary tools are available for a q-analog version of the

results presented above. It seems evident that such relations must exist, extending to the corresponding q-analog special function identities, the intuitive interrelationships of the contraction limit (which also exists for quantum groups).

It would be interesting to verify whether or not these conjectured q-analog relations actually exist.

Acknowledgment. It is our hope that this contribution to Dick Askey will be of interest not only to him, but also to his many friends and students who have benefitted by his extensive knowledge, highly developed skills, and many contributions to special function theory over the years.

REFERENCES

[1] E. P. WIGNER, *Application of Group Theory to the Special Functions of Mathematical Physics*, Princeton University, Princeton, NJ, 1955, unpublished lecture notes.

[2] N. I. VILENKIN, *Special Functions and the Theory of Group Representations*, Nauka, Moscow, 1965.

[3] R. A. ASKEY, T. H. KOORNWINDER, AND W. SCHEMPP, EDS., *Special Functions: Group Theoretical Aspects and Applications*, Reidel, Dordrecht, Holland, 1984, pp. 129–162.

[4] J. A. GAUNT, *The triplets of helium*, Trans. Roy. Soc., A228 (1929), pp. 151–196.

[5] W. T. SHARP, *Racah Algebra and the Contraction of Groups*, Department of Mathematics, University of Toronto, Toronto, Ontario, 1984 (based on the doctoral thesis of W. T. Sharp carried out under the direction of E. P. Wigner, Princeton University, 1960).

[6] M. DANOS, A. A. STAHLHOFEN, AND L. C. BIEDENHARN, *Intrinsic sticking in dt muon-catalyzed fusion: interplay of atomic, molecular and nuclear Phenomena*, Ann. Phys., 192 (1989), pp. 158–203.

[7] A. K. ÇIFTÇI, *A New Calculation Technique for Three-Body Final States*, Ph.D. thesis, Department of Physics, Duke University, Durham, NC, 1991.

[8] S. J. JONES, *Muon-catalysed fusion revisited*, Nature, 321 (1986), p. 127.

[9] L. I. PONOMAREV, *Muon catalysed fusion*, Contemp. Phys., 31, (1990), pp. 219–245.

[10] V. G. DRINFELD, *Quantum Groups*, ICM Proceedings, Berkeley, CA, 1986, pp. 798–820.

[11] M. JIMBO, *A q-difference analogue of $U_q g$ and the Yang–Baxter equation*, Lett. Math. Phys., 10 (1985), pp. 63–69.

[12] L. C. BIEDENHARN AND J. D. LOUCK, *Angular momentum in quantum physics*, in Encyclopedia of Mathematics and Its Applications, Vol. 8, Addison-Wesley Advanced Book Program, Reading, MA, 1981; reprinted by Cambridge University Press, Cambridge, 1989.

[13] E. P. WIGNER, *On the matrices which reduce the Kronecker products of representations of S. R. groups*, 1940, unpublished; in Quantum Theory of Angular Momentum, L. C. Biedenharn and H. van Dam, eds., Academic Press, New York, 1965, pp. 87–133.

[14] L. C. BIEDENHARN, A. GIOVANNINI, AND J. D. LOUCK, *Canonical definition of Wigner coefficients in U_n*, J. Math. Phys., 8 (1967), p. 691.

[15] E. P. WIGNER, *Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra*, Academic Press, New York, 1959; English translation by J. J. Griffin.

[16] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1944.

[17] A. ERDELYI, *Higher Transcendental Functions*, Vols. 1 and 2, McGraw-Hill, New York, 1953; *Tables of Integral Transforms*, Vols. 1 and 2, McGraw-Hill, New York, 1954.

[18] E. INÖNÜ AND E. P. WIGNER, *On the contraction of groups and their representations*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), p. 510.

[19] I. E. SEGAL, *A class of operator algebras which are determined by groups*, Duke Math. J., 18 (1951), p. 221.

[20] L. C. BIEDENHARN, *Wigner coefficients for the R_4 group and some applications*, J. Math. Phys., 2 (1961), p. 433.

[21] L. C. BIEDENHARN AND J. D. LOUCK, *The Racah–Wigner algebra in quantum theory*, in Vol. 9, Addison-Wesley, Advanced Book Program, Reading, MA, 1981.

[22] J. D. TALMAN, *Special Functions: A Group Theoretic Approach*, W. A. Benjamin, Reading, MA, 1968.

[23] L. C. BIEDENHARN AND M. E. ROSE, *Theory of Angular Correlation of Nuclear Radiations*, Rev. Modern Phys., 25 (1953), p. 729.

[24] L. C. BIEDENHARN, *Angular Correlations in Nuclear Spectroscopy*, in Nuclear Spectroscopy II, Chap. V, Fay Ajzenberg-Selove, ed., Academic Press, New York, 1960.

[25] A. D. JACKSON AND L. C. MAXIMON, *Integrals of products of Bessel functions*, SIAM J. Math. Anal., 3 (1972), p. 446.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 288–303, March 1994

004

ASYMPTOTIC APPROXIMATIONS FOR SYMMETRIC ELLIPTIC INTEGRALS*

B. C. CARLSON† AND J. L. GUSTAFSON†

Abstract. Symmetric elliptic integrals, which have been used as replacements for Legendre's integrals in recent integral tables and computer codes, are homogeneous functions of three or four variables. When some of the variables are much larger than the others, this paper presents asymptotic approximations with error bounds. In most cases they are derived from a uniform approximation to the integrand. As an application the symmetric elliptic integrals of the first, second, and third kinds are proved to be linearly independent with respect to coefficients that are rational functions.

Key words. elliptic integral, asymptotic approximation, inequalities, hypergeometric R-function

AMS subject classifications. primary 33A25, 41A60, 26D15; secondary 33A30, 26D20

1. Introduction.

A recent table of elliptic integrals [9]–[13] uses symmetric standard integrals instead of Legendre's integrals because permutation symmetry makes it possible to unify many of the formulas in previous tables. Fortran codes for numerical computation of the symmetric integrals, which are homogeneous functions of three or four variables, can be found in several major software libraries as well as in the supplements to [9] and [10]. For analytical purposes it is desirable to know how the homogeneous functions behave when some of the variables are much larger than the others. For all such cases we list in §2 asymptotic approximations (sometimes two or three approximations of different accuracy), always with error bounds. Proofs are discussed in §3. In most cases the approximations are obtained by replacing the integrand by a uniform approximation. Many of the results found by a different method in [16] have been improved by sharpening the error bounds or by finding bounds for incomplete elliptic integrals that are still useful for the complete integrals, which are then not listed separately. Cases not considered in [16] include two for a completely symmetric integral of the second kind and two for a symmetric integral of the third kind in which two variables are much larger than the other two.

We assume that x, y, z are nonnegative and that at most one of them is 0. The symmetric integral of the first kind,

$$R_F(x,y,z) = \frac{1}{2}\int_0^\infty [(t+x)(t+y)(t+z)]^{-1/2}dt, \tag{1}$$

is homogeneous of degree $-\frac{1}{2}$ in x, y, z and satisfies $R_F(x,x,x) = x^{-1/2}$. The symmetric integral of the third kind,

$$R_J(x,y,z,p) = \frac{3}{2}\int_0^\infty [(t+x)(t+y)(t+z)]^{-1/2}(t+p)^{-1}dt, \qquad p>0, \tag{2}$$

is homogeneous of degree $-3/2$ in x, y, z, p and satisfies $R_J(x,x,x,x) = x^{-3/2}$. If $p = z$, R_J reduces to an integral of the second kind,

$$R_D(x,y,z) = R_J(x,y,z,z) = \frac{3}{2}\int_0^\infty [(t+x)(t+y)]^{-1/2}(t+z)^{-3/2}dt, \qquad z>0, \tag{3}$$

* Received by the editors March 20, 1992; accepted for publication (in revised form) November 6, 1992. This work was supported by the Director of Energy Research, Office of Basic Energy Sciences. The Ames Laboratory is operated for the U.S. Department of Energy by Iowa State University under contract W-7405-ENG-82.

† Ames Laboratory and Department of Mathematics, Iowa State University, Ames, Iowa 50011-3020.

which is symmetric in x and y only. If two variables of R_F are equal, the integral becomes an elementary function,

$$R_C(x,y) = R_F(x,y,y) = \frac{1}{2}\int_0^\infty (t+x)^{-1/2}(t+y)^{-1}dt, \qquad y>0. \tag{4}$$

If $x<y$, it is an inverse trigonometric function,

$$R_C(x,y) = (y-x)^{-1/2}\arccos(x/y)^{1/2}, \tag{5}$$

and if $x>y$, it is an inverse hyperbolic function,

$$R_C(x,y) = (x-y)^{-1/2}\operatorname{arccosh}(x/y)^{1/2} = (x-y)^{-1/2}\ln\frac{\sqrt{x}+\sqrt{x-y}}{\sqrt{y}}. \tag{6}$$

If the second argument of R_C is negative, the Cauchy principal value is [18, eq. (4.8)]

$$R_C(x,-y) = \left(\frac{x}{x+y}\right)^{1/2} R_C(x+y,y), \qquad y>0. \tag{7}$$

If the fourth argument of R_J is negative, the Cauchy principal value is given by [18, eq. (4.6)]

$$\begin{aligned}(y+p)R_J(x,y,z,-p) &= (q-y)R_J(x,y,z,q) - 3R_F(x,y,z)\\ &\quad + 3\left(\frac{xyz}{xz+pq}\right)^{1/2} R_C(xz+pq,pq), \qquad p>0,\end{aligned} \tag{8}$$

where $q-y = (z-y)(y-x)/(y+p)$. If we permute the values of x,y,z so that $(z-y)(y-x)\geq 0$, then $q\geq y>0$.

A completely symmetric integral of the second kind is not as convenient as R_D for use in tables because its representation by a single integral is more complicated [7, eq. (9.1-9)]:

$$R_G(x,y,z) = \frac{1}{4}\int_0^\infty [(t+x)(t+y)(t+z)]^{-1/2}\left(\frac{x}{t+x}+\frac{y}{t+y}+\frac{z}{t+z}\right)t\,dt. \tag{9}$$

It is symmetric and homogeneous of degree $\frac{1}{2}$ in x,y,z, and it satisfies $R_G(x,x,x) = x^{1/2}$. It has a nice representation by a double integral that expresses the surface area of an ellipsoid [7, eq. (9.4-6)]. It is related to R_D and R_F by (58) and by

$$R_G(x,y,z) = x(y+z)R_D(y,z,x) + y(z+x)R_D(z,x,y) + z(x+y)R_D(x,y,z), \tag{10}$$

$$R_G(x,y,0) = xy[R_D(0,x,y) + R_D(0,y,x)]. \tag{11}$$

Legendre's complete elliptic integrals K and E are given by

$$K(k) = R_F(0,1-k^2,1), \tag{12}$$

$$\begin{aligned}E(k) &= 2R_G(0,1-k^2,1)\\ &= \frac{1-k^2}{3}[R_D(0,1-k^2,1) + R_D(0,1,1-k^2)],\end{aligned} \tag{13}$$

$$K(k) - E(k) = \frac{k^2}{3}R_D(0,1-k^2,1), \tag{14}$$

$$E(k) - (1-k^2)K(k) = \frac{k^2(1-k^2)}{3}R_D(0,1,1-k^2). \tag{15}$$

Approximations and inequalities for K, E, and some combinations thereof are given in [1]–[3]. If the error terms in (30), (31), and (53) are omitted, the approximations reduce to the leading terms of well-known series expansions of K and E for k near 1 [15, p. 54] [4, eqs. 900.06, 900.10]. If the series for K is truncated after any number of terms, simple bounds for the *relative* error are given in [14, eq. (1.17)]. A generalization of this series to $R_F(x,y,z)$ with $x,y \ll z$ is given in [14, eqs. (1.14)–(1.16)], again with simple bounds for the relative error of truncation.

The various functions designated by R with a letter subscript are special cases of the multivariate hypergeometric R-function,

$$R_{-a}(b_1,\dots,b_n;\, z_1,\dots,z_n),$$

which is symmetric in the indices $1,\dots,n$ and is homogeneous of degree $-a$ in the variables $z_1,\dots,z_n$. Best regarded as the Dirichlet average of x^{-a} [7, §5.9], it is a symmetric variant of the function known as Lauricella's F_D. By the method of Mellin transforms, series expansions that converge rapidly if some of the z's are much larger than the others and if the parameters satisfy $\sum_{i=1}^{n} b_i > a > 0$ are obtained in [8, eqs. (4.16)–(4.19)]. Thus the leading terms of these series provide asymptotic approximations for all except R_G among the functions

$$(16)\quad R_F(x,y,z) = R_{-\frac{1}{2}}(\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2};\, x,y,z), \qquad R_C(x,y) = R_{-\frac{1}{2}}(\tfrac{1}{2},1;\, x,y),$$

$$(17)\quad R_J(x,y,z,p) = R_{-\frac{3}{2}}(\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2},1;\, x,y,z,p), \qquad R_D(x,y,z) = R_{-\frac{3}{2}}(\tfrac{1}{2},\tfrac{1}{2},\tfrac{3}{2};\, x,y,z),$$

$$(18)\quad R_G(x,y,z) = R_{\frac{1}{2}}(\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2};\, x,y,z).$$

However, error bounds for the approximations are more easily derived by the methods of the present paper. Another function that is used repeatedly in obtaining error bounds is [7, Ex. 9.8-5]

$$(19)\quad R_{-1}(\tfrac{1}{2},\tfrac{1}{2},1;\, x,y,z) = \int_0^\infty [(t+x)(t+y)]^{-1/2}(t+z)^{-1}\,dt$$

$$(20)\quad = 2R_C((\sqrt{xy}+z)^2, (\sqrt{x}+\sqrt{y})^2 z).$$

In §4 the asymptotic approximations are applied to show that $R_F(x,y,z)$, $R_D(x,y,z)$, $R_J(x,y,z,p)$, and $(xyz)^{-1/2}$ are linearly independent with respect to coefficients that are rational functions of x,y,z, and p. The appendix contains some elementary inequalities that are used in obtaining error bounds.

The results in this paper provide upper and lower approximations that approach the elliptic integrals as selected ratios of the variables approach zero. Approximations that approach the integrals as all variables approach a common value have been found by other methods. For example, the theory of hypergeometric mean values yields upper and lower algebraic approximations for all the integrals in this paper [5, Thm. 2], while truncation of Taylor series about the arithmetic mean of the variables gives approximations with errors that may be positive or negative. Successive applications of the duplication theorem for R_F, making its three variables approach equality, provide ascending and descending sequences of successively sharper (and successively more complicated) algebraic approximations to R_F and R_C [6]. Transcendental approximations that approach R_F when only two of its variables approach equality are furnished by

$$(21)\quad R_C\left(x, \frac{y+z}{2}\right) \le R_F(x,y,z) \le R_C(x, \sqrt{yz}), \qquad yz \ne 0,$$

which follows from (71). The inequalities can be sharpened by first using Landen or Gauss transformations of R_F [7, §9.5] to make y and z approach equality. If $x = 0$, the Gauss transformation reduces to replacing $\sqrt{y}$ and $\sqrt{z}$ by their arithmetic and geometric means, and each R_C-function becomes $\pi/2$ divided by the square root of its second argument. Therefore, in the complete case the procedure reduces to the algorithm of the arithmetic-geometric mean [7, eqs. (6.10-6), (9.2-3)] and provides ascending and descending sequences of algebraic approximations, of which leading members are shown in (33).

2. Results. We assume throughout that x, y, and z are nonnegative and that at most one of them is 0. The last argument of R_C, R_D, and R_J is assumed to be positive (see (7) and (8)).

C1. $R_C(x,y)$ with $x \ll y$.

$$R_C(x,y) = \frac{\pi}{2\sqrt{y}} - \frac{\sqrt{x}}{y} + \frac{\pi x\theta}{4y^{3/2}}, \tag{22}$$

where $1/(1+\sqrt{x/y}) \le \theta \le 1$ with equalities if and only if $x = 0$.

C2. $R_C(x,y)$ with $y \ll x$. Two approximations of different accuracies are

$$R_C(x,y) = \frac{1}{2\sqrt{x}}\left(\ln\frac{4x}{y} + \frac{y}{2x-y}\ln\frac{\theta_1 x}{y}\right), \tag{23}$$

$$= \frac{1}{2\sqrt{x}}\left[\left(1+\frac{y}{2x}\right)\ln\frac{4x}{y} - \frac{y}{2x} + \frac{3y^2}{4x(2x-y)}\ln\frac{\theta_2 x}{y}\right], \tag{24}$$

where $1 < \theta_i < 4$ for $i = 1, 2$. The first approximation implies

$$R_C(x,y) < \frac{1}{2\sqrt{x}(1-y/2x)}\ln\frac{4x}{y}. \tag{25}$$

F1. $R_F(x,y,z)$ with $x, y \ll z$. Let $a = (x+y)/2$, $g = \sqrt{xy}$, and $\rho = \max\{x,y\}/z$. Then

$$R_F(x,y,z) = \frac{1}{2\sqrt{z}}\left(\ln\frac{8z}{a+g} + \frac{r}{2z}\right), \tag{26}$$

where

$$\frac{g}{1-g/z}\ln\frac{2z}{a+g} < r < \frac{a}{1-a/2z}\ln\frac{8z}{a+g}.$$

The upper bound implies

$$R_F(x,y,z) < \frac{1}{2\sqrt{z}(1-a/2z)}\ln\frac{8z}{a+g}. \tag{27}$$

A sharper lower bound and a higher-order approximation are given by

$$R_F(x,y,z) = \frac{1}{2\sqrt{z}}\left(\ln\frac{8z}{a+g} + \frac{ar_1}{2z}\right) \tag{28}$$

$$= \frac{1}{2\sqrt{z}}\left[\left(1+\frac{a}{2z}\right)\ln\frac{8z}{a+g} - \frac{2a-g}{2z} + \frac{3(3a^2-g^2)r_2}{16z^2}\right], \tag{29}$$

where

$$\ln\frac{z}{2a} < \frac{\ln(1/\rho)}{1-\rho} < r_i < \frac{1}{1-a/2z}\ln\frac{8z}{a+g}, \qquad i=1,2.$$

By (12) this implies (since $4k^2 < 4-k'^2$ if $k^2<1$)

$$K(k) = \ln\frac{4}{k'} + \frac{k'^2}{4-k'^2}\ln\frac{\theta_1}{k'} \tag{30}$$

$$= \left(1+\frac{k'^2}{4}\right)\ln\frac{4}{k'} - \frac{k'^2}{4} + \frac{9k'^4}{16(4-k'^2)}\ln\frac{\theta_2}{k'}, \tag{31}$$

where $0<k'=\sqrt{1-k^2}$ and $1<\theta_i<4$ for $i=1,2$.

F2. $R_F(x,y,z)$ with $z \ll x,y$. Let $a=(x+y)/2$ and $g=\sqrt{xy}$. Then

$$R_F(x,y,z) = R_F(x,y,0) - \frac{\sqrt{z}}{g} + \frac{\pi z\theta}{4g^{3/2}}, \tag{32}$$

where $1/(1+\sqrt{z/g}) < \theta < a/g$. Note that $R_F(x,y,0) = \pi/2\mathrm{AGM}(\sqrt{x},\sqrt{y})$, where AGM denotes Gauss's arithmetic-geometric mean [7, eqs. (6.10-6), (9.2-3)], and hence

$$\frac{1}{\sqrt{a}} \le \sqrt{\frac{2}{a+g}} \le \frac{2}{\sqrt{(a+g)/2}+\sqrt{g}} \le \frac{2}{\pi}R_F(x,y,0) \le \left(\frac{2}{ag+g^2}\right)^{1/4} \le \frac{1}{\sqrt{g}}, \tag{33}$$

with equalities if and only if $x=y$.

D1. $R_D(x,y,z)$ with $x,y \ll z$. Let $a=(x+y)/2$ and $g=\sqrt{xy}$. Then

$$R_D(x,y,z) = \frac{3}{2z^{3/2}}\left(\ln\frac{8z}{a+g} - 2 + \frac{\theta}{z}\ln\frac{2z}{a+g}\right), \tag{34}$$

where

$$\frac{g}{1-g/z} < \theta < \frac{3a}{2(1-a/z)}.$$

D2. $R_D(x,y,z)$ with $z \ll x,y$. Let $a=(x+y)/2$ and $g=\sqrt{xy}$. Then

$$R_D(x,y,z) = \frac{3}{\sqrt{xyz}}\left(1 - \frac{\pi\theta}{2}\sqrt{\frac{z}{g}}\right), \tag{35}$$

where

$$1 - \frac{4}{\pi}\sqrt{\frac{z}{g}} < \theta < \frac{a}{g}.$$

A higher-order approximation is

$$R_D(x,y,z) = \frac{3}{\sqrt{xyz}} - R_D(0,x,y) - R_D(0,y,x) + \frac{3\pi\theta\sqrt{z}}{2g^2(1+\sqrt{z/g})}, \tag{36}$$

where

$$\frac{1}{\sqrt{2/3+\sqrt{z/g}}} < \theta < \frac{3a}{2g(1+\sqrt{z/g})}.$$

An approximation of still higher order is

$$R_D(x,y,z) = \frac{3}{\sqrt{xyz}} - \frac{6}{xy}R_G(x,y,0) + \frac{6a\sqrt{z}}{g^3}\left(1 - \frac{\pi\theta}{4}\sqrt{\frac{z}{a}}\right), \tag{37}$$

where we have used (11) and where

$$\frac{1}{1+\sqrt{z/a}} < \theta < \left(\frac{a}{g}\right)^{3/2}\left(3 - \frac{g^2}{a^2}\right).$$

D3. $R_D(x,y,z)$ with $y,z \ll x$. Let $a = (y+z)/2$ and $g = \sqrt{yz}$. Then

$$R_D(x,y,z) = \frac{3}{\sqrt{x}}\left(\frac{1}{g+z} - \frac{r}{4x}\right), \tag{38}$$

where

$$\frac{1}{1-g/x}\ln\frac{2x}{a+g} - \frac{2z}{g+z} < r < \frac{1}{1-a/2x}\ln\frac{8x}{a+g}.$$

D4. $R_D(x,y,z)$ with $x \ll y,z$. Let $a = (y+z)/2$ and $g = \sqrt{yz}$. Then

$$R_D(x,y,z) = R_D(0,y,z) + \frac{3\sqrt{x}}{gz}\left(-1 + \frac{\pi\theta}{4}\sqrt{\frac{x}{a}}\right), \tag{39}$$

where

$$\frac{1}{1+\sqrt{x/a}} < \theta < \left(\frac{a}{g}\right)^{3/2}\left(1 + \frac{y}{a}\right).$$

J1. $R_J(x,y,z,p)$ with $x,y,z \ll p$. Let $a = (x+y+z)/3$ and $b = (\sqrt{3}/2)(xy+xz+yz)^{1/2}$. Then

$$R_J(x,y,z,p) = \frac{3}{p}R_F(x,y,z) + \frac{3\pi}{2p^{3/2}}(-1+r), \tag{40}$$

where

$$\frac{\sqrt{b/p}}{1+\sqrt{b/p}} < r < \frac{3}{2}\,\frac{\sqrt{a/p}}{1+\sqrt{a/p}}.$$

In the complete case a sharper result is

$$R_J(x,y,0,p) = \frac{3}{p}\left(R_F(x,y,0) - \frac{\pi}{2\sqrt{p}}\right)\left(1 + \frac{\theta/p}{1-\theta/p}\right), \tag{41}$$

where $\sqrt{xy} \le \theta \le (x+y)/2$, with equalities if and only if $x = y$.

J2. $R_J(x,y,z,p)$ with $p \ll x,y,z$. Let $g = (xyz)^{1/3}$, $3h^{-1} = x^{-1}+y^{-1}+z^{-1}$, and $\lambda = \sqrt{xy}+\sqrt{xz}+\sqrt{yz}$. Note that g is the geometric mean and h is the harmonic mean; whence $g \ge h$ with equality if and only if $x = y = z$. Then

$$R_J(x,y,z,p) = \frac{3}{2\sqrt{xyz}}\left(\ln\frac{4g}{p} - 2 + r\right), \tag{42}$$

where

$$-\ln\frac{g}{h} < r < \frac{3p}{2(g-p)}\ln\frac{g}{p}.$$

A higher-order approximation is

$$R_J(x,y,z,p) = \frac{3}{2\sqrt{xyz}}\ln\frac{4xyz}{p\lambda^2} + 2R_J(x+\lambda, y+\lambda, z+\lambda, \lambda) + \frac{3pr}{4\sqrt{xyz}}, \tag{43}$$

where

$$\frac{2}{g-p}\ln\frac{g}{p} < r < \frac{3}{h-p}\ln\frac{h}{p}.$$

The second term in the approximation is independent of p but is otherwise as complicated as the function being approximated. The same is true of an even more accurate approximation [16, Thm. 11] in which the error is of order p instead of $p\ln p$ and the leading term involves R_C.

J3. $R_J(x,y,z,p)$ with $x,y \ll z,p$. Let $a = (x+y)/2$ and $g = \sqrt{xy}$. Then

$$R_J(x,y,z,p) = \frac{3}{2\sqrt{zp}}\left[\ln\frac{8z}{a+g} - 2R_C\left(1,\frac{p}{z}\right) + \frac{\theta}{p}\ln\frac{2p}{a+g}\right], \tag{44}$$

where

$$\frac{g}{1-g/p} < \theta < \frac{a}{1-a/p}\left(1+\frac{p}{2z}\right).$$

J4. $R_J(x,y,z,p)$ with $z,p \ll x,y$. Let $a = (x+y)/2$, $g = \sqrt{xy}$, $b = \sqrt{3p(p+2z)}/2$, and $d = (z+2p)/3$. Then

$$R_J(x,y,z,p) = \frac{3}{g}R_C(z,p) - \frac{3\theta}{g-p}\left[R_C(z,g) - \frac{p}{g}R_C(z,p)\right], \tag{45}$$

where $1 \le \theta \le a/g$ with equalities if and only if $x = y$. Since $z \ll g$, $R_C(z,g)$ can be estimated from (22). In the complete case (45) reduces to

$$R_J(x,y,0,p) = \frac{3\pi}{2\sqrt{xyp}}\left(1 - \frac{\theta\sqrt{p}}{\sqrt{g}+\sqrt{p}}\right), \tag{46}$$

with θ as before. A higher-order approximation is

$$R_J(x,y,z,p) = \frac{3}{g}R_C(z,p) - \frac{6}{xy}R_G(x,y,0) + \frac{3\pi\theta}{2xy}, \tag{47}$$

where we have used (11) and where

$$\frac{\sqrt{b}}{1+\sqrt{b/g}} < \theta < \frac{3a}{2g}\frac{\sqrt{d}}{1+\sqrt{d/g}}.$$

J5. $R_J(x,y,z,p)$ with $x \ll y,z,p$. Let $a = (y+z)/2$ and $g = \sqrt{yz}$. Then

$$R_J(x,y,z,p) = R_J(0,y,z,p) + \frac{3\sqrt{x}}{gp}\left(-1 + \frac{\pi\theta}{4}\sqrt{\frac{x}{g}}\right), \tag{48}$$

where

$$\frac{\sqrt{g/a}}{1+\sqrt{x/a}} < \theta < \frac{a}{g} + \frac{g}{p}.$$

J6. $R_J(x,y,z,p)$ with $y,z,p \ll x$. Let $a = (y+z)/2$ and $g = \sqrt{yz}$. Then

$$R_J(x,y,z,p) = \frac{3}{\sqrt{x}}\left[R_C((g+p)^2, 2(a+g)p) - \frac{r}{4}\right], \tag{49}$$

where

$$\frac{1}{x-g}\ln\frac{2x}{a+g} - \frac{2p}{x}R_C((g+p)^2,2(a+g)p) < r < \frac{1}{x-a/2}\ln\frac{8x}{a+g}.$$

In the complete case this reduces to

$$R_J(x,0,z,p) = \frac{3}{\sqrt{xp}}R_C(p,z) - \frac{3s}{4x^{3/2}}, \tag{50}$$

where

$$\ln\frac{4x}{z} - 2\sqrt{p}R_C(p,z) < s < \frac{1}{1-z/4x}\ln\frac{16x}{z}.$$

G1. $R_G(x,y,z)$ with $x,y \ll z$. Let $a = (x+y)/2$ and $g = \sqrt{xy}$. Then

$$R_G(x,y,z) = \frac{\sqrt{z}}{2}\left(1 + \frac{r}{2z}\right), \tag{51}$$

where

$$\frac{a+g}{2}\ln\frac{2z}{a+g} + 2g - \frac{4a}{3} < r < (3a-g)\ln\frac{2z}{a+g} + 2g - \frac{a}{3}.$$

In the right-hand inequality it is assumed that $5a < z$. A sharper result for the complete case is

$$R_G(0,y,z) = \frac{\sqrt{z}}{2} + \frac{y}{8\sqrt{z}}\left(\ln\frac{16z}{y} - 1 + \frac{ys}{2z}\right), \tag{52}$$

where

$$\frac{3}{4}\ln\frac{z}{y} < s < \frac{1}{1-y/z}\left(\ln\frac{16z}{y} - \frac{13}{6}\right).$$

By (13) this follows from

$$E(k) = 1 + \frac{k'^2}{2}\left(\ln\frac{4}{k'} - \frac{1}{2} + k'^2 r\right), \tag{53}$$

where $0 < k' = \sqrt{1-k^2} \ll 1$ and

$$\frac{3}{8}\ln\frac{1}{k'} < r < \frac{1}{k(1+k)}\left(\ln\frac{4}{k'} - \frac{13}{12}\right).$$

G2. $R_G(x,y,z)$ with $z \ll x,y$. Let $a = (x+y)/2$ and $g = \sqrt{xy}$. Then

$$R_G(x,y,z) = R_G(x,y,0) + \pi\theta z/8, \tag{54}$$

where

$$\frac{1}{\sqrt{a}}\left(1 - \frac{4}{\pi}\sqrt{\frac{z}{a}}\right) < \theta < \left(\frac{2}{ag+g^2}\right)^{1/4} \le \frac{1}{\sqrt{g}}$$

with equality if and only if $x = y$.

3. Proofs. Most of the results in §2 are obtained by replacing an integrand f by an approximation f_a, writing $\int f = \int f_a + \int (f - f_a)$, and finding upper and lower bounds for $\int (f - f_a)$. All integrals are taken over the positive real line. The function f_a is usually chosen to be a uniform approximation $f_a = f_i + f_o - f_m$, where f_i is an approximation in the inner region, f_o is an approximation in the outer region, and f_m is an approximation in the overlap, or matching, region. For instance, if $f(t) = [(t+x)(t+y)(t+z)]^{-1/2}$, with $x, y \ll z$, we get f_i by neglecting t compared to z, f_o by neglecting x and y compared to t, and f_m by doing both. A first example of this process is the proof of Lemma 1.

LEMMA 1. *If $x \geq 0$, $y \geq 0$, and $0 < x + y \ll z$, then*

$$\int_0^\infty \frac{dt}{\sqrt{(t+x)(t+y)}(t+z)} = \frac{1}{z-\theta} \ln \frac{2z}{a+g}, \tag{55}$$

where $\sqrt{xy} = g \leq \theta \leq a = (x+y)/2$ with equalities if and only if $x = y$.

Proof. Let

$$f(t) = \frac{1}{\sqrt{(t+x)(t+y)}(t+z)}, \qquad f_i(t) = \frac{1}{z\sqrt{(t+x)(t+y)}},$$
$$f_o(t) = \frac{1}{t(t+z)}, \qquad f_m(t) = \frac{1}{zt}.$$

Taking $f_a = f_i + f_o - f_m$, we find

$$\int_0^\infty f_a(t)dt = \frac{1}{z} \ln \frac{2z}{a+g}$$

and

$$f - f_a = \frac{t}{z(t+z)} \left(\frac{1}{t} - \frac{1}{\sqrt{(t+x)(t+y)}} \right).$$

Inequality (64) in the appendix implies

$$f - f_a = \frac{\theta}{z\sqrt{(t+x)(t+y)}(t+z)}, \qquad g \leq \theta \leq a,$$

and thus

$$\int f = \int f_a + \int (f - f_a) = \int f_a + \frac{\theta}{z} \int f = \frac{1}{1-\theta/z} \int f_a = \frac{1}{z-\theta} \ln \frac{2z}{a+g}. \qquad \square$$

As a second example, in which Lemma 1 is used, consider $R_F(x,y,z)$ with $x, y \ll z$. Let

$$f(t) = \frac{1}{\sqrt{(t+x)(t+y)(t+z)}}, \qquad f_i(t) = \frac{1}{\sqrt{(t+x)(t+y)z}},$$
$$f_o(t) = \frac{1}{t\sqrt{t+z}}, \qquad f_m(t) = \frac{1}{\sqrt{z}t}.$$

Taking $f_a = f_i + f_o - f_m$, we find (with a and g the same as before)

$$\int_0^\infty f_a(t)dt = \frac{1}{\sqrt{z}} \ln \frac{8z}{a+g}$$

and

$$f - f_a = \left(\frac{1}{\sqrt{z}} - \frac{1}{\sqrt{t+z}}\right)\left(\frac{1}{t} - \frac{1}{\sqrt{(t+x)(t+y)}}\right).$$

Inequalities (61) and (64) imply

$$\frac{g}{2\sqrt{z(t+x)(t+y)}(t+z)} < f - f_a < \frac{a}{2z\sqrt{(t+x)(t+y)(t+z)}}.$$

Hence, by Lemma 1,

$$\frac{g}{2\sqrt{z}(z-g)} \ln\frac{2z}{a+g} < \int (f - f_a) < \frac{a}{2z}\int f < \frac{a/2z}{1-a/2z}\int f_a,$$

where the last inequality follows from the next to last. We complete the proof of (26) by noting that

$$2R_F(x,y,z) = \int f = \int f_a + \int (f - f_a).$$

Equations (28) and (29) are obtained from [14, eqs. (2.15), (3.25)] with $w = \infty$. To derive (32) we construct $f_a = f_i + f_o - f_m$ as usual and find bounds for $\int (f - f_a)$ by using (60) and (65). To simplify the upper bound we note that $R_F(x,y,z) \le R_F(x,y,0)$ and use (33).

Equations (22), (23), (24), and (25) follow from (32), (26), (29), and (27), respectively, by replacing x by y, replacing z by x, and simplifying.

Among the approximations for R_D, we need discuss only (35) and (37) since (34), (36), (38), and (39) follow from (44), (47), (49), and (48), respectively, by putting $p = z$ and simplifying. To prove (35) we let

$$f(t) = \frac{1}{\sqrt{(t+x)(t+y)}(t+z)^{3/2}}, \qquad f_i(t) = \frac{1}{g(t+z)^{3/2}},$$

choose $f_a = f_i$, and apply (65) to get

$$\frac{t}{g(t+g)(t+z)^{3/2}} \le f_a - f \le \frac{at}{g^2\sqrt{(t+x)(t+y)}(t+z)^{3/2}},$$

$$\frac{1}{g(g-z)\sqrt{t+z}}\left(\frac{g}{t+g} - \frac{z}{t+z}\right) \le f_a - f < \frac{a}{g^2\sqrt{t+z}(t+g)},$$

$$\frac{2}{g-z}\left[R_C(z,g) - \frac{\sqrt{z}}{g}\right] \le \int (f_a - f) < \frac{2a}{g^2} R_C(z,g).$$

Use of (22) completes the proof. Approximation (37) follows from applying (39) to two terms on the right side of

$$R_D(x,y,z) = 3(xyz)^{-1/2} - R_D(z,x,y) - R_D(z,y,x), \tag{56}$$

an identity that comes from [7, eqs. (5.9-5) and (6.8-15)].

In discussing approximations for R_J, we define

$$f(t) = \frac{1}{\sqrt{(t+x)(t+y)(t+z)}(t+p)}$$

and construct f_i, f_o, and f_m for each case in the manner described at the beginning of this section. For example, if $x, y, z \ll p$, then f_i is obtained by neglecting t compared to p. Unless otherwise stated, we define $f_a = f_i + f_o - f_m$, take $\int f_a$ as an approximation to $\int f$, and find bounds for $\int(f - f_a)$ by using the inequalities in the appendix.

To prove (40) we use (69). To prove (41) we use (64) and note that $\int(f - f_a) = (\theta/p)\int f$. Before discussing (42), we consider (43), in which the error bounds are easily found by using (70). Finding $\int f_a$ requires an integration by parts and a formula of which we omit the proof,

$$\begin{aligned}(57)\qquad &\int_0^\infty (\ln\, t)\frac{d}{dt}[(t+x)(t+y)(t+z)]^{-1/2}dt \\ &\quad = \frac{1}{\sqrt{xyz}}\ln\frac{\lambda^2}{4xyz} - \frac{4}{3}R_J(x+\lambda, y+\lambda, z+\lambda, \lambda),\end{aligned}$$

where $\lambda = \sqrt{xy} + \sqrt{xz} + \sqrt{yz}$. To have a simpler approximation (42), we define $f_a = f_i + f_o - f_m$ and $\phi_a = f_i + f_s - f_m$, where f_o has been replaced by

$$f_s(t) = \frac{1}{t(t+g)^{3/2}}.$$

Then

$$\int \phi_a = \frac{1}{\sqrt{xyz}}\left(\ln\frac{4g}{p} - 2\right),$$

and an upper bound for $\int(f-\phi_a)$ is found by using $\sqrt{(t+x)(t+y)(t+z)} \geq (t+g)^{3/2}$ and (63). To find a lower bound, we note that $f - f_a > 0$, whence

$$f - \phi_a = f - f_a + f_o - f_s > f_o - f_s.$$

A lower bound for $\int(f_o - f_s)$ follows from (73).

The straightforward proof of (44) uses (64), (67), and Lemma 1. For the elementary approximation (45) we choose $f_a = f_i$ and use (66). For the more accurate approximation (47) we take $f_a = f_i + f_o - f_m$ and evaluate $\int f_a$ by integrating by parts. The error bounds follow from (66) and (69) with two variables equated. To find the error bounds for (48), we use (68), (60), and (71) to prove

$$\frac{\sqrt{t}}{(t+x)(t+a)} < \frac{2gp}{x}(f - f_a) < \left(\frac{a}{g} + \frac{g}{p}\right)\frac{1}{\sqrt{t+z}(t+g)}.$$

After integration, (22) is used to complete the proof. In the case of (49), where $\int(f_o - f_m)$ is infinite, we choose $f_a = f_i$ and evaluate $\int f_a$ by (20). It follows from (61) that

$$\frac{1}{2\sqrt{x(t+y)(t+z)}}\left(\frac{1}{t+x} - \frac{p}{x(t+p)}\right) < f_a - f < \frac{1}{2x\sqrt{(t+x)(t+y)(t+z)}},$$

where we have replaced $t/(t+p)$ by 1 in the upper bound and $x/(x-p)$ by 1 in the lower bound. We then use (20), (55), and (27).

The function R_G can be expressed in terms of R_F and R_D by (17) and [7, Table 9.3-1]:

$$2R_G(x,y,z) = zR_F(x,y,z) - \frac{1}{3}(z-x)(z-y)R_D(x,y,z) + \sqrt{\frac{xy}{z}}. \tag{58}$$

Applying (26) and (34), we obtain (51). The error bounds have been substantially simplified by using the numerical value of $\ln 2$ and assuming $5a < z$ in the upper bound. It is not hard to obtain (53) from a well-known infinite series [15, p. 54] for $E(k)$ by using the inequality

$$1 + \frac{3}{8}k'^2 < {}_2F_1(\tfrac{1}{2}, \tfrac{3}{2}; 2; k'^2) < (1-k'^2)^{-1/2} = 1/k, \qquad 0 < k' < 1,$$

for the hypergeometric function ${}_2F_1$. Unfortunately, (58) does not lead to simple error bounds for (54). Instead, we define $f(z) = R_G(x,y,z)$ and find from [7, eqs. (5.9-9), (6.8-6)] that

$$f'(z) = \frac{1}{8}\int_0^\infty \frac{t\,dt}{\sqrt{(t+x)(t+y)}(t+z)^{3/2}}.$$

Since this is a strictly decreasing function of z, the mean value theorem yields $f(z) = f(0) + zf'(\zeta)$ where

$$f'(z) < f'(\zeta) < f'(0) = \tfrac{1}{4}R_F(x,y,0).$$

By (71) and (5) we see that

$$f'(z) \geq \frac{1}{8}\int_0^\infty \frac{t\,dt}{(t+a)(t+z)^{3/2}} = \frac{1}{4(a-z)}[-\sqrt{z} + aR_C(z,a)].$$

Use of (33) and (22) completes the proof of (54).

4. Application to linear independence. In [7, Thm. 9.2-1] it is shown that $R_F(x,y,z)$, $R_G(x,y,z)$, an integral of the third kind called $R_H(x,y,z,p)$, and the algebraic function $(xyz)^{-1/2}$ are linearly independent with respect to coefficients that are rational functions of x,y,z,p. It then follows [7, §9.2] that every elliptic integral can be expressed in terms of R_F, R_G, R_H, and elementary functions. From (58) and a known relation expressing R_H in terms of R_J and R_F, we may conclude that every elliptic integral can be expressed in terms of R_F, R_D, R_J, and elementary functions. To reach the same conclusion without invoking R_G and R_H, we shall use the results of this paper to prove the linear independence of R_F, R_D, R_J, and $(xyz)^{-1/2}$ with respect to coefficients that are rational functions.

THEOREM 1. *The functions $R_F(x,y,z)$, $R_D(x,y,z)$, $R_J(x,y,z,p)$, and $(xyz)^{-1/2}$ are linearly independent with respect to coefficients that are rational functions of x,y,z, and p.*

Proof. Let α, β, γ, and δ be rational functions of x,y,z, and p. We need to prove that

$$\alpha R_F(x,y,z) + \beta R_D(x,y,z) + \gamma R_J(x,y,z,p) + \delta\,(xyz)^{-1/2} \equiv 0 \tag{59}$$

if and only if α, β, γ, and δ are identically 0. We may assume that these coefficients are polynomials since we can multiply all terms by the denominator of any rational

function. As $p \to 0$, (42) shows that $R_J(x,y,z,p)$ involves $\ln p$, whereas all other quantities are polynomials in p, whence $\gamma \equiv 0$. As $z \to \infty$ we have

$$\alpha = az^m(1+O(1/z)), \qquad \beta = bz^n(1+O(1/z)),$$

where m and n are nonnegative integers and a and b are polynomials in x, y, and p. Using (26) and (34) and multiplying all terms by $2z^{3/2}$, we find

$$az^{m+1}\left[\ln\frac{8z}{a+g}+O\left(\frac{\ln z}{z}\right)\right]+3bz^n\left[\ln\frac{8z}{a+g}-2+O\left(\frac{\ln z}{z}\right)\right]+2\delta\,(xy)^{-1/2}z\equiv 0.$$

Cancellation of the leading terms in $\ln z$ requires $az^{m+1}+3bz^n \equiv 0$, implying that $n = m+1$ and $a \equiv -3b$ and leaving

$$O(z^m \ln z) - 6bz^{m+1} + 2\delta\,(xy)^{-1/2}z \equiv 0.$$

Because the second term is of different order from the first and does not have a square root in common with the third, it follows that $b \equiv 0$, whence also $a \equiv 0$. Since the leading terms of the polynomials α and β are identically 0, so too are α and β. Finally, with only one term remaining in (59), we have $\delta \equiv 0$. □

It is an open question whether Theorem 1 is still true if the coefficients are algebraic functions instead of rational functions. However, polynomial coefficients suffice (see the first paragraph of [7, §9.2]) to prove that every elliptic integral can be expressed in terms of R_F, R_D, R_J, and elementary functions.

Appendix: Elementary inequalities. Assuming x, y, z, and t are positive, we list and prove some inequalities that are used in this paper to obtain error bounds:

$$\frac{x}{2\sqrt{t}(t+x)} < \frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t+x}} < \frac{x}{2t\sqrt{t+x}}, \tag{60}$$

$$\frac{t}{2\sqrt{x}(t+x)} < \frac{1}{\sqrt{x}} - \frac{1}{\sqrt{t+x}} < \frac{t}{2x\sqrt{t+x}}, \tag{61}$$

$$\frac{1}{t^{3/2}} - \frac{1}{(t+x)^{3/2}} = \frac{\theta x}{t^{3/2}(t+x)}, \qquad 1 < \theta < \frac{3}{2}, \tag{62}$$

$$\frac{1}{x^{3/2}} - \frac{1}{(t+x)^{3/2}} = \frac{\theta t}{x^{3/2}(t+x)}, \qquad 1 < \theta < \frac{3}{2}. \tag{63}$$

In the next five inequalities let $a = (x+y)/2$ and $g = \sqrt{xy}$. Inequalities become equalities in (64), (65), and (66) if and only if $x = y$.

$$\frac{1}{t} - \frac{1}{\sqrt{(t+x)(t+y)}} = \frac{\theta}{t\sqrt{(t+x)(t+y)}}, \qquad g \le \theta \le a, \tag{64}$$

$$\frac{t}{g(t+g)} \le \frac{1}{\sqrt{xy}} - \frac{1}{\sqrt{(t+x)(t+y)}} \le \frac{at}{g^2\sqrt{(t+x)(t+y)}}, \tag{65}$$

or, alternatively,

$$\frac{1}{\sqrt{xy}} - \frac{1}{\sqrt{(t+x)(t+y)}} = \frac{\theta t}{g(t+g)}, \qquad 1 \le \theta \le \frac{a}{g}, \tag{66}$$

$$\frac{1}{\sqrt{xy}} - \frac{1}{\sqrt{t+x}(t+y)} = \frac{\theta t}{\sqrt{x}y(t+y)}, \qquad 1 < \theta < 1 + \frac{y}{2x}, \tag{67}$$

$$\frac{1}{\sqrt{xyz}} - \frac{1}{\sqrt{(t+x)(t+y)}(t+z)} = \frac{\theta t}{gz\sqrt{(t+x)(t+y)}}, \qquad 1 < \theta < \frac{a}{g} + \frac{g}{z}. \tag{68}$$

Finally we have

$$\frac{b}{t^{3/2}(t+b)} < \frac{1}{t^{3/2}} - \frac{1}{\sqrt{(t+x)(t+y)(t+z)}} < \frac{3a}{2t^{3/2}(t+a)}, \tag{69}$$

where $a = (x+y+z)/3$ and $b = \sqrt{3(xy+xz+yz)}/2$, and

$$\frac{t}{g^{3/2}(t+g)} < \frac{1}{\sqrt{xyz}} - \frac{1}{\sqrt{(t+x)(t+y)(t+z)}} < \frac{3t}{2g^{3/2}(t+h)}, \tag{70}$$

where $g = (xyz)^{1/3}$ and $3h^{-1} = x^{-1} + y^{-1} + z^{-1}$.

To prove (60) we write

$$\frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t+x}} = \frac{\sqrt{t+x} - \sqrt{t}}{\sqrt{t(t+x)}} = \frac{x}{\sqrt{t(t+x)}(\sqrt{t+x} + \sqrt{t})}$$

and replace the last denominator factor by either $2\sqrt{t}$ or $2\sqrt{t+x}$. Interchange of t and x leads from (60) to (61). To prove (62) let $y = \sqrt{1+x/t}$ and write

$$\frac{t^{3/2}(t+x)}{x}\left(\frac{1}{t^{3/2}} - \frac{1}{(t+x)^{3/2}}\right) = \frac{y^2}{y^2-1}\left(1 - \frac{1}{y^3}\right) = 1 + \frac{1}{y(y+1)},$$

which increases from 1 to 3/2 as t increases from 0 to ∞ and y decreases from ∞ to 1. Interchange of t and x leads from (62) to (63).

If the left side of (64) is put over a common denominator, it suffices to observe that

$$t + g \le \sqrt{(t+x)(t+y)} \le t + a. \tag{71}$$

The left inequality is enough to prove the left inequality in (65). To prove the right inequality in (65), we define

$$\phi(t) = \left(\sqrt{(t+x)(t+y)} - \sqrt{xy}\right)/t$$

and note that $\phi(t)$ tends to a/g as $t \to 0$ and to 1 as $t \to \infty$. Differentiation shows that ϕ decreases monotonically, because

$$t^2\sqrt{(t+x)(t+y)}\phi' = -(ta+g^2) + [(ta+g^2)^2 - t^2(a^2-g^2)]^{1/2} \le 0,$$

with equality if and only if $x = y$. Because of (71), (65) implies (66).

Equation (67) is proved by solving for θ and using (61). Likewise, (68) is proved by solving for

$$\theta = \phi(t) + \frac{t}{t+z}$$

and using the result just established that $1 \le \phi(t) \le a/g$.

To prove (69) we use Maclaurin's inequality [17, Thm. 52] to find that

$$t^3 + 2bt^2 + 4b^2t/3 < (t+x)(t+y)(t+z) \le (t+a)^3,$$

and hence

$$\sqrt{t}(t+b) < \sqrt{(t+x)(t+y)(t+z)} \le (t+a)^{3/2}. \tag{72}$$

Inequality (69) follows from this and (62).

The proof of (70) uses Maclaurin's inequality and the inequality of arithmetic and geometric means to get

$$\frac{t+g}{g} \le \left[\frac{(t+x)(t+y)(t+z)}{xyz}\right]^{1/3} = \left[\left(1+\frac{t}{x}\right)\left(1+\frac{t}{y}\right)\left(1+\frac{t}{z}\right)\right]^{1/3} \le 1+\frac{t}{h},$$

with equalities if and only if $x = y = z$, whence

$$(t+g)^{3/2} \le \sqrt{(t+x)(t+y)(t+z)} \le \left(\frac{g}{h}\right)^{3/2}(t+h)^{3/2}. \tag{73}$$

Two applications of (63) complete the proof of (70).

Acknowledgment. We thank Arthur Gautesen for suggesting the use of uniform approximations.

REFERENCES

[1] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Functional inequalities for complete elliptic integrals*, SIAM J. Math. Anal., 21 (1990), pp. 536–549.

[2] ———, *Functional inequalities for hypergeometric functions and complete elliptic integrals*, SIAM J. Math. Anal., 23 (1992), pp. 512–524.

[3] 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.

[4] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists*, 2nd ed., Springer-Verlag, New York, 1971.

[5] B. C. Carlson, *Some inequalities for hypergeometric functions*, Proc. Amer. Math. Soc., 17 (1966), pp. 32–39.

[6] ———, *Inequalities for a symmetric elliptic integral*, Proc. Amer. Math. Soc., 25 (1970), pp. 698–703.

[7] ———, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[8] ———, *The hypergeometric function and the R-function near their branch points*, Rend. Sem. Mat. Univ. Politec. Torino, Fascicolo speciale (1985), pp. 63–89.

[9] ———, *A table of elliptic integrals of the second kind*, Math. Comp., 49 (1987), pp. 595–606; supplement, pp. S13–S17.

[10] ———, *A table of elliptic integrals of the third kind*, Math. Comp., 51 (1988), pp. 267–280; supplement, pp. S1–S5.

[11] ———, *A table of elliptic integrals: Cubic cases*, Math. Comp., 53 (1989), pp. 327–333.

[12] ———, *A table of elliptic integrals: One quadratic factor*, Math. Comp., 56 (1991), pp. 267–280.

[13] B. C. CARLSON, *A table of elliptic integrals: Two quadratic factors*, Math. Comp., 59 (1992), pp. 165–180.
[14] B. C. CARLSON AND J. L. GUSTAFSON, *Asymptotic expansion of the first elliptic integral*, SIAM J. Math. Anal., 16 (1985), pp. 1072–1092.
[15] A. CAYLEY, *Elliptic Functions*, 2nd ed., Dover, New York, 1961.
[16] J. L. GUSTAFSON, *Asymptotic Formulas for Elliptic Integrals*, Ph. D. thesis, Iowa State University, Ames, IA, 1982.
[17] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, 2nd ed., Cambridge University Press, London, 1959.
[18] D. G. ZILL AND B. C. CARLSON, *Symmetric elliptic integrals of the third kind*, Math. Comp., 24 (1970), pp. 199–214.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 304–321, March 1994

005

UNIFORM AIRY-TYPE EXPANSIONS OF INTEGRALS*

A. B. OLDE DAALHUIS† AND N. M. TEMME‡

Abstract. A new method for representing the remainder and coefficients in Airy-type expansions of integrals is given. The quantities are written in terms of Cauchy-type integrals and are natural generalizations of integral representations of Taylor coefficients and remainders of analytic functions. The new approach gives a general method for extending the domain of the saddle-point parameter to unbounded domains. As a side result the conditions under which the Airy-type asymptotic expansion has a double asymptotic property become clear. An example relating to Laguerre polynomials is worked out in detail. How to apply the method to other types of uniform expansions, for example, to an expansion with Bessel functions as approximants, is explained. In this case the domain of validity can be extended to unbounded domains and the double asymptotic property can be established as well.

Key words. uniform asymptotic expansions of integrals, Airy approximation, Bessel function, Laguerre polynomial, Bessel approximation

AMS subject classifications. 41A60, 30E20, 33A40, 33A65

1. Introduction.

Many problems in mathematical physics and special functions lead to integral representations of the form

$$F(z,\alpha) = \int_{\mathcal{C}} e^{zf(x,\alpha)} g(x)\,dx, \tag{1.1}$$

where $\mathcal{C}$ is a contour in the complex plane, z is a large parameter, and f and g are analytic functions on a neighborhood of $\mathcal{C}$. In Airy-type expansions f depends on a parameter α, the *saddle-point parameter*, that describes the location of the saddle points. For a critical value of α, say, $\alpha = 0$, two saddle points coalesce with each other. With the cubic transformation $x \mapsto w$, given by

$$f(x,\alpha) = \tfrac{1}{3}w^3 - b^2 w + c \tag{1.2}$$

and suggested by Chester, Friedman, and Ursell [3], an asymptotic expansion for large values of z in terms of Airy functions can be obtained, this expansion being uniformly valid with respect to α as α ranges over a connected set containing the critical value 0 in its interior. The parameters b and c are determined explicitly from the requirement that the transformation (1.2) is analytic on a neighborhood of the two saddle points. Transformation (1.2) yields the standard form

$$\frac{1}{2\pi i}\int_{\mathcal{L}} e^{z(\frac{1}{3}w^3 - b^2 w)} h_0(w)\,dw, \tag{1.3}$$

where

$$h_0(w) = g(x(w))\frac{dx}{dw}.$$

*Received by the editors September 15, 1990; accepted for publication (in revised form) March 30, 1992.

†Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742-2431.

‡Centrum voor Wiskunde en Informatica, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (nicot@cwi.nl).

The phase function has two saddle points at $w = \pm b$. In the transformed integral (1.3) we call b the saddle-point parameter. The integral (1.3) has a turning-point character: the behavior changes strongly when b varies from real to imaginary values. When $b = 0$, the saddle points coalesce at $w = 0$.

The method for obtaining the Airy-type expansion, based on an integration-by-parts method, is introduced for a different class of integrals in Bleistein [1]. It is described for the Airy case in Bleistein and Handelsman [2, §9.2], in Olver [10, §§9.12, 9.13], and in Wong [19, §7.5].

The purpose of the paper is to present a new method for representing the remainder (and the coefficients) in Airy-type expansions. Two new aspects with respect to the saddle-point parameter are introduced in this way.

(i) A general method is described for extending the domain of this parameter to unbounded domains, by taking into account the singularities of the integrand function (especially the distance between the singularities and the relevant saddle point). The extension is possible since the order estimates of the remainder include information on the behavior of the remainder as the saddle-point parameter tends to infinity.

(ii) The method clearly describes the condition needed for the double asymptotic property of the expansion. That is, under certain conditions, the roles of the large parameter and the saddle-point parameter may be interchanged in describing the asymptotic phenomena. For instance, our analysis shows that the Airy-type expansion of the Laguerre polynomials given in [7] does not have the double asymptotic property, although the domain of uniform validity is indeed unbounded, as is claimed in [7].

Our method is based on a new class of rational functions with which the remainders in the expansions can be represented in a manner that is analogous to the representation of the remainder in the Taylor series of an analytic function. The rational functions do not depend on the integrand function and can be used as a general tool in treating uniform Airy-type expansions. The method is mainly of theoretical interest and delivers only order estimates for the remainders. In §8 we describe a method for obtaining strict error bounds for remainders of Airy-type expansions.

Our methods are not restricted to Airy-type expansions. In §7 we consider some other types of uniform expansions. In particular, a uniform expansion in terms of Bessel functions is considered. In this case the extension of the domain can be obtained, as can the double asymptotic property.

2. Related and earlier results. Airy-type expansions occur in the asymptotic theory of differential equations, for instance in turning-point problems; see [10, chap. 11]. In this case the estimation of remainders in terms of realistic and strict error bounds is well developed. Moreover, Olver extended the domains of the large parameter and the analogue of the saddle-point parameter to large areas in the complex plane.

The situation for integrals is quite different. Although the uniform Airy-type expansions have been extensively studied, a general theory for obtaining computable strict and realistic error bounds is still missing. This problem is more difficult than that for the case of differential equations. In transforming a given integral to a standard form by means of a mapping $x \mapsto w$ as in (1.2), a mapping $\alpha \mapsto b$ is implicitly introduced. Because of these two mappings, the function $h_0(w)$ in (1.3) may be difficult to handle. In corresponding problems in differential equations only the mapping $\alpha \mapsto b$ (or a related one) has to be considered. Another point is that in the theory of differential equations several techniques for bounding the remainders exist, but these techniques cannot be translated to the treatment of remainders of expansions obtained

through integrals. An example is Olver's method that is based on bounding the remainders by using Volterra integral equations.

In [3] the analytical properties of the mapping (1.2) are considered locally around the relevant saddle points; in Friedman [8] another proof is given. Levinson [9] gives a fundamental mapping theorem that generalizes the mapping (1.2) considerably; see also [19, §7.6]. In Qu and Wong [11] an iterative method is used for proving the local analytic property of mappings that are more complicated than those defined by (1.2) (there is a pole in the neighborhood of the coalescing saddle points). The transformation (1.2) is discussed in terms of conformal mappings on unbounded domains for special cases; for instance, in Copson [4] for an integral defining the Bessel function $J_\nu(z)$, in [10] for the Anger function (a function related to the Bessel function), and in [7] for integrals defining the Laguerre polynomials.

Recent examples of the construction of strict bounds in uniform asymptotic expansions of integrals are presented in Shivakumar and Wong [12] and in Frenzen [5] for Legendre-function expansions and in Frenzen and Wong [6] for Jacobi polynomials. The expansions are not of the compound type that follows from the Bleistein method, and a restricted number of terms in the expansions are considered. Another approach is given in Ursell [17] for Legendre functions, where uniform bounds are obtained by applying the maximum-modulus theorem. Ursell's method does not give sharp computable estimates of the remainders, and extension of the bounded domain of z to an unbounded domain is indicated without proof. In Ursell [18] the Airy-type expansion is discussed by using the maximum-modulus principle for complex values of the saddle-point parameter. The possibility of extending the validity to unbounded domains is mentioned again. Earlier, in Ursell [16], the Airy-type expansion is compared with the steepest-descent expansion, giving a continuation to unbounded domains. Qualitative results are obtained for the coefficient functions and the remainders; the Bleistein sequence is not used.

In the Anger function example in [10] the region of the saddle-point parameter is extended to an unbounded real domain by giving order estimates of the remainder. Olver's technique is based on estimating remainders of Taylor series. The expansion is not of the Bleistein type but is obtained by expanding the integrand function at a saddle point inside the interval of integration. The analysis shows that the distance between singular points of the integrand function and that saddle point plays a crucial role, although the singularities are not mentioned explicitly.

In the treatment of Laguerre polynomials in [7] order estimates for the remainders are also given, and there is a claim of uniform validity with respect to the saddle-point parameter in an unbounded real domain. The claim does not follow from investigating the singularities of the integrand function. In the present paper we take into account the singularities, and we show that the claim is indeed correct.

Soni and Soni [14] give new representations of the coefficients and remainders of Airy-type expansions; these representations are based on an expansion of the integrand function in terms of a class of polynomials. The paper is a continuation of earlier papers by Soni and Sleeman [13] and Soni and Temme [15]. The coefficients and remainder are written as contour integrals of the integrand function and rational functions related with the polynomials. New order estimates of the remainder have been derived for a finite domain of the saddle-point parameter.

3. Uniform Airy-type expansion. Let

$$F(z,b) = \frac{1}{2\pi i}\int_{\mathcal{L}} e^{z(\frac{1}{3}w^3 - b^2 w)} h_0(w)\,dw, \tag{3.1}$$

where $h_0(w)$ is an analytic function on a neighborhood of $\mathcal{L}$, with $\mathcal{L}$ a suitable contour that begins at $\infty\exp(-\frac{1}{3}\pi i)$ and ends at $\infty\exp(\frac{1}{3}\pi i)$. When $b \in [0,\infty)$ we take $\mathcal{L}$ the steepest-descent contour through b, which is given by $\mathcal{L} = \{w = x + iy \in \mathbb{C} \mid y^2 = 3x^2 - 3b^2\}$ (see Fig. 3.1), such that $\operatorname{Im}(\frac{1}{3}w^3 - b^2w) = 0$ and $\frac{1}{3}w^3 - b^2w$ attains its maximum on $\mathcal{L}$ at b.

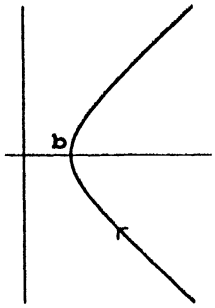


FIG. 3.1. *Steepest-descent curve $\mathcal{L}$ when $b \in [0,\infty)$.*

When $b \in [0, i\infty)$ we take $\mathcal{L} = \{w = x + iy \in \mathbb{C} \mid 3yx^2 = (y \pm ib)^2(y \mp 2ib)\}$, the steepest-descent contour through $\pm b$ (see Fig. 3.2).

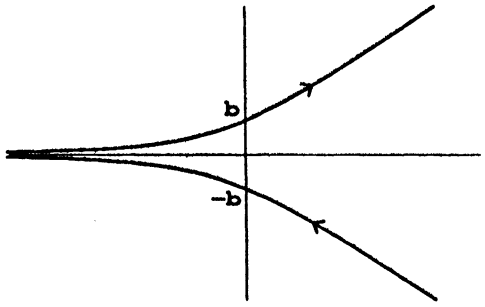


FIG. 3.2. *Steepest-descent curve $\mathcal{L}$ when $b \in [0, i\infty)$.*

It is not necessary to restrict our analysis to these contours of integration, but using these steepest-descent contours makes the following calculations less complicated.

We use Bleistein's method for obtaining an asymptotic expansion, defining $g_n(w)$, $h_{n+1}(w)$, $n = 0, 1, 2, \ldots$, by writing

$$\begin{aligned} h_n(w) &= \alpha_n + \beta_n w + (w^2 - b^2)g_n(w), \\ h_{n+1}(w) &= \frac{d}{dw}g_n(w), \end{aligned} \tag{3.2}$$

with α_n, β_n following from substitution of $w = \pm b$. If we use (3.2) in (3.1) and integrate n times by parts, we obtain

$$F(z,b) = \operatorname{Ai}(z^{2/3}b^2)\sum_{k=0}^{n-1}(-1)^k\alpha_k z^{-k-1/3} - \operatorname{Ai}'(z^{2/3}b^2)\sum_{k=0}^{n-1}(-1)^k\beta_k z^{-k-2/3} + \varepsilon_n, \tag{3.3}$$

where

$$\varepsilon_n = (-1)^n z^{-n}\frac{1}{2\pi i}\int_{\mathcal{L}} e^{z(\frac{1}{3}w^3 - b^2w)}h_n(w)dw \tag{3.4}$$

and where $\mathrm{Ai}(z)$ is the Airy function and $\mathrm{Ai}'(z)$ is its derivative. The functions $h_n(w)$ share, by inheritance, the analytic properties of h_0 on the same neighborhood of $\mathcal{L}$.

Estimates of $|\varepsilon_n|$, for large values of z and for $|b|$ bounded, given in the literature are usually of the form

$$|\varepsilon_n| \leq \frac{M_n}{z^{n+1/3}}\widetilde{\alpha}_n(b)|\widetilde{\mathrm{Ai}}(z^{2/3}b^2)| + \frac{N_n}{z^{n+2/3}}\widetilde{\beta}_n(b)|\widetilde{\mathrm{Ai}}'(z^{2/3}b^2)|, \tag{3.5}$$

where M_n and N_n depend on n and where $\widetilde{\alpha}_n$, $\widetilde{\beta}_n$ are related to the coefficients in (3.2). Furthermore,

$$\begin{aligned} \widetilde{\mathrm{Ai}}\,(u) &= \begin{cases} \mathrm{Ai}\,(u) & \text{if } u \geq 0, \\ \left[\mathrm{Ai}^2(u) + \mathrm{Bi}^2(u)\right]^{1/2} & \text{if } u < 0, \end{cases} \\ \widetilde{\mathrm{Ai}}'\,(u) &= \begin{cases} \mathrm{Ai}'(u) & \text{if } u \geq 0, \\ \left[\mathrm{Ai}'^2(u) + \mathrm{Bi}'^2(u)\right]^{1/2} & \text{if } u < 0. \end{cases} \end{aligned} \tag{3.6}$$

A proof of an estimate like (3.5) is given in [7], with

$$\widetilde{\alpha}_n(b) = \begin{cases} 1 & \text{if } 0 < b < \xi, \\ |\alpha_n| & \text{if } b > \xi, \end{cases} \qquad \widetilde{\beta}_n(b) = \begin{cases} 1 & \text{if } 0 < b < \xi, \\ |\beta_n| & \text{if } b > \xi, \end{cases}$$

where ξ is a fixed positive number.

Notice that the influence of large $|b|$ in (3.5) is not clear. We assume that the function h_0 of (3.1) depends on the saddle-point parameter b. Usually this is a consequence of the transformation to the standard form (3.1) by the mapping defined in (1.2). Also, when h_0 does not depend on b, all functions h_n obtained by recursion from (3.2) do depend on b.

For bounded $|b|$ an estimate like (3.5) holds for rather mild conditions on h_0. However, for obtaining uniformly valid estimates when b runs through an unbounded interval, we need more information on h_0. In the following sections we obtain estimates of $|h_n(w)|$ by formulating conditions on h_0 on discs with centers $\pm b$. These discs have radius $\rho(b)$, which indeed may be a function of b.

For obtaining estimates of ε_n of (3.4) holding in unbounded b-intervals, we now introduce a new class of rational functions.

4. Intermezzo: A new class of rational functions. We introduce a class of rational functions that satisfy the following theorem.

THEOREM 4.1. *Let*

$$R_0(u,w,b) = \frac{1}{u-w}, \tag{4.1.a}$$

$$R_{n+1}(u,w,b) = \frac{-1}{u^2-b^2}\frac{d}{du}R_n(u,w,b), \qquad n = 0,1,2,\ldots, \tag{4.1.b}$$

where $u,w,b \in \mathbb{C}$, $u \neq w$, $u^2 \neq b^2$. Let $h_n(w)$ be defined by the recursive scheme (3.2), *with $h_0(w)$ a given analytic function in a domain G. Then we have*

$$h_n(w) = \frac{1}{2\pi i}\int_C R_n(u,w,b)h_0(u)du, \tag{4.2}$$

where C is a simple closed contour in G that encircles the points w and $\pm b$.

Proof.

$$
\begin{aligned}
h_n(w) &= \frac{1}{2\pi i}\int_C R_0(u,w,b)h_n(u)du = \frac{1}{2\pi i}\int_C R_0(u,w,b)\frac{d}{du}g_{n-1}(u)du \\
&= \frac{1}{2\pi i}\int_C R_1(u,w,b)h_{n-1}(u)du - \frac{1}{2\pi i}\int_C R_1(u,w,b)(\alpha_{n-1}+\beta_{n-1}u)du \\
&=_* \frac{1}{2\pi i}\int_C R_1(u,w,b)h_{n-1}(u)du \\
&\vdots \\
&= \frac{1}{2\pi i}\int_C R_n(u,w,b)h_0(u)du.
\end{aligned}
$$

In * we use the fact that the rational function $R_1(u,w,b)(\alpha_{n-1}+\beta_{n-1}u)$ is $\mathcal{O}(u^{-2})$ as $|u|\to\infty$ and that all the poles of this function are inside C. Thus the integral of this function along C vanishes (use the transformation $u\mapsto u^{-1}$, which is well defined at $u=\infty$ and yields an integral with no singularities inside the contour of integration). □

COROLLARY 4.2. *Let $A_n(u,b)$, $B_n(u,b)$ be defined by the recursion in* (4.1.b), *with initial values*

$$
A_0(u,b) = \frac{u}{u^2-b^2}, \qquad B_0(u,b) = \frac{1}{u^2-b^2}. \tag{4.3}
$$

Then for $n=0,1,2,\ldots$, the coefficients α_n, β_n of (3.2) *can be written as*

$$
\alpha_n = \frac{1}{2\pi i}\int_C A_n(u,b)h_0(u)du, \qquad \beta_n = \frac{1}{2\pi i}\int_C B_n(u,b)h_0(u)du, \tag{4.4}
$$

where C is a simple closed contour in G that encircles the points $\pm b$.

We observe that the rational functions defined by (4.1) are independent of the function h_0 and that representation (4.2) can be considered as the analogue of the Cauchy integral defining the remainder of a Taylor series. An estimate of h_n, the integrand function of (3.4), will be obtained as in Cauchy's inequality for bounding the coefficients of a Taylor series.

By induction with respect to n, it follows that R_n has an expansion of the form

$$
R_n(u,w,b) = \sum_{i=0}^{n-1}\sum_{j=0}^{k_{n,i}} \frac{C_{ij}u^{i-j}}{(u-w)^{n+1-i-j}(u^2-b^2)^{n+i}}, \qquad n=1,2,\ldots, \tag{4.5}
$$

with $k_{n,i}=\min(i,n-1-i)$ and where C_{ij} do not depend on u, w, and b.

We conclude this section by giving estimates for R_n and for integrals of this function; these can be proved easily with (4.5).

(i) Let $w\in\mathbb{C}$ such that $|w-b|=\mathcal{O}(b)$ as $b\to\infty$, and let Γ be a simple closed contour that encircles b and w. Then for $n=1,2,\ldots,$

$$
\frac{1}{2\pi i}\int_\Gamma R_n(u,w,b)du = \mathcal{O}(b^{-3n}) \tag{4.6}
$$

as $b\to\infty$.

(ii) Let $b \in \mathbb{C}$ and $\Omega(b) = \{(u,w) \in \mathbb{C}^2 \mid |u-b| = \rho(b),\ |w-b| \le \frac{1}{2}\rho(b)\}$, such that $\rho(b) = \mathcal{O}(|b|^\theta)$ as $b \to \infty$, where $-\frac{1}{2} < \theta \le 1$. Then we can assign numbers A_n independent of b, such that

$$\sup_{(u,w)\in\Omega(b)} |R_n(u,w,b)| \le A_n |b|^{-(1+2\theta)n-\theta} \quad \text{as} \quad b \to \infty. \tag{4.7}$$

5. Extension of the domain of validity. In this section we prove that, under certain circumstances, expansion (3.3) holds uniformly with respect to the saddle-point parameter b in unbounded domains.

For defining the radius $\rho(b)$ of the discs mentioned at the end of §3, we first define

$$\rho_0(b) = \min\{|w \pm b| \mid w \text{ is a singularity of } h_0(w)\} \tag{5.1}$$

and we assume that, for large $|b|$, we have $\rho_0(b) \ge \delta_0 |b|^\theta$, where the constants δ_0 and θ satisfy $\delta_0 > 0$, $\theta > -\frac{1}{2}$. This is the essential assumption on $h_0(w)$ in the neighborhood of the saddle points.

Now we take $\rho(b) \le \rho_0(b)$ such that $\rho(b) \sim \delta |b|^\theta$ as $b \to \infty$, where the constant $\delta > 0$. We take $\theta \le 1$ as large as possible, and we drop the restriction $\theta \le 1$ after Theorem 5.2. Notice that we concentrate on estimates with $|b| \to \infty$ and that we do not give details for b in compacta.

Next we introduce upper bounds for the $h_n(w)$, $n = 0, 1, 2, \ldots$. Thus let

$$\widetilde{h}_n = \sup_{|w\pm b| \le (1/2)\rho(b)} |h_n(w)|. \tag{5.2}$$

Notice that $h_0(w)$ is analytic on $|w \pm b| < \rho(b)$; thus $\widetilde{h}_0$ is finite.

For obtaining estimates of $\widetilde{h}_n$ in terms of $\widetilde{h}_0$ let Γ be a circle around $\pm b$ with radius $\rho(b)$ and let $|w \mp b| \le \frac{1}{2}\rho(b)$. We require $\theta \le 1$ to ensure that both saddle points are not inside the circle Γ. This is possible by choosing δ appropriately. Then, if we use (4.6), we have

$$\begin{aligned}
h_n(w) &= \frac{1}{2\pi i}\int_\Gamma R_0(u,w,b) h_n(u)\,du \\
&= \frac{1}{2\pi i}\int_\Gamma R_1(u,w,b) h_{n-1}(u)\,du - \frac{1}{2\pi i}\int_\Gamma R_1(u,w,b)(\alpha_{n-1}+\beta_{n-1}u)\,du \\
&=_{(4.6)} \frac{1}{2\pi i}\int_\Gamma R_1(u,w,b) h_{n-1}(u)\,du + \widetilde{h}_{n-1}\mathcal{O}(b^{-3}) \\
&\ \ \vdots \\
&=_{(4.6)} \frac{1}{2\pi i}\int_\Gamma R_n(u,w,b) h_0(u)\,du + \widetilde{h}_{n-1}\mathcal{O}(b^{-3}) + \cdots + \widetilde{h}_0\mathcal{O}(b^{-3n})
\end{aligned}$$

as $b \to \infty$. So by induction and (4.7) we have proved the following theorem.

THEOREM 5.1. *Let $\widetilde{h}_n$, $n = 0, 1, 2, \ldots$, be the upper bound of $h_n(w)$, defined in* (5.2). *Then we have the estimate*

$$\widetilde{h}_n \le C_n |b|^{-(1+2\theta)n} \widetilde{h}_0 \quad \text{as} \quad b \to \infty, \tag{5.3}$$

where C_n does not depend on b.

Now we shall prove that ε_n can be bounded as follows:

$$|\varepsilon_n| \le C_n(|b|+1)^{-(1+2\theta)n}\widetilde{h}_0 z^{-n-1/3}\widetilde{\mathrm{Ai}}(z^{2/3}b^2), \tag{5.4}$$

with a slightly different C_n that does not depend on b and z.

In order to use the preceding estimates, we split up the contour $\mathcal{L}$ into $\mathcal{L}'$ and $\mathcal{L}''$. In the case that $b \in [0,\infty)$ we take $\mathcal{L}' = \{w \in \mathcal{L} \mid |w-b| \le \frac{1}{2}\rho(b)\}$, and in the case that $b \in [0,i\infty)$ we take $\mathcal{L}' = \{w \in \mathcal{L} \mid |w \pm b| \le \frac{1}{2}\rho(b)\}$. We define $\mathcal{L}'' = \mathcal{L} - \mathcal{L}'$ and introduce the corresponding integrals:

$$\begin{aligned}
\varepsilon_{n|_{\mathcal{L}'}} &= (-1)^n z^{-n}\frac{1}{2\pi i}\int_{\mathcal{L}'} e^{z(\frac{1}{3}w^3-b^2w)}h_n(w)dw,\\
\varepsilon_{n|_{\mathcal{L}''}} &= (-1)^n z^{-n}\frac{1}{2\pi i}\int_{\mathcal{L}''} e^{z(\frac{1}{3}w^3-b^2w)}h_n(w)dw.
\end{aligned} \tag{5.5}$$

In the Appendix we formulate conditions on $h_0(w)$ such that when $\theta > -\frac{1}{2}$ the estimate of $|\varepsilon_{n|_{\mathcal{L}''}}|$ is exponentially small compared with the estimate of $|\varepsilon_{n|_{\mathcal{L}'}}|$ as $z \to \infty$ uniformly with respect to b.

The proof of (5.4) for large b is divided into separate cases: (i) $b \in [0,\infty)$ and (ii) $b \in [0,i\infty)$. We first consider case (i). With (5.3) we have

$$\begin{aligned}
|\varepsilon_{n|_{\mathcal{L}'}}| &\le C_n z^{-n}|b|^{-(1+2\theta)n}\widetilde{h}_0\frac{1}{2\pi i}\int_{\mathcal{L}'} e^{z(\frac{1}{3}w^3-b^2w)}dw\\
&\le C_n z^{-n-1/3}|b|^{-(1+2\theta)n}\widetilde{h}_0\mathrm{Ai}(z^{2/3}b^2).
\end{aligned}$$

In case (ii) we write $w = x+iy$ and we define $\mathcal{L}'_+ = \{y > 0 \mid \text{there exists } x \in \mathbb{R} : x+iy \in \mathcal{L}'\}$. Simple transformations give

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{\mathcal{L}'} e^{z(\frac{1}{3}w^3-b^2w)}h_n(w)dw =\\
&\quad\frac{1}{2\pi i}\int_{\mathcal{L}'_+} e^{-z(y+ib)^2f(y)}g(y)(e^{-\frac{2}{3}zb^3}h_n(w) - e^{+\frac{2}{3}zb^3}h_n(\overline{w}))dy\\
&\quad+\frac{1}{2\pi}\int_{\mathcal{L}'_+} e^{-z(y+ib)^2f(y)}(e^{-\frac{2}{3}zb^3}h_n(w) + e^{+\frac{2}{3}zb^3}h_n(\overline{w}))dy,
\end{aligned}$$

where

$$f(y) = \frac{2(2y-ib)^2}{9y}\sqrt{\frac{y-2ib}{3y}}, \qquad g(y) = \sqrt{\frac{y-2ib}{3y}} + \frac{ib(y+ib)}{3y^2}\sqrt{\frac{3y}{y-2ib}}.$$

Note that the functions have real arguments and that $g(y) \ge 0$. Thus with (5.3) we have

$$\begin{aligned}
&\left|\frac{1}{2\pi i}\int_{\mathcal{L}'} e^{z(\frac{1}{3}w^3-b^2w)}h_n(w)dw\right|\\
&\le \frac{1}{2\pi}\int_{\mathcal{L}'_+} e^{-z(y+ib)^2f(y)}(1+g(y))(|h_n(w)| + |h_n(\overline{w})|)dy\\
&\le_{(5.3)} C_n|b|^{-(1+2\theta)n}\widetilde{h}_0\frac{1}{\pi}\int_0^\infty e^{-z(y+ib)^2f(y)}(1+g(y))dy\\
&\le C'_n|b|^{-(1+2\theta)n}\widetilde{h}_0\frac{1}{\sqrt{zb/i}}\\
&\sim_* \pi^{1/2}C'_n|b|^{-(1+2\theta)n}\widetilde{h}_0 z^{-1/3}\widetilde{\mathrm{Ai}}(z^{2/3}b^2),
\end{aligned}$$

as $z \to \infty$. In * we have used the relation $\widetilde{\mathrm{Ai}}(x) \sim \pi^{-1/2}(-x)^{-1/4}$ as $x \to -\infty$; see [10, p. 395].

In the Appendix we prove that

$$|\varepsilon_{n|_{\mathcal{L}''}}| \leq C_n e^{-\lambda(z-\mu)|b|^{2\theta+1}} \widetilde{h}_0 z^{-n-4/3} \widetilde{\mathrm{Ai}}(z^{2/3}b^2), \tag{5.6}$$

where the positive C_n, λ, and μ do not depend on b and z and where $|b| \geq c > 0$. These estimates show that (5.4) is valid. Thus we have proved the following theorem.

THEOREM 5.2. *Let $F(z,b)$ be of the form (3.1), where $h_0(w)$ satisfies the conditions mentioned in the beginning of this section and in the Appendix. Then we have (3.3) as a uniform asymptotic expansion for $F(z,b)$, where (5.4) is an estimate for $|\varepsilon_n|$ as $z \to \infty$ uniformly with respect to $b \in [0,\infty) \cup [0, i\infty)$ and where $\widetilde{h}_0$ is given in (5.2).*

Now we drop the restriction $\theta \leq 1$. In the case that $\theta > 1$, the analysis that leads to Theorem 5.2 is much easier; every time $1 + 2\theta$ occurs it can be replaced with the larger factor 3θ.

Remark 1. With the conditions of Theorem 5.2 it follows that expansion (3.3) has a double asymptotic property: the roles of b and z can be interchanged. The double asymptotic property is lost in the example considered in §6.

Remark 2. An estimate like (5.4) has been derived in [10, p. 360] for a particular example. There the estimate for the remainder of an expansion of the Anger function $\mathrm{A}_{-\nu}(\nu \operatorname{sech} \alpha)$ reads

$$\varepsilon_n(\alpha,\nu) = (1+\xi)^{-\theta(n+1)} \nu^{-\frac{1}{3}(n+1)} \mathrm{Qi}_n(\nu^{\frac{2}{3}}\xi) \mathcal{O}(1),$$

where $\mathrm{Qi}_n(z)$ is a special function, $\frac{2}{3}\xi^{3/2} = \alpha - \tanh\alpha$, and $\theta = -\frac{1}{4}$. This estimate holds as $\nu \to \infty$ uniformly with respect to $\alpha \in [0,\infty)$ or $\xi \in [0,\infty)$. Indeed, the value $\theta = -\frac{1}{4}$ is related to the distance between the relevant saddle point and the nearest singularities of the integrand function, which is of order $\xi^{-1/4}$ as $\xi \to \infty$.

6. Laguerre polynomials: A boundary case. In this section we show that, in certain circumstances, the condition $\theta > -\frac{1}{2}$ of Theorem 5.2 can be replaced with $\theta = -\frac{1}{2}$. We demonstrate this feature by considering a recent expansion for the Laguerre polynomials.

First we summarize the main steps for obtaining an Airy-type expansion of the Laguerre polynomials. More details are given in [7] and [19]. Laguerre polynomials have the following integral representation:

$$(-1)^N 2^\alpha e^{-zt/2} L_N^{(\alpha)}(zt) = \frac{1}{2\pi i} \int_{+\infty}^{(1+)} e^{zf(x,t)} (1-x^2)^{\frac{\alpha-1}{2}} dx, \tag{6.1}$$

where the contour of integration begins and ends at $+\infty$ and encircles 1 in the positive direction and where

$$f(x,t) = \frac{1}{4}\ln\left(\frac{1+x}{1-x}\right) - \frac{1}{2}xt \tag{6.2}$$

and $z = 4N + 2\alpha + 2$, $\alpha > -1$, and $t \geq 1$. Again, we use the transformation

$$f(x,t) = \tfrac{1}{3}w^3 - b^2 w. \tag{6.3}$$

The x-saddle points $\pm\sqrt{1-1/t}$ should correspond with $\pm b$. It follows that

$$b^3 = \tfrac{3}{4}\left(\sqrt{t^2-t} - \operatorname{arccosh}\sqrt{t}\right). \tag{6.4}$$

With transformation (6.3) we have for (6.1)

$$(-1)^N 2^\alpha e^{-zt/2} L_N^{(\alpha)}(zt) = \frac{1}{2\pi i}\int_{\mathcal{L}} e^{z(\frac{1}{3}w^3 - b^2 w)} h_0(w)\,dw, \tag{6.5}$$

where

$$h_0(w) = (1-x^2)^{\frac{\alpha-1}{2}}\frac{dx}{dw} = 2\frac{(1-x^2)^{\frac{\alpha+1}{2}}(w^2-b^2)}{1-t(1-x^2)} \tag{6.6}$$

and $\mathcal{L}$ is given in Fig. 3.1. Again, using (3.2) in (6.5), we obtain

$$\begin{aligned}(-1)^N 2^\alpha e^{-zt/2} L_N^{(\alpha)}(zt) = {}& \operatorname{Ai}(z^{2/3}b^2)\sum_{k=0}^{n-1}(-1)^k \alpha_k z^{-k-1/3} \\ & - \operatorname{Ai}'(z^{2/3}b^2)\sum_{k=0}^{n-1}(-1)^k \beta_k z^{-k-2/3} + \varepsilon_n,\end{aligned} \tag{6.7}$$

where ε_n is as in (3.4).

To apply the analysis of §5, we locate the relevant singular points of $h_0(w)$. Let $x_0 = \sqrt{1-1/t}$ be the positive x-saddle point when $t > 1$. The point x_0 is mapped to $w(x_0) = b$ by the mapping given in (6.3), when the logarithmic function takes its principal value. However, the points x_0 at other sheets of the Riemann surface of the log function are singular points of the mapping (6.3). Then the phase of $1 - x_0$ is, for instance, 2π. When $b = 0$ the singularities $w = S_\pm$ nearest to b satisfy $\frac{1}{3}S_\pm^3 = \pm\frac{1}{2}\pi i$, whereas

$$S_\pm \sim b \pm \sqrt{\frac{\pi i}{2b}} \quad \text{as} \quad b \to \infty. \tag{6.8}$$

Thus $\rho_0(b)$ of (5.1) is of order $b^{-1/2}$ as $b \to \infty$.

As before, we want to split up $\mathcal{L}$ into $\mathcal{L}'$ and $\mathcal{L}''$, and define $\varepsilon_{n|_{\mathcal{L}'}}$, $\varepsilon_{n|_{\mathcal{L}''}}$ similar to (5.5). So define $\mathcal{L}' = \{w \in \mathcal{L} \mid |w-b| \le \delta b^\theta\}$, where the constants δ and θ satisfy $\delta > 0$ and $-\frac{1}{2} < \theta \le 1$, in order that the estimate of $|\varepsilon_{n|_{\mathcal{L}''}}|$ is exponentially small compared with the estimate of $|\varepsilon_{n|_{\mathcal{L}'}}|$ as $z \to \infty$ uniformly with respect to b. We choose θ close to $-\frac{1}{2}$ fixed.

Let Γ_θ be a closed contour which encircles $\mathcal{L}'$ such that

$$\text{length } \Gamma_\theta = \mathcal{O}(b^\theta), \quad \text{distance}(\Gamma_\theta, \mathcal{L}') \sim cb^{-1/2} \quad \text{as} \quad b \to \infty$$

and such that $h_0(w)$ is analytic on $\overline{I(\Gamma_\theta)}$, where $\overline{I(\Gamma_\theta)}$ is the closure of the interior of Γ_θ. Then straightforward calculations give that

$$\sup_{w \in \overline{I(\Gamma_\theta)}} |h_0(w)| \le C_0 b^{(\theta+\frac{1}{2})\alpha}|h_0(b)| \quad \text{as} \quad b \to \infty, \tag{6.9}$$

where C_0 does not depend on b. Further calculations, similar to those in §5 yield, for $n = 1, 2, \ldots$,

$$\sup_{w \in \overline{I(\Gamma_\theta)}} |h_n(w)| \leq C_n b^{(\theta+\frac{1}{2})(\alpha+1)} |h_0(b)| \quad \text{as} \quad b \to \infty, \tag{6.10}$$

where, here and below, C_n denotes a generic quantity that does not depend on b and z. Notice that, in contrast to (5.3), the power of b is positive and does not depend on n. These estimates yield

$$|\varepsilon_{n|_{\mathcal{L}'}}| \leq C_n z^{-n-1/3} b^{(\theta-1)\alpha+(\theta-\frac{1}{2})} \mathrm{Ai}(z^{2/3} b^2). \tag{6.11}$$

In (6.11) we used

$$h_0(b) = t^{\frac{(1-\alpha)}{2}} \frac{\sqrt{2b}}{(t-1)^{1/4} t^{3/4}}.$$

In the Appendix we prove that

$$|\varepsilon_{n|_{\mathcal{L}''}}| \leq C_n' z^{-n-4/3} e^{-\lambda(z-2)b^{2\theta+1}} |h_0(b)| \mathrm{Ai}(z^{2/3} b^2), \tag{6.12}$$

where the positive C_n' and λ do not depend on b and z. It follows that we can assign numbers C_n, independent of z and b, such that

$$|\varepsilon_n| \leq C_n z^{-n-1/3} (b+1)^{(\theta-1)\alpha+(\theta-\frac{1}{2})} \mathrm{Ai}(z^{2/3} b^2), \tag{6.13}$$

as $z \to \infty$ uniformly with respect to $b \in [0, \infty)$. A similar approach can be used for $b \in [0, i\tau]$, where $0 < \tau < (\frac{3}{8}\pi)^{1/3}$, τ fixed.

We can compare this estimate with the estimate given in [7] and [19], which is of the form (3.5). First, we notice that (6.13) is not in the form of the first neglected terms of expansion (6.7). But with (6.10) it easily follows that the first neglected terms can be estimated by the right-hand side of (6.13). Regardless, (6.13) clearly shows why expansion (6.7) holds uniformly with respect to b in an unbounded domain. Secondly, in (6.13) the influence of b is more transparent than in the right-hand side of (3.5).

7. Other uniform expansions generated by the Bleistein method. In this section we show that the methods used for the Airy-type expansions are quite general and can be applied to other uniform expansions of integrals of the form

$$\int_C e^{zf(x,b)} h_0(x) dx \tag{7.1}$$

with coinciding saddle points and singularities. In this section we work out an example of uniform expansions in terms of Bessel functions. In [7] such an expansion of the Laguerre polynomials is given. Let

$$F(z, A) = \frac{1}{2\pi i} \int_{-\infty}^{(0+)} w^{-\alpha-1} h_0(w) e^{\frac{1}{2}z(w-A^2/w)} dw, \tag{7.2}$$

where the contour of integration begins and ends at $-\infty$ and encircles the origin in the positive direction. We assume that $h_0(w)$ is analytic on a neighborhood of the contour of integration, and we let $z > 0$, $iA > 0$, and $\alpha > -1$. Notice that $\pm iA$ are

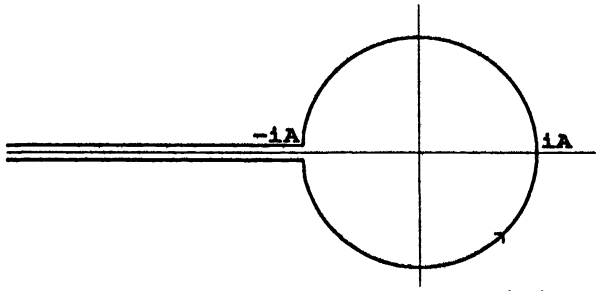


FIG. 7.1. *Steepest-descent curve for integral* (7.2).

the saddle points of the integral. We choose the contour of integration through these saddle points, and the steepest-descent path looks like Fig. 7.1.

The recursion in connection with integral (7.2) is

(7.3)
$$h_n(w) = \alpha_n + \frac{\beta_n}{w} + \left(1 + \frac{A^2}{w^2}\right) g_n(w), \qquad h_{n+1}(w) = w^{\alpha+1} \frac{d}{dw} \left[w^{-\alpha-1} g_n(w)\right],$$

and if we integrate n times by parts, we obtain the expansion

$$F(z, A) = \frac{J_\alpha(zA)}{A^\alpha} \sum_{k=0}^{n-1} (-1)^k \alpha_k \left(\frac{2}{z}\right)^k + \frac{J_{\alpha+1}(zA)}{A^{\alpha+1}} \sum_{k=0}^{n-1} (-1)^k \beta_k \left(\frac{2}{z}\right)^k + \varepsilon_n, \tag{7.4}$$

where

$$\varepsilon_n = (-1)^n \left(\frac{2}{z}\right)^n \frac{1}{2\pi i} \int_{-\infty}^{(0+)} w^{-\alpha-1} h_n(w) e^{\frac{1}{2} z(w - A^2/w)} dw \tag{7.5}$$

and where $J_\alpha(z)$ and $J_{\alpha+1}(z)$ are Bessel functions of the first kind. Since zA is purely imaginary, modified Bessel functions occur in the expansion.

The class of rational functions generated by (7.3) is recursively defined by

$$\begin{aligned} Q_0(u, w, A) &= \frac{1}{u - w}, \\ Q_{n+1}(u, w, A) &= \frac{-1}{1 + A^2/u^2} \left(\frac{\alpha + 1}{u} + \frac{d}{du}\right) Q_n, \qquad n = 0, 1, 2, \ldots. \end{aligned} \tag{7.6}$$

By induction with respect to n it follows that Q_n has an expansion of the form

$$Q_n(u, w, A) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-i} \frac{C_{ij} (A^2/u^2)^i}{(u - w)^{n+1-i-j} u^{i+j} (1 + A^2/u^2)^{n+i}}, \qquad n = 1, 2, \ldots, \tag{7.7}$$

where the C_{ij} do not depend on u, w, and A.

Again, we concentrate on the influence of A on the expansion (7.4), especially when $|A|$ is large.

If Γ is a simple closed contour that encircles iA and w and with $-iA$ in its exterior, then we can prove, just as for (4.6), that

$$\frac{1}{2\pi i} \int_\Gamma Q_n(u, w, A) du = \mathcal{O}(|A|^{-n}) \quad \text{as} \quad |A| \to \infty. \tag{7.8}$$

As before, we want to split up $\mathcal{L}$ into $\mathcal{L}'$ and $\mathcal{L}''$. We assume that, for large $|A|$, the distance from the singularities of $h_0(w)$ to the saddle points $\pm iA$ is at least $\delta|A|^\theta$, where the constants δ, θ satisfy $\delta > 0$, $\frac{1}{2} < \theta \leq 1$. Consequently, we take $\mathcal{L}' = \{w \in \mathcal{L} \mid |w - iA| \leq \frac{1}{2}\delta|A|^\theta\}$ and $\mathcal{L}'' = \mathcal{L} - \mathcal{L}'$ such that the estimate of $|\varepsilon_{n|_{\mathcal{L}''}}|$ is exponentially small compared with the estimate of $|\varepsilon_{n|_{\mathcal{L}'}}|$ as $z \to \infty$ uniformly with respect to $iA \in [c, \infty)$, where $c > 0$ fixed. In fact, we need a growth condition on $h_0(w)$ on a prescribed neighborhood of $\mathcal{L}''$, which is similar to the condition mentioned in the Appendix.

If we set $\Omega(A) = \{(u, w) \in \mathbb{C}^2 \mid |u - iA| = \frac{3}{4}\delta|A|^\theta,\ |w - iA| \leq \frac{1}{2}\delta|A|^\theta\}$, we can prove

$$\sup_{(u,w)\in\Omega(A)} |Q_n(u, w, A)| \leq C_n|A|^{(1-2\theta)n-\theta}, \tag{7.9}$$

where C_n does not depend on A. Finally, we define

$$\widetilde{h}_0 = \sup_{|w\pm iA|\leq(1/2)\delta|A|^\theta} |h_0(w)|. \tag{7.10}$$

With (7.8), (7.9), and straightforward calculations similar to those leading to (5.3), we obtain, for $n = 1, 2, \ldots$,

$$\sup_{|w-iA|\leq(1/2)\delta|A|^\theta} |h_n(w)| \leq C_n|A|^{(1-2\theta)n}\widetilde{h}_0 \quad \text{as} \quad |A| \to \infty, \tag{7.11}$$

where C_n does not depend on A. With the aid of these estimates we obtain as the main result of this section

$$|\varepsilon_n| \leq C_n(|A| + 1)^{(1-2\theta)n-\alpha}\widetilde{h}_0 z^{-n}|J_\alpha(zA)| \tag{7.12}$$

as $z \to \infty$ uniformly with respect to $iA \in [0, \infty)$, where C_n does not depend on A and z.

In the case that $\theta > 1$ we can use the same analysis that leads to (7.12), but every time $1 - 2\theta$ occurs it has to be replaced with $-\theta$.

A similar approach is possible for real values of A.

8. Strict upper bounds of the remainder. In this section we assume that we have quantitative information on the functions $h_n(w)$ and that we can construct upper bounds for the remainders ε_n of (3.4). The simplest case is that we know that $|h_n(w)|$ is bounded on $\mathcal{L}$. If $b \geq 0$, an upper bound for ε_n can be easily expressed in terms of this bound and of the Airy function $\mathrm{Ai}(z^{2/3}b^2)$. When $b^2 < 0$ (the oscillatory case), the bound can be expressed in terms of the modulus function $\left[\mathrm{Ai}^2(z^{2/3}b^2) + \mathrm{Bi}^2(z^{2/3}b^2)\right]^{1/2}$ (see also (3.6)).

When the maximal value of $|h_n(w)|$ occurs at $w = w_0$, with w_0 far away from the saddle point $w = b$, the upper bound obtained in this way may be quite inaccurate. The fact is that the main contributions to the integral (3.4) come from a small neighborhood of b, especially when z is large. To obtain realistic upper bounds of $|\varepsilon_n|$ we describe a different approach in which we also allow unbounded functions $h_n(w)$. We concentrate on the case $b \geq 0$.

The contour $\mathcal{L}$ can be parameterized by writing $w = x + iy$, $3x^2 - y^2 = 3b^2$. By using this and integrating with respect to y, the integral (3.1) can be written in the form

$$\varepsilon_n = (-1)^n z^{-n} \frac{e^{-\frac{2}{3}zb^3}}{2\pi i} \int_{-\infty}^{\infty} e^{-z\phi(y)} H_n(y)\, dy, \tag{8.1}$$

where

$$\phi(y) = (\tfrac{8}{9}y^2 + \tfrac{2}{3}b^2)\sqrt{\tfrac{1}{3}y^2 + b^2} - \tfrac{2}{3}b^3, \qquad H_n(y) = h_n(x+iy)\left[\tfrac{dx}{dy} + i\right],$$

and

$$\frac{dx}{dy} = \frac{\frac{1}{3}y}{\sqrt{\frac{1}{3}y^2 + b^2}}.$$

When $h_0 = 1$ and $n = 0$ we obtain the real representation for the Airy function:

$$\mathrm{Ai}(z^{\frac{2}{3}}b^2) = \frac{z^{\frac{1}{3}}e^{-\frac{2}{3}zb^3}}{2\pi}\int_{-\infty}^{\infty} e^{-z\phi(y)}\,dy. \tag{8.2}$$

To bound ε_n we assume that for fixed b the function $H_n(y)$ is majorized by

$$|H_n(y)| \le M_n e^{\sigma_n \phi(y)}, \qquad -\infty < y < \infty, \tag{8.3}$$

where M_n and σ_n are nonnegative numbers that may depend on b. Observe that, in fact, only the even part of the function $H_n(y)$ needs to be bounded in this way; when $h_n(w)$ is a real function, the even part of $H_n(y)$ equals the imaginary part. The best strategy is to start with M_n and to define it slightly larger than $|H_n(0)| = |h_n(b)|$ (when this quantity vanishes a minor modification is needed), say, $M_n = 1.25|h_n(b)|$. Next we determine the smallest number σ_n that satisfies the upper bound in (8.3). When $|h_n(w)|$ is bounded on $\mathcal{L}$ and assumes its maximal value on $\mathcal{L}$ at $w = w_0 = x_0 + iy_0$, one may take $M_n = |H_n(y_0)|$ and $\sigma_n = 0$. However, as mentioned previously, when y_0 is not close to zero, the resulting bound may be unrealistic. When $\sigma_n > 0$, the argument of the exponential function in the right-hand side of (8.3) is unbounded; thus we accept unbounded functions $|h_n(w)|$. Observe that far away from the origin the estimate (8.3) may be very rough, but there the contribution to the integral (8.1) is negligible, especially when z is large.

Using (8.3) in (8.1) (when $z > \sigma_n$), we obtain with (8.2) the estimate

$$|\varepsilon_n| \le M_n z^{-n}(z-\sigma_n)^{-\frac{1}{3}}\,\mathrm{Ai}\left((z-\sigma_n)^{\frac{2}{3}}b^2\right)e^{-\frac{2}{3}b^3\sigma_n}, \qquad z > \sigma_n, \quad b \ge 0. \tag{8.4}$$

The factor M_n contains the information on the parameter b; especially, it contains the information on whether or not the expansion holds uniformly on unbounded b-domains and has the double asymptotic property.

This bound is computable when the function $h_n(w), w \in \mathcal{L}$ is computable. Representation (4.2) may be helpful in computing $h_n(w)$. We expect that the bound in (8.4) is realistic for a wide class of functions $h_0(w)$.

When the function $h_n(w)$ grows too fast with b, the number σ_n may be an unbounded function of b. In that case the bound in (8.4) loses its uniform character. For example, when $h_0(w) = \exp(-w^2b^2)$ it is easily verified that the minimal value of σ_0 that satisfies (8.3) is $\sigma_0 = (2/243)b$.

When $b \in [0, i\infty)$ a similar approach is possible by majorizing the function $h_n(w)$ on the contour of Fig. 3.2. The analysis and the resulting bounds are slightly more complicated. Details will not be given.

9. An example. We consider the function

$$F(z,b) = \frac{1}{2\pi i}\int_{\mathcal{L}} e^{z(\frac{1}{3}w^3 - b^2w)}\frac{1}{w-b-1}dw, \tag{9.1}$$

where $b \in [0, \infty)$ and $\mathcal{L}$ is the steepest-descent contour shown in Fig. 3.1. $F(z,b)$ can be written as an integral of the Airy function, that is,

$$F(z,b) = e^{(b+1)\zeta}\left[-e^{\frac{1}{3}z(b+1)^3} + z^{-\frac{1}{3}}\int_{\zeta}^{\infty} e^{(b+1)t}\mathrm{Ai}(tz^{-\frac{1}{3}})\,dt\right], \qquad \zeta = zb^2.$$

In this example we have $h_0(w,b) = 1/(w-b-1)$. Thus the quantities introduced in (5.1) and (5.2) are as follows: $\rho_0(b) = 1$, $\theta = 0$, and $\widetilde{h}_0 = 2$. It is easily verified that

$$h_1(w) = \frac{-1}{(2b+1)(w-b-1)^2}, \qquad h_1(b) = \frac{-1}{2b+1},$$
$$h_2(w) = 2\frac{b^2+4b+2-(b+1)w}{(2b+1)^3(w-b-1)^3}, \qquad h_2(b) = -2\frac{3b+2}{(2b+1)^3}.$$

Further calculations show that

$$\begin{aligned} F(z,b) &= \mathrm{Ai}(z^{2/3}b^2)\alpha_0 z^{-1/3} - \mathrm{Ai}'(z^{2/3}b^2)\beta_0 z^{-2/3} + \varepsilon_1 \\ &= \mathrm{Ai}(z^{2/3}b^2)\left(\alpha_0 - \frac{\alpha_1}{z}\right)z^{-1/3} - \mathrm{Ai}'(z^{2/3}b^2)\left(\beta_0 - \frac{\beta_1}{z}\right)z^{-2/3} + \varepsilon_2, \end{aligned} \tag{9.2}$$

with

$$\begin{aligned} \alpha_0 &= -\frac{b+1}{2b+1}, & \beta_0 &= -\frac{1}{2b+1}, \\ \alpha_1 &= -\frac{2b^2+2b+1}{(2b+1)^3}, & \beta_1 &= -2\frac{b^2+b}{(2b+1)^3}, \\ \alpha_2 &= -4\frac{3b^3+5b^2+3b+1}{(2b+1)^5}, & \beta_2 &= -2\frac{6b^2+10b+5}{(2b+1)^5}. \end{aligned} \tag{9.3}$$

We can determine the numbers M_n, σ_n occurring in (8.3), but already for this simple example optimal values have to be computed numerically. Analytical bounds of $\operatorname{Im} H_n(y)$ are easily obtained, however. For example, we have (recall that $x = \sqrt{(1/3)y^2 + b^2}$)

$$\operatorname{Im} H_0(y) = \operatorname{Im}\left[\left(i + \frac{dx}{dy}\right)\frac{1}{x+iy-b-1}\right] = \frac{b^2-(b+1)x}{x[(x-b-1)^2+3x^2-3b^2]}, \qquad x \geq b$$

(changing to x gives better formulas). When $b \geq \frac{1}{2}$ we have $|\operatorname{Im} H_0(y)| \leq |h_0(b)|$; when $b \in [0, \frac{1}{2})$ the maximal value of $|\operatorname{Im} H_0(y)|$ is slightly larger than $|h_0(b)|$. Similar results hold for $n = 1, 2$, where the critical b-values are $b = \frac{1}{3}, b = (\sqrt{7}-1)/6 = 0.27$, respectively. It follows that in this example the remainders can be estimated in terms of the first neglected terms of the asymptotic expansion (note that $h_n(b) = \alpha_n + b\beta_n$):

$$\begin{aligned} |\varepsilon_1| &\leq |h_1(b)|z^{-4/3}\mathrm{Ai}(z^{2/3}b^2), & b &\geq \tfrac{1}{3}, \\ |\varepsilon_2| &\leq |h_2(b)|z^{-7/3}\mathrm{Ai}(z^{2/3}b^2), & b &\geq \tfrac{\sqrt{7}-1}{6}. \end{aligned} \tag{9.4}$$

These estimates may be compared with the order estimates (5.4) obtained from less qualitative information on the functions $h_n(w)$.

Appendix. We formulate conditions on $h_0(w)$ such that the estimate of $|\varepsilon_{n|_{\mathcal{L}''}}|$ is exponentially small compared with the estimate of $|\varepsilon_{n|_{\mathcal{L}'}}|$ as $z \to \infty$ uniformly with respect to b. We take $\mathcal{L}$, $\mathcal{L}'$, $\mathcal{L}''$, $\rho(b)$, δ, θ, $\varepsilon_{n|_{\mathcal{L}'}}$, $\varepsilon_{n|_{\mathcal{L}''}}$, and $\widetilde{h}_n$ as in §5. Define

$$\mathcal{R}(w,b,p,q,r) = r|w^{-q}e^{p(\frac{1}{3}w^3 - b^2w + \frac{2}{3}b^3)}|.$$

We assume that $h_0(w)$ is an analytic function on a neighborhood $\Omega_0(b)$ of $\mathcal{L}''$, such that for every $w \in \mathcal{L}''$ a disc with center w and radius $\mathcal{R}$ is contained in $\Omega_0(b)$, where $r > 0$ and $p,\ q \geq 0$ do not depend on b and w. Note that, since $w \in \mathcal{L}''$, $\mathcal{R}$ may be exponentially small as $|w| \to \infty$. Furthermore, we assume that there are constants $\sigma \geq 0$ and $C_0 > 0$ such that

$$|h_0(w)| \leq C_0 \widetilde{h}_0 |e^{-\sigma(\frac{1}{3}w^3 - b^2 w + \frac{2}{3}b^3)}| \quad \forall w \in \Omega_0(b) \cup \mathcal{L},\ b \in [0, \infty). \tag{A.1}$$

Thus we allow functions $h_0(w)$ to be exponentially large as $|w| \to \infty$.

We define recursively neighborhoods $\Omega_n(b)$ of $\mathcal{L}''$ for $n = 0, 1, 2, \ldots$. Let $\Omega_{n+1}(b)$ be those $w \in \Omega_n(b)$ such that the disc with center w and radius $2^{-(n+1)}\mathcal{R}$ is contained in $\Omega_n(b)$.

Next, let $w \in \Omega_n(b)$ and let Γ be the circle with center w and radius $2^{-n}\mathcal{R}$. The following two weak asymptotic estimates are simply proved with (4.5):

$$\frac{1}{2\pi i}\int_\Gamma R_n(u, w, b) u^m du = \mathcal{O}(|e^{-\frac{1}{3}w^3 + b^2 w - \frac{2}{3}b^3}|), \tag{A.2}$$

$$\sup_{u \in \Gamma} |R_n(u, w, b)| = \mathcal{O}(|e^{-((n+1)p + \frac{1}{2})(\frac{1}{3}w^3 - b^2 w + \frac{2}{3}b^3)}|) \tag{A.3}$$

as $|b| \to \infty$ uniformly with respect to $w \in \Omega_n(b)$ and $m \in \{0, 1\}$.

Now we can estimate $h_n(w)$ on $\Omega_n(b)$.

$$\begin{aligned}
h_n(w) &= \frac{1}{2\pi i}\int_\Gamma R_0(u, w, b) h_n(u) du \\
&= \frac{1}{2\pi i}\int_\Gamma R_1(u, w, b) h_{n-1}(u) du - \frac{1}{2\pi i}\int_\Gamma R_1(u, w, b)(\alpha_{n-1} + \beta_{n-1} u) du \\
&=_{\text{(A.2)}} \frac{1}{2\pi i}\int_\Gamma R_1(u, w, b) h_{n-1}(u) du + \widetilde{h}_{n-1}\mathcal{O}(|e^{-\frac{1}{3}w^3 + b^2 w - \frac{2}{3}b^3}|) \\
&\ \ \vdots \\
&=_{\text{(A.2)}} \frac{1}{2\pi i}\int_\Gamma R_n(u, w, b) h_0(u) du + (\widetilde{h}_{n-1} + \cdots + \widetilde{h}_0)\mathcal{O}(|e^{-\frac{1}{3}w^3 + b^2 w - \frac{2}{3}b^3}|) \\
&=_{\text{(A.3) \& (5.3)}} \widetilde{h}_0 \mathcal{O}(|e^{-(np+1+\sigma)(\frac{1}{3}w^3 - b^2 w + \frac{2}{3}b^3)}|).
\end{aligned}$$

Thus with (5.3) we have proved that

$$|h_n(w)| \leq C_n \widetilde{h}_0 |e^{-(np+1+\sigma)(\frac{1}{3}w^3 - b^2 w + \frac{2}{3}b^3)}| \tag{A.4}$$

for all $w \in \Omega_n(b) \cup \mathcal{L}$ and $b \in [0, \infty) \cup [0, i\infty)$.

For $b \geq c > 0$ it is not difficult to prove that $\mathcal{L}'' = \{\sqrt{(1/3)y^2 + b^2} + iy \,|\, |y| \geq \delta' b^\theta\}$ for a certain positive δ' that does not depend on b. With the notation of §8 we have

$$\varepsilon_{n|\mathcal{L}''} = (-1)^n z^{-n} \frac{e^{-\frac{2}{3}zb^3}}{2\pi i}\int_{\delta' b^\theta}^\infty e^{-z\phi(y)}\big[H_n(y) + H_n(-y)\big]\, dy.$$

We choose $z > np + 2 + \sigma$ and estimate $|\varepsilon_{n|\mathcal{L}''}|$:

$$\begin{aligned}
|\varepsilon_{n|\mathcal{L}''}| &\leq_{(A.4)} C_n'\widetilde{h}_0 e^{-\frac{2}{3}zb^3} z^{-n} \int_{\delta' b^\theta}^{\infty} e^{-(z-np-1-\sigma)\phi(y)}\, dy \\
&\leq C_n'\widetilde{h}_0 e^{-\frac{2}{3}zb^3} z^{-n} \int_{\delta' b^\theta}^{\infty} e^{-(z-np-1-\sigma)by^2}\, dy \\
&\leq C_n'\widetilde{h}_0 e^{-\frac{2}{3}zb^3} z^{-n} \frac{e^{-\delta'^2(z-np-1-\sigma)b^{2\theta+1}}}{2\delta' b^{\theta+1}(z-np-1-\sigma)} \\
&\leq C_n''\widetilde{h}_0\, \mathrm{Ai}(z^{\frac{2}{3}}b^2) z^{-n-\frac{4}{3}} e^{-\delta'^2(z-np-2-\sigma)b^{2\theta+1}}.
\end{aligned}$$

With similar estimates for $b \in [ic, i\infty)$ we have proved

$$|\varepsilon_{n|\mathcal{L}''}| \leq C_n\widetilde{h}_0\, \widetilde{\mathrm{Ai}}(z^{\frac{2}{3}}b^2) z^{-n-\frac{4}{3}} e^{-\delta'^2(z-np-2-\sigma)|b|^{2\theta+1}}, \tag{A.5}$$

where the constants δ' and C_n do not depend on b and z.

Remark. For the boundary case that has been handled in §6 it is not difficult to prove that $p = \sigma = 0$, and (6.10) shows that in (A.4) $\widetilde{h}_0$ can be replaced by $(b+1)^{(\theta+1/2)(\alpha+1)}|h_0(b)|$, and further calculations show that in (A.5) $\widetilde{h}_0$ can be replaced by $|h_0(b)|$.

Acknowledgment. We appreciate the remarks and suggestions of the referees regarding earlier versions of the paper.

REFERENCES

[1] N. Bleistein, *Uniform asymptotic expansions of integrals with stationary points near algebraic singularity*, Comm. Pure Appl. Math., 19 (1966), pp. 353–370.

[2] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, New York 1975.

[3] C. Chester, B. Friedman, and F. Ursell, *An extension of the method of steepest descents*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 599–611.

[4] E. T. Copson, *Asymptotic Expansions*, Cambridge Tracts in Math. and Math. Phys., 55, Cambridge University Press, London, 1965.

[5] C. L. Frenzen, *Error bounds via complete monotonicity for a uniform asymptotic expansion of the Legendre Function* $P_n^{-m}(\cosh z)$, in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York 1990, pp. 587–599.

[6] C. L. Frenzen and R. Wong, *A uniform asymptotic expansion of the Jacobi polynomials with error bounds*, Canad. J. Math., 37 (1985), pp. 979–1007.

[7] ———, *Uniform asymptotic expansions of Laguerre polynomials*, SIAM J. Math. Anal., 19 (1988), pp. 1232–1248.

[8] B. Friedman, *Stationary phase with neighboring critical points*, J. Soc. Indust. Appl. Math., 7 (1959), pp. 280–289.

[9] N. Levinson, *Transformation of an analytic function of several variables to a canonical form*, Duke Math. J., 28 (1961), pp. 345–353.

[10] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[11] C. K. Qu and R. Wong, *Transformations to canonical form for uniform asymptotic expansions*, J. Math. Anal. Appl., 149 (1990), pp. 210–219.

[12] P. N. Shivakumar and R. Wong, *Error bounds for a uniform asymptotic expansion of the Legendre function* $P_n^{-m}(\cosh z)$, Quart. Appl. Math., 46 (1988), pp. 473–488.

[13] K. Soni and B. D. Sleeman, *On uniform asymptotic expansions and associated polynomials*, J. Math. Anal. Appl., 124 (1987), pp. 561–583.

[14] K. Soni and R. P. Soni, *A system of polynomials associated with the Chester, Friedman, and Ursell technique*, in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, 1990, pp. 417–440.

[15] K. SONI AND N. M. TEMME, *On a biorthogonal system associated with uniform asymptotic expansions*, IMA J. Appl. Math., 44 (1990), pp. 1–25.

[16] F. URSELL, *Integrals with a large parameter. The continuation of uniformly asymptotic expansions*, Proc. Cambridge Philos. Soc., 61 (1965), pp. 113–128.

[17] ———, *Integrals with a large parameter: Legendre functions of large degree and fixed order*, Math. Proc. Camb. Phil. Soc., 95 (1984), 367–380.

[18] ———, *Integrals with a large parameter and the maximum-modulus principle*, in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, 1990, pp. 477–489.

[19] R. WONG, *Asymptotic Approximations of Integrals*, Academic Press, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 322–353, March 1994

006

UNIFORM ASYMPTOTIC SOLUTIONS OF SECOND-ORDER LINEAR DIFFERENTIAL EQUATIONS HAVING A SIMPLE POLE AND A COALESCING TURNING POINT IN THE COMPLEX PLANE*

T. M. DUNSTER†

Abstract. The asymptotic behavior, as a parameter $u \to \infty$, of solutions of second-order linear differential equations having a simple pole and a coalescing turning point is considered. Uniform asymptotic approximations are constructed in terms of Whittaker's confluent hypergeometric functions, which are uniformly valid in a complex domain that includes both the pole and the turning point. Explicit error bounds for the difference between the approximations and the exact solutions are established. These results extend previous real-variable results of F. W. J. Olver and J. J. Nestor to the complex plane.

Key words. turning point theory, differential equations in the complex plane

AMS subject classifications. 34E20, 34A20, 30E10

1. Introduction. In this paper we seek asymptotic solutions, as $u \to \infty$, of second-order linear differential equations of the form

$$\frac{d^2w}{dz^2} = \{u^2 f(a,z) + g(a,z)\}\, w, \tag{1.1}$$

where u and a are real parameters and the independent variable z lies in some complex domain $\mathbf{D}$ (which may be unbounded). The particular case we shall consider is where $f(a,z)$ has a simple zero located on the nonnegative real z-axis at $z = z_t(a)$, with $f(a,z)$ having no other zeros in $\mathbf{D}$. We assume that $(z - z_t(a))^{-1} f(a,z)$ is either real and negative on the positive real axis (we shall call this case I) or real and positive on the positive real axis (case II). These two cases are considered separately in §§2 and 4, respectively.

The position $z = z_t(a)$ of the turning point of the equation is assumed to be a continuous real function of a, which tends to $z = 0$ as a approaches a critical value, say, a_0 . At the point $z = 0$ we assume that $f(a,z)$ has a simple pole, except in the critical case $a = a_0$ when the turning point coalesces with the pole. By an appropriate scaling, we may assume without loss of generality that $a_0 = 0$. We shall then examine (1.1) for a lying in some closed interval $[\,0, A\,]$ (with A a fixed positive number).

Both the pole and the turning point are to lie in $\mathbf{D}$, and both $f(a,z)$ and $g(a,z)$ are to be holomorphic in $\mathbf{D}$ and continuous functions of a and z, simultaneously, except possibly at $z = 0$ (where they may have poles of certain orders). In particular, the function $g(a,z)$ may either be analytic at $z = 0$ or have a simple or double pole there. Moreover, we assume that $\lim_{z\to 0} z^2 g(a,z)$ is independent of a and that

$$\lim_{z\to 0} z^2 g(a,z) \geq -\tfrac{1}{4} \tag{1.2}$$

(see (2.7) in §2). The reason for the restriction (1.2) is that we require the solutions of (1.1) to be monotonic near the regular singularity $z = 0$; they would be oscillatory otherwise.

In the real-variable case Nestor [4] derived uniform asymptotic approximations for solutions of second-order linear ordinary differential equations having a coalescing

* Received by the editors April 15, 1992; accepted for publication (in revised form) May 6, 1993. This research was supported by the National Science Foundation under grant DMS-9102834.

† Department of Mathematics, San Diego State University, San Diego, California 92182-0314.

turning point and simple pole, in terms of Whittaker functions. In this paper we extend Nestor's results to the complex plane, constructing asymptotic solutions to (1.1) also in terms of Whittaker functions, which are valid in certain subdomains of $\mathbf{D}$, uniformly for $u > 0, a \in [0, A]$. Not only are the present results more general in this regard, but in some instances they are valid for a larger parameter range and include a full set of numerically satisfactory solutions. In particular, for what is equivalent to our case II, Nestor imposes the restriction $m \leq \frac{1}{2}$ and does not construct a uniform asymptotic solution that is always recessive at the pole $z = 0$. By working in the complex plane we are able to overcome both of these restrictions.

It is worth noting that having asymptotic approximations that are valid in the complex plane can be of importance in subsequent identification of standard solutions of (1.1) with the asymptotic solutions. This identification is often greatly facilitated by using complex variables.

Examples of differential equations that are of the form (1.1) are the differential equation satisfied by the Jacobi polynomials and the differential equation satisfied by the Mathieu functions (in their algebraic form). The latter equation is the motivation for the present investigation; having asymptotic solutions in a complex domain containing all the critical points will allow a certain analytic continuation, which in turn should provide asymptotic information on the characteristic exponent of Mathieu's equation.

The new results in this paper in effect unify two current asymptotic theories of a turning point in the complex plane and a simple pole in the complex plane, given in Olver's book [5, Chaps. 11 and 12], and can be regarded as complementary to currently existing uniform asymptotic theories concerning coalescing critical points (see [1], [3], [4], [6]). The first such investigation is in the famous paper of Olver [6], who constructed uniform asymptotic approximations of equations having two coalescing turning points. By an appropriate Liouville transformation Olver arrives at equations of the form

$$d^2W/d\zeta^2 = \left\{\pm u^2(\alpha^2 - \zeta^2) + \psi(u, \alpha, \zeta)\right\} W \tag{1.3}$$

and derives asymptotic solutions in terms of parabolic cylinder functions. These are uniformly valid for $u > 0$ and α lying in some closed interval that contains the critical value $\alpha = 0$ (when the two turning points at $\zeta = \pm\alpha$ coalesce). Olver treats the case for which the principal part of (1.3) is real, that is, for ζ either real or purely imaginary.

In a review paper of 1975 on asymptotics [8], Olver mentions that one of the unsolved problems is an extension of the coalescing turning point theory to the complex plane. The results of this paper partly achieve this goal. To see this, consider (1.3) with ζ now regarded as complex, with $\arg(\zeta) \leq \pi/2$. By the simple Liouville transformation on (1.3)

$$z = \zeta^2, \qquad w(z) = \zeta^{1/2} W(\zeta), \tag{1.4}$$

we arrive at the equation

$$\frac{d^2w}{dz^2} = \left\{\pm u^2\left(\frac{a-z}{4z}\right) - \frac{3}{16z^2} + \frac{\psi(u, \alpha, \zeta)}{4z}\right\} w \qquad (a = \alpha^2), \tag{1.5}$$

where $\arg(z) \leq \pi$. If $\psi(u, \alpha, \zeta)$ is an analytic function of z at $z = 0$, then (1.5) is a special case of the general class of equation we are investigating (satisfying (1.2)),

the $\pm$ corresponding to cases I and II, respectively. Other ranges of $\arg(\zeta)$ can be considered either similarly or by using appropriate connection formulas for the special functions under consideration.

Olver applies his results of [6] to the associated Legendre equation [7] and to Whittaker's equation [9]. The present theory can also be applied to the associated Legendre equation.

2. Case I: $(z - z_t(a))^{-1} f(a, z) < 0$ on the positive real z-axis. Our first step is to make the following Liouville transformation:

$$f(a,z)\left(\frac{dz}{d\xi}\right)^2 = \frac{\alpha - \xi}{\xi}, \tag{2.1a}$$

$$W(\xi) = \left(\frac{d\xi}{dz}\right)^{1/2} w(z), \tag{2.1b}$$

where α is a nonnegative parameter that will be specified shortly, so that $z = 0$ corresponds to $\xi = 0$. Integration of (2.1a) yields the relationship

$$\int_\alpha^\xi \left\{\frac{\tau - \alpha}{\tau}\right\}^{1/2} d\tau = \int_{z_t}^z \{-f(a,t)\}^{1/2} dt, \tag{2.2}$$

the lower limits of integration being chosen so that $z = z_t(a)$ corresponds to $\xi = \alpha$. Explicit integration then gives

$$\xi^{1/2}(\xi - \alpha)^{1/2} - \frac{\alpha}{2}\ln\left(\frac{2\xi - \alpha + 2\xi^{1/2}(\xi - \alpha)^{1/2}}{\alpha}\right) = \int_{z_t}^z \{-f(a,t)\}^{1/2}\, dt. \tag{2.3}$$

Branches for the points $z = 0$ ($\xi = 0$) and $z = z_t(a)$ ($\xi = \alpha$) must be chosen so that $\xi(z)$ is an analytic function of z at both $z = 0$ and $z = z_t(a)$. We temporarily introduce cuts along the real z- and ξ-axes, from $\xi = -\infty$ to $\xi = \alpha$ and from $z = -\infty$ to $z = z_t(a)$. Our choice of branches is such that both sides of (2.2) (and (2.3)) are real and positive when z is lying in the real interval $z > z_t$ and are continuous elsewhere.

We now define α so that $z = 0$ corresponds to $\xi = 0$: from (2.2) we see that this is achieved by specifying

$$\int_0^\alpha \left\{\frac{\alpha - \tau}{\tau}\right\}^{1/2} d\tau = \int_0^{z_t} \{f(a,t)\}^{1/2}\, dt, \tag{2.4}$$

which from (2.3) gives the following definition of α as a continuous nonnegative real function of a:

$$\alpha = \frac{2}{\pi}\int_0^{z_t} \{f(a,t)\}^{1/2}\, dt. \tag{2.5}$$

We shall assume that α is a strictly increasing function of a. Note that α tends to zero as a approaches the critical value 0. In this limiting case the $z - \xi$ relationship is given simply by

$$\xi = \int_0^z \{-f(0,t)\}^{1/2}\, dt. \tag{2.6}$$

We denote by Δ the ξ domain corresponding to the z domain $\mathbf{D}$. Therefore, of course, both the critical points $\xi = 0, \alpha$ lie in Δ. Also, we denote $\alpha = \Lambda$ as the corresponding value of $a = A$.

Let m be a nonnegative real number such that

$$\lim_{z\to 0} z^2 g(a,z) = m^2 - \tfrac{1}{4}. \tag{2.7}$$

Then, the effect of the preceding transformations is to yield the new differential equation

$$\frac{d^2W}{d\xi^2} = \left\{u^2\left(\frac{\alpha-\xi}{\xi}\right) + \frac{m^2-\frac{1}{4}}{\xi^2} + \frac{\psi(\alpha,\xi)}{\xi}\right\} W, \tag{2.8}$$

where, with dots representing differentiation with respect to ξ,

$$\frac{\psi(\alpha,\xi)}{\xi} = \dot{z}^{1/2}\frac{d^2}{d\xi^2}(\dot{z}^{-1/2}) + g\dot{z}^2 - \frac{m^2-\frac{1}{4}}{\xi^2}. \tag{2.9}$$

From (2.1a) and (2.9) we find that

$$\psi(\alpha,\xi) = \frac{\alpha^2 + 4\xi^2 - 16m^2(\alpha-\xi)^2}{16\xi(\alpha-\xi)^2} + \frac{(\alpha-\xi)(4ff'' + 16f^2g - 5f'^2)}{16f^3}, \tag{2.10}$$

where primes represent differentiation with respect to z. If one recalls that $z(\xi)$ is analytic Δ, it is straightforward to show from (2.7) and (2.9) that $\psi(\alpha,\xi)$ is an analytic function in Δ. Moreover, one can show that under the preceding conditions $\psi(\alpha,\xi)$ is continuous for $\xi \in \Delta$ and $\alpha \in [0,\Lambda]$. The proof of this is a fairly straightforward exercise using Cauchy integral representations for z and its ξ-derivatives. The corresponding proof for the real-variable case is considerably more difficult (cf. [6, pp. 142–150]).

If one neglects the term $\psi(\alpha,\xi)/\xi$ in equation (2.8), the resulting "comparison equation" has solutions that can be expressed as a linear combination of any pair of the three Whittaker functions $M_{u\alpha i/2,m}(2u\xi e^{\pi i/2}), W_{\pm u\alpha i/2,m}(2u\xi e^{\pm\pi i/2})$. (The notation we use is the standard one (see, for example, [5, p. 260]).) With this in mind we seek three asymptotic solutions of (2.8) of the form

$$W^{(j)}(u,\alpha,\xi) = \mathcal{U}^{(j)}_{u\alpha/2,m}(2u\xi) + \varepsilon^{(j)}(u,\alpha,\xi) \qquad (j = 0,1,2), \tag{2.11}$$

where we define

$$\mathcal{U}^{(0)}_{k,m}(z) = e^{-k\pi/2}e^{-\pi i(m/2+1/4)}\frac{|\Gamma(m+ik+\frac{1}{2})|}{\Gamma(1+2m)}M_{ik,m}\left(ze^{\pi i/2}\right), \tag{2.12}$$

$$\mathcal{U}^{(1)}_{k,m}(z) = e^{k\pi/2}e^{i\theta}W_{-ik,m}\left(ze^{-\pi i/2}\right), \tag{2.13}$$

and

$$\mathcal{U}^{(2)}_{k,m}(z) = e^{k\pi/2}e^{-i\theta}W_{ik,m}\left(ze^{\pi i/2}\right). \tag{2.14}$$

Here, for convenience, we have introduced the parameter

$$\theta = \arg\Gamma\left(m + ik + \frac{1}{2}\right) - \frac{m\pi}{2} - \frac{\pi}{4}, \tag{2.15}$$

with the principal value taken. The normalizing constants taken in (2.12)–(2.14) are selected to ensure that the three functions form a numerically satisfactory set both for complex values of z and large positive values of k. This is of crucial importance in subsequent construction of error bounds.

We do not attempt to obtain asymptotic expansions. The reason for this is the same as that for the case of two coalescing turning points; see [6, Part D]. In short, error bounds for asymptotic expansions would not be uniformly valid in unbounded domains. This problem was not encountered in the problem of a coalescing turning point and double pole [1], [3].

From the well-known connection and analytic continuation formulas of Whittaker functions (see [5, pp. 261–262]), one easily finds that the three functions are related by

$$\mathcal{U}_{k,m}^{(0)}(z) = \mathcal{U}_{k,m}^{(1)}(z) + \mathcal{U}_{k,m}^{(2)}(z). \tag{2.16}$$

Next, we record the following asymptotic forms of the functions as $z \to 0, \infty$, which we require:

$$\mathcal{U}_{k,m}^{(1)}(z) \sim e^{i\theta} z^{-ik} e^{iz/2} \qquad (z \to \infty, \quad -\pi < \arg z < 2\pi), \tag{2.17}$$

$$\mathcal{U}_{k,m}^{(2)}(z) \sim e^{-i\theta} z^{ik} e^{-iz/2} \qquad (z \to \infty, \quad -2\pi < \arg z < \pi), \tag{2.18}$$

$$\mathcal{U}_{k,m}^{(0)}(z) \sim e^{-k\pi/2} \frac{\left|\Gamma(m+ik+\frac{1}{2})\right|}{\Gamma(1+2m)} z^{m+1/2} \qquad (z \to 0), \tag{2.19}$$

$$\mathcal{U}_{k,m}^{(1,2)}(z) \sim \mp i e^{k\pi/2} \frac{\Gamma(2m)}{\left|\Gamma(m+ik+\frac{1}{2})\right|} z^{1/2-m} \qquad (z \to 0, \quad m > 0), \tag{2.20}$$

$$\mathcal{U}_{k,0}^{(1,2)}(z) \sim \mp i e^{k\pi/2} \frac{1}{\left|\Gamma(ik+\frac{1}{2})\right|} z^{1/2} \ln\left(\frac{1}{z}\right) \qquad (z \to 0). \tag{2.21}$$

An important observation is that $\mathcal{U}_{k,m}^{(0)}(z)$ is recessive at $z = 0$ (with respect to the other solutions), $\mathcal{U}_{k,m}^{(1)}(z)$ is recessive at $z = i\infty$, and $\mathcal{U}_{k,m}^{(2)}(z)$ is recessive at $z = -i\infty$.

We shall use what can now be regarded as a standard method, due to Olver, of obtaining bounds for the error terms $\varepsilon^{(j)}(u, \alpha, \xi)$ in (2.11). First, we find a differential equation for the error terms, and this differential equation is then re-expressed as a Volterra integral equation. A bound for a solution of this integral equation may be found by the method of successive approximations (by using [5, Thm. 10.2, p. 220]). To use this theorem we require suitably defined real-valued auxiliary functions $E_{k,m}^{(j)}(z)$, $M_{k,m}^{(j)}(z)$, and $\theta_{k,m}^{(j)}(z)$, which satisfy[1]

$$\left|\mathcal{U}_{k,m}^{(j+1)}(z)\right| = E_{k,m}^{(j+1)}(z)^{-1} M_{k,m}^{(j)}(z) \sin\theta_{k,m}^{(j)}(z) \tag{2.22}$$

and

$$\left|\mathcal{U}_{k,m}^{(j-1)}(z)\right| = E_{k,m}^{(j-1)}(z)^{-1} M_{k,m}^{(j)}(z) \cos\theta_{k,m}^{(j)}(z). \tag{2.23}$$

[1] In §§2 and 3 we shall suppose j is enumerated modulo 3.

We shall define weight functions $E_{k,m}^{(j)}(z)$ $(j = 0, 1, 2)$ that have an asymptotic behavior that is similar to that of the corresponding functions $|\mathcal{U}_{k,m}^{(j)}(z)|^{-1}$ (with regard to both the complex variable z and the real parameter k). Once these weight functions are prescribed, the modulus and phase functions are implicitly given by (2.22) and (2.23), viz.,

$$M_{k,m}^{(j)}(z) = \left\{ E_{k,m}^{(j-1)}(z)^2 \left| \mathcal{U}_{k,m}^{(j-1)}(z) \right|^2 + E_{k,m}^{(j+1)}(z)^2 \left| \mathcal{U}_{k,m}^{(j+1)}(z) \right|^2 \right\}^{1/2}, \tag{2.24}$$

$$\theta_{k,m}^{(j)}(z) = \tan^{-1} \left\{ \frac{E_{k,m}^{(j+1)}(z) \left| \mathcal{U}_{k,m}^{(j+1)}(z) \right|}{E_{k,m}^{(j-1)}(z) \left| \mathcal{U}_{k,m}^{(j-1)}(z) \right|} \right\}. \tag{2.25}$$

To motivate our choice of weight functions let us briefly examine the general asymptotic behavior of the Whittaker functions $\mathcal{U}_{k,m}^{(j)}(z)$ as $k \to \infty$ and $|z| \to \infty$. The functions $\mathcal{U}_{k,m}^{(j)}(z)$ satisfy the differential equation

$$\frac{d^2U}{dz^2} = \left\{ -\left(\frac{z - 4k}{4z} \right) + \frac{m^2 - \frac{1}{4}}{z^2} \right\} U. \tag{2.26}$$

By an application of [5, Chap. 6, Thm. 11.1], with the identification in [5, Chap. 6, eq. (11.01)], i.e.,

$$f(z) = -\left(\frac{z - 4k}{4z} \right), \qquad g(z) = \frac{m^2 - \frac{1}{4}}{z^2}, \tag{2.27}$$

one establishes the existence of two solutions of (2.26) of the form

$$U_k^{(j)}(z) = \left(\frac{z}{z - 4k} \right)^{1/4} \exp\left\{ i\Phi_k^{(j)}(z) \right\} \left\{ 1 + \varepsilon_k^{(j)}(z) \right\} \qquad (j = 1, 2). \tag{2.28}$$

The functions $\Phi_k^{(j)}(z)$ appearing in (2.28) are given by

$$\Phi_k^{(j)}(z) = \frac{1}{2} \int_{4k}^{z} \left(\frac{t - 4k}{t} \right)^{1/2} dt, \tag{2.29}$$

(cf. [1, eq. (5.7)]). These functions have branch points at $z = 0, 4k$, and their branches will be specified shortly in such a way that Im $\Phi_k^{(1)}(z) \to \infty$, Im $\Phi_k^{(2)}(z) \to -\infty$ as $z \to i\infty$, and Im $\Phi_k^{(1)}(z) \to -\infty$, Im $\Phi_k^{(2)}(z) \to \infty$ as $z \to -i\infty$. We shall also specify branches for a third function $\Phi_k^{(0)}(z)$.

The error terms $\varepsilon_k^{(j)}(z)$ are uniformly bounded by [5, Chap. 6, eq. (11.07)] in a certain domain that contains $z = \pm i\infty$ (but not the points $z = 0, 4k$), and these bounds imply that in this domain

$$\varepsilon_k^{(j)}(z) = o(1) \qquad (k \to \infty \text{ or } |z| \to \infty). \tag{2.30}$$

With the choice of branches for (2.29) just described we see that $U_k^{(1)}(z)$ is recessive at $z = i\infty$ and $U_k^{(2)}(z)$ is recessive at $z = -i\infty$. Consequently, $\mathcal{U}_{k,m}^{(1)}(z)$ is a multiple of $U_k^{(1)}(z)$ and $\mathcal{U}_{k,m}^{(2)}(z)$ is a multiple of $U_k^{(2)}(z)$.

We introduce cuts for the branch points of $\Phi_k^{(j)}(z)$ $(j = 0, 1, 2)$ as follows. With respect to the branch point at $z = 4k$ we introduce a cut along a curve on which Im $\Phi_k^{(j)}(z) = 0$. There are three such curves emanating from $z = 4k$, one in the upper half plane (which we label $\mathcal{C}_2$), its conjugate in the lower half plane (labelled $\mathcal{C}_1$), and one along the positive real axis from $z = 4k$ to $z = \infty$ (labelled $\mathcal{C}_0$). The curves $\mathcal{C}_1$ and $\mathcal{C}_2$ emanate from $z = 4k$ at an angle of $\mp 2\pi/3$ with the positive real axis, respectively, and are asymptotic to the lines Im $z = \mp 2k\pi$ as Re $z \to -\infty$ (see Fig. 1).

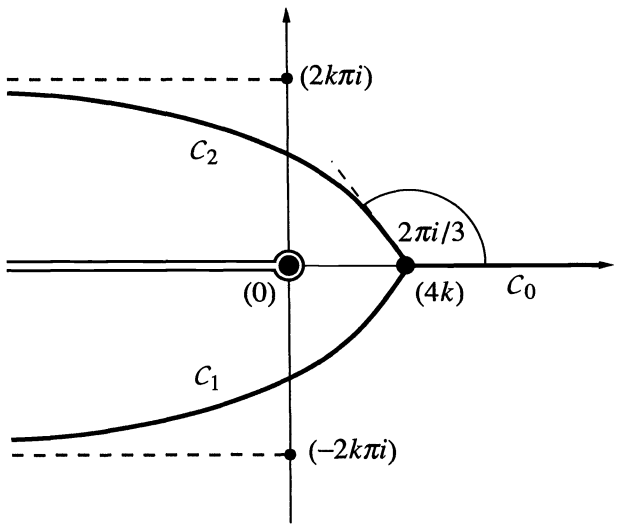


FIG. 1. *z-plane.*

We define $\Phi_k^{(j)}(z)$ $(j = 0, 1, 2)$ to have the branch cut $\mathcal{C}_j$, and with respect to the branch point at $z = 0$ we introduce, for all three functions, a cut along the negative real z-axis from $z = 0$ to $z = -\infty$. The branches in (2.29) are now selected so that the three functions are continuous in their respective cut planes, such that Im $\Phi_k^{(1)}(z) \to \infty$ as $z \to i\infty$, Im $\Phi_k^{(2)}(z) \to \infty$ as $z \to -i\infty$, and Im $\Phi_k^{(0)}(z) \to -\infty$ as $z \to \pm i\infty$.

We now are in a position to define weight functions for $\mathcal{U}_{k,m}^{(j)}(z)(j = 0, 1, 2)$ and we shall use the functions $\Phi_{k+1}^{(j)}(z)$ to do this. The branch cuts associated with these functions emanating from $z = 4k + 4$ divide the z-plane into three regions, which we denote by $\mathcal{S}_k^{(j)}$ (see Fig. 2). Each of the functions $\mathcal{U}_{k,m}^{(j)}(z)$ is recessive in $\mathcal{S}_k^{(j)}$ and dominant in $\mathcal{S}_k^{(j-1)} \cup \mathcal{S}_k^{(j+1)}$. The reason we use $\Phi_{k+1}^{(j)}(z)$ instead of $\Phi_k^{(j)}(z)$ is that the region $\mathcal{S}_k^{(0)}$ as it is now defined does not vanish as $k \to 0$.

We denote by $\mathrm{S}_\alpha^{(j)}(\alpha = 2k/u)$ the regions in the ξ-plane $(\xi = z/(2u))$ corresponding to $\mathcal{S}_k^{(j)}$. The so-called level curves Im $\Phi_{u\alpha/2+1}^{(j)}(2u\xi)$ = constant are important in our asymptotic analysis, and some of these curves are indicated in Fig. 3. In this figure the heavy lines emanating from the point $\xi = \alpha + 2/u$ are the curves Im $\Phi_{u\alpha/2+1}^{(j)}(2u\xi) = 0$, which form the boundaries of $\mathrm{S}_\alpha^{(j)}$.

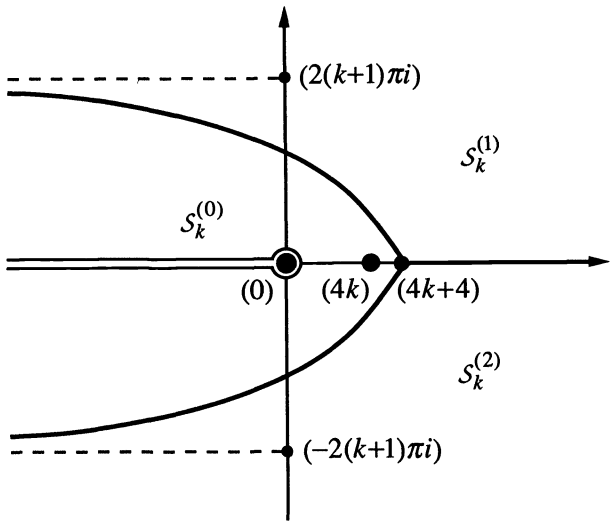


FIG. 2. *z-plane.*

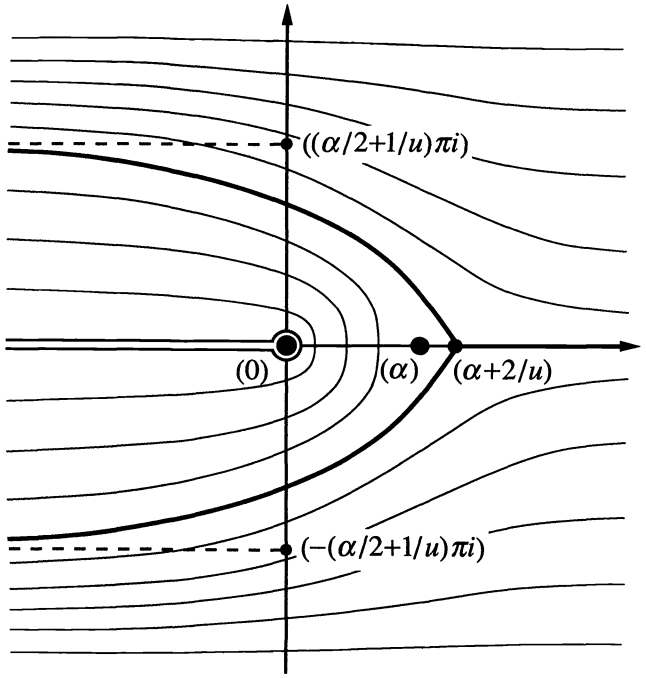


FIG. 3. *ξ-plane: level curves for* $\operatorname{Im} \Phi^{(j)}_{ue/2+1}(2u\xi) = \text{constant}$.

With (2.16) and (2.28) in mind we define

$$E^{(j)}_{k,m}(z) = \left|\exp\left\{-i\Phi^{(j)}_{k+1}(z)\right\}\right| \qquad \left(z \in \mathcal{S}^{(1)}_k \cup \mathcal{S}^{(2)}_k\right) \tag{2.31}$$

for $j = 0, 1, 2$. Note that with our choice of branches $E_{k,m}^{(j)}(z) \geq 1$ for $z \in \mathcal{S}_k^{(j)}$ and $E_{k,m}^{(j)}(z) \leq 1$ for $z \in \mathcal{S}_k^{(j-1)} \cup \mathcal{S}_k^{(j+1)}$.

Unfortunately, (2.31) is not an appropriate definition for the weight functions in $\mathcal{S}_k^{(0)}$ because the Whittaker functions $\mathcal{U}_{k,m}^{(j)}(z)$ have an asymptotic behavior that is different from that of (2.28) near the singularity $z = 0$. Therefore, we must give a different definition for $E_{k,m}^{(j)}(z)$ in $\mathcal{S}_k^{(0)}$. Our choices should reflect both the k-asymptotic behavior $\mathcal{U}_{k,m}^{(j)}(z)$ in $\mathcal{S}_k^{(0)}$ and the behavior of these functions as $z \to 0$ (see (2.19)–(2.21)). Also, bear in mind that $E_{k,m}^{(j)}(z)$ must be continuous for $z > 0$, and so our definitions must be such that $E_{k,m}^{(j)}(z) \to 1$ as z approaches either of the two curves that form the boundary of $\mathcal{S}_k^{(0)}$, i.e., as $\operatorname{Im} \Phi_{k+1}^{(j)}(z) \to 0$.

We begin by defining a positive real-valued radial function $R_k(z)$ having the properties

$$R_k(z) \sim |z| \qquad (z \to 0), \tag{2.32}$$

$$R_k(z) \to \infty \qquad \left(\left|\exp\left\{-i\Phi_{k+1}^{(j)}(z)\right\}\right| \to 1\right). \tag{2.33}$$

The following function satisfies these criteria:

$$R_k(z) = \left(\frac{\left|\exp\left\{-i\Phi_{k+1}^{(0)}(0)\right\}\right| - 1}{\left|\exp\left\{-i\Phi_{k+1}^{(0)}(z)\right\}\right| - 1}\right) \frac{|z|}{1+|z|}. \tag{2.34}$$

Given any point $z \in \mathcal{S}_k^{(0)}$, we define a domain $\mathcal{D}_k^{(0)}(z)$ of a complex t-plane to be the set of points satisfying

$$\operatorname{Im}\Phi_{k+1}^{(0)}(t) > \operatorname{Im}\Phi_{k+1}^{(0)}(z), \tag{2.35}$$

$$|t| < R_k(z), \tag{2.36}$$

$$-\pi < \arg t < \pi \tag{2.37}$$

(see Fig. 4(a)). Note that the level curve $\operatorname{Im} \Phi_{k+1}^{(0)}(t) = \operatorname{Im}\Phi_{k+1}^{(0)}(z)$ forms part of the boundary of $\mathcal{D}_k^{(0)}(z)$.

Similarly, given any point $z \in \mathcal{S}_k^{(0)}$, we define $\mathcal{D}_k^{(1,2)}(z)$ to be the union of the set of points in the t-plane satisfying $\operatorname{Im} \Phi_{k+1}^{(1)}(t) = 0$ with those satisfying

$$\operatorname{Im}\Phi_{k+1}^{(1)}(t) > \operatorname{Im}\Phi_{k+1}^{(1)}(z), \tag{2.38}$$

$$|t| \geq R_k(z), \tag{2.39}$$

$$-\pi < \arg t < \pi \tag{2.40}$$

(see Fig. 4(b)).

Following Olver [6], we introduce so-called balancing functions given by

$$\Omega_d(z) = \begin{cases} |z|^{-d}, & 0 < |z| \leq 1, \\ 1, & |z| > 1, \end{cases} \tag{2.41}$$

where d is real and nonnegative. With the preceding definitions we now can define our weight functions $E_{k,m}^{(j)}(z)$ for $z \in \mathcal{S}_k^{(0)}$. For each $z \in \mathcal{S}_k^{(0)}$ ($|z| > 0$) we define

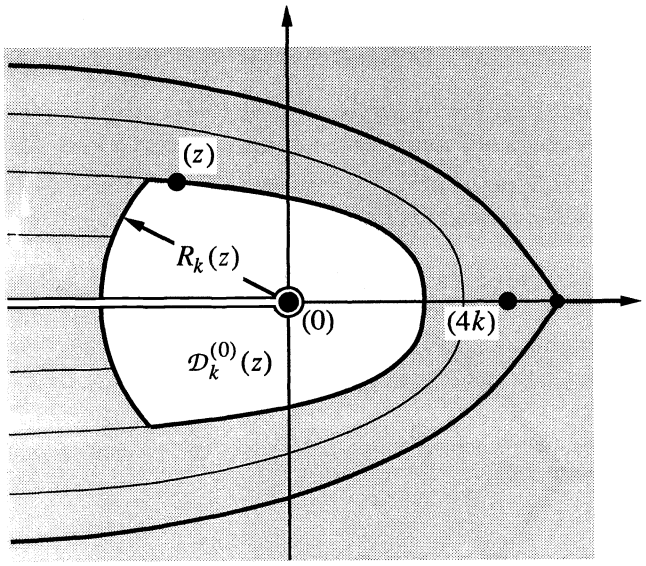


FIG. 4(a). *t-plane.*

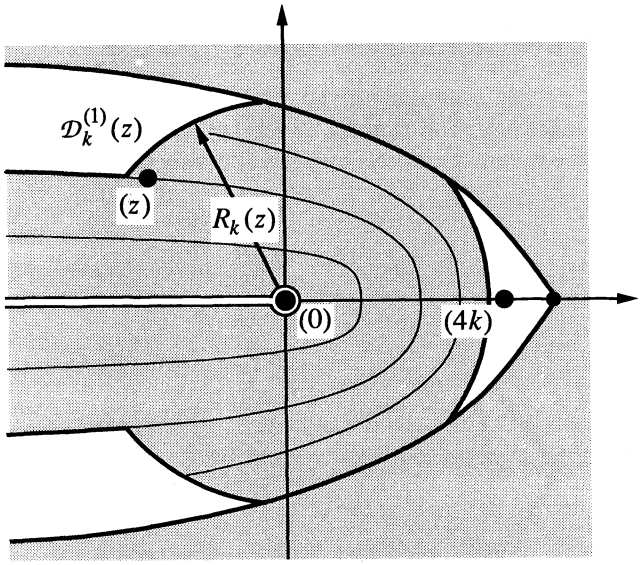


FIG. 4(b). *t-plane.*

$$e_{k,m}^{(j)} E_{k,m}^{(j)}(z)^{-1} = \sup_{t \in D_k^{(j)}(z)} \left\{ \Omega_{1/2}(t) \left| \mathcal{U}_{k,m}^{(j)}(t) \right| \right\} \qquad (j = 0, 1, 2), \tag{2.42}$$

where

$$e_{k,m}^{(0)} = \sup_{t \in D_k^{(0)}(4(k+1))} \left\{ \Omega_{1/2}(t) \left| \mathcal{U}_{k,m}^{(j)}(t) \right| \right\}, \tag{2.43}$$

$$e_{k,m}^{(1)} = e_{k,m}^{(2)} = \sup_{\mathrm{Im}\Phi_{k+1}^{(1)}(t)=0} \left\{ \Omega_{1/2}(t) \left| \mathcal{U}_{k,m}^{(1)}(t) \right| \right\}. \tag{2.44}$$

(The reason for introducing the factor $\Omega_{1/2}(t)$ in (2.42) will be explained shortly.) The coefficients given by (2.43) and (2.44) are introduced so that $E_{k,m}^{(j)}(z) \to 1$ as z approaches any boundary point of $\mathcal{S}_k^{(0)}$. To see this note that as a consequence of the property (2.33) of $R_k(z)$ and the definitions (2.35)–(2.40)

$$\mathcal{D}_k^{(0)}(z^*) = \mathcal{D}_k^{(0)}(4(k+1)), \qquad \mathcal{D}_k^{(1,2)}(z^*) = \left\{ t : \mathrm{Im}\Phi_{k+1}^{(1)}(t) = 0 \right\} \tag{2.45}$$

for any point z^* on the boundary of $\mathcal{S}_k^{(0)}$.

It is straightforward to show from (2.36), (2.39), and (2.42) that

$$e_{k,m}^{(j)} E_{k,m}^{(j)}(z)^{-1} \sim |z|^{-1/2} \left| \mathcal{U}_{k,m}^{(j)}(z) \right| \qquad (z \to 0, j = 0,1,2) \tag{2.46}$$

and hence from (2.24) that

$$M_{k,m}^{(j)}(z) \sim \left\{ \left(e_{k,m}^{(j-1)} \right)^2 + \left(e_{k,m}^{(j+1)} \right)^2 \right\}^{1/2} |z|^{1/2} \qquad (z \to 0). \tag{2.47}$$

The reason for introducing the factor $\Omega_{1/2}(t)$ in (2.42) was to ensure that $\Omega_{1/2}(t)$ $|\mathcal{U}_{k,m}^{(1,2)}(t)| \to \infty$ as $t \to 0$, which would not be true without this factor if $0 \le m \le \frac{1}{2}$ (see (2.20) and (2.21)). The importance of this is that the supremum in (2.42) for $j = 1,2$ is attained on the boundary given by (2.39) when $|t|$ is sufficiently small, which, in turn, is a sufficient condition for (2.46) to hold when $j = 1,2$.

Having defined the weight functions for all values of z ($|z| > 0$), we define modulus and phase functions for the derivatives of $\mathcal{U}_{k,m}^{(j)}(z)$ ($j = 0,1,2$) by

$$\left| \mathcal{U}_{k,m}^{(j+1)'}(z) \right| = E_{k,m}^{(j+1)}(z)^{-1} N_{k,m}^{(j)}(z) \sin \omega_{k,m}^{(j)}(z), \tag{2.48}$$

$$\left| \mathcal{U}_{k,m}^{(j-1)'}(z) \right| = E_{k,m}^{(j-1)}(z)^{-1} N_{k,m}^{(j)}(z) \cos \omega_{k,m}^{(j)}(z). \tag{2.49}$$

We now substitute (2.11) into the differential equation (2.8), yielding the following inhomogeneous differential equation for the error term:

$$\begin{aligned} &\frac{\partial^2 \varepsilon^{(j)}(u,\alpha,\xi)}{\partial \xi^2} - \left\{ u^2 \left(\frac{\alpha - \xi}{\xi} \right) + \frac{m^2 - \frac{1}{4}}{\xi^2} \right\} \varepsilon^{(j)}(u,\alpha,\xi) \\ &\quad = \frac{\psi(\alpha,\xi)}{\xi} \left\{ \mathcal{U}_{u\alpha/2,m}^{(j)}(2u\xi) + \varepsilon^{(j)}(u,\alpha,\xi) \right\}, \end{aligned} \tag{2.50}$$

which by variation of parameters can be re-expressed as the integral equation

$$\varepsilon^{(j)}(u,\alpha,\xi) = \int_{\mathcal{P}^{(j)}} K(\xi,v) \frac{\psi(\alpha,v)}{v} \left\{ \mathcal{U}_{u\alpha/2,m}^{(j)}(2uv) + \varepsilon^{(j)}(u,\alpha,v) \right\} dv. \tag{2.51}$$

The kernel $K(\xi,v)$ is given by

$$K(\xi,v) = \frac{\mathcal{U}_{u\alpha/2,m}^{(j)}(2uv)\mathcal{U}_{u\alpha/2,m}^{(j\pm1)}(2u\xi) - \mathcal{U}_{u\alpha/2,m}^{(j)}(2u\xi)\mathcal{U}_{u\alpha/2,m}^{(j\pm1)}(2uv)}{\mathcal{W}\left(\mathcal{U}_{u\alpha/2,m}^{(j)}(2uv), \mathcal{U}_{u\alpha/2,m}^{(j\pm1)}(2uv) \right)}, \tag{2.52}$$

where $\mathcal{W}$ denotes the Wronskian and the choice of suffix is $j+1$ when $\xi \in \mathrm{S}_\alpha^{(j)} \cup \mathrm{S}_\alpha^{(j+1)}$ and is $j-1$ when $\xi \in \mathrm{S}_\alpha^{(j)} \cup \mathrm{S}_\alpha^{(j-1)}$. From well-known results we find that

$$\left|\mathcal{W}\left(\mathcal{U}^{(j)}_{u\alpha/2,m}(2uv), \mathcal{U}^{(j\pm1)}_{u\alpha/2,m}(2uv)\right)\right| = 2u. \tag{2.53}$$

The path of integration $\mathcal{P}^{(j)}$ in (2.51) runs from $v = \tilde{\xi}^{(j)}$ to $v = \xi$, where $\tilde{\xi}^{(0)} = 0$ and $\tilde{\xi}^{(1,2)}$ is some suitably chosen reference point in $\mathrm{S}_\alpha^{(1,2)}$ (possibly at infinity) such that

(i) $\mathcal{P}^{(j)}$ consists of a finite chain of R_2 arcs;

(ii) $\operatorname{Im} \Phi^{(j)}_{u\alpha/2+1}(2uv)$ is nonincreasing as v passes along $\mathcal{P}^{(j)}$ from $\tilde{\xi}^{(j)}$ to ξ;

(iiia) for the segment of $\mathcal{P}^{(0)}$ in $\mathrm{S}_\alpha^{(0)}$, $|v|$ is nondecreasing as v passes along $\mathcal{P}^{(0)}$ from 0;

(iiib) if $\xi \in \mathrm{S}_\alpha^{(0)}$, $|v|$ is nonincreasing as v passes to ξ along the segment of $\mathcal{P}^{(1,2)}$ lying in $\mathrm{S}_\alpha^{(0)}$.

(Here, for simplicity, we identify the region $\mathrm{S}_\alpha^{(0)}$ in the v-plane as being equivalent to $\mathrm{S}_\alpha^{(0)}$ in the ξ-plane.) The reason for these conditions is that $E^{(j)}_{u\alpha/2,m}(2uv)$ is nonincreasing and $E^{(j\pm1)}_{u\alpha/2,m}(2uv)$ is nondecreasing as v passes along $\mathcal{P}^{(j)}$ from $\tilde{\xi}^{(j)}$ to ξ ($\xi \in \mathrm{S}_\alpha^{(j)} \cup \mathrm{S}_\alpha^{(j\pm1)}$); these monotonicity conditions allow us to apply [5, Thm. 10.2, p. 220]. The subsequent error bounds will be uniformly valid in subdomains $\Delta^{(j)}$ ($j = 0, 1, 2$) defined as the set of points in Δ that can be linked to $\tilde{\xi}^{(j)}$ by a path $\mathcal{P}^{(j)}$ satisfying the conditions (i)–(iii).

Before we state our theorem on error bounds, we introduce some terms that appear. We define

$$F(u, \alpha, \xi) = \int \frac{\psi(\alpha, \xi)}{\xi \Omega_\delta(2u\xi)}\, d\xi \tag{2.54}$$

and assume that this integral converges uniformly with respect to α at $\xi = 0$ for some suitably chosen δ in the interval

$$0 \le \delta < \frac{2}{3}. \tag{2.55}$$

Recall that $\psi(\alpha, \xi)$ is analytic in Δ. If $\psi(\alpha, \xi) = O(\xi)$ as $\xi \to 0$, then, of course, we may prescribe $\delta = 0$ (and hence $\Omega_\delta(2u\xi) \equiv 1$ in (2.54)).

Following [1, eq. (5.13)], we introduce the following constant:

$$\kappa_m = \kappa_{m,1}\kappa_{m,2}, \tag{2.56}$$

where

$$\kappa_{m,1} = \sup\left\{\Omega_\delta(z)(1+k)^{-1/3} \sum_{j=0}^{2} E^{(j)}_{k,m}(z)^2 \left|\mathcal{U}^{(j)}_{k,m}(z)\right|^2\right\} \qquad (m > 0), \tag{2.57}$$

$$\kappa_{m,2} = \sup\left\{E^{(0)}_{k,m}(z)^{-1} E^{(1)}_{k,m}(z)^{-1}\right\} \qquad (m > 0), \tag{2.58}$$

the supremum being taken over all $k \ge 0$ and all $|z| > 0$ for (2.57), and the supremum being taken over all $k \ge 0$ and all $z \in \mathcal{S}_k^{(0)} \setminus \{0\}$ for (2.58).

For the case $m = 0$ we select any δ' satisfying

$$0 < \delta' < \frac{1}{3} \tag{2.59}$$

and then define

$$\kappa_{0,1} = \sup\left\{\Omega_\delta(z)\Omega_{\delta'}(z)(1+k)^{-1/3}\sum_{j=0}^{2} E_{k,0}^{(j)}(z)^2\left|\mathcal{U}_{k,0}^{(j)}(z)\right|^2\right\}, \tag{2.60}$$

$$\kappa_{0,2} = \sup\left\{\Omega_{\delta'}(z)^{-1}E_{k,0}^{(0)}(z)^{-1}E_{k,0}^{(1)}(z)^{-1}\right\}, \tag{2.61}$$

where the suprema are taken over the same ranges as for (2.57) and (2.58), respectively. The factor $(1+k)^{-1/3}$ appearing in (2.57) and (2.60) is needed to ensure that these suprema exist. A consequence of this is that the error bounds are necessarily weakened by a factor of $(\frac{1}{2}u\alpha+1)^{1/3}$ (see (2.63), which follows). An indication of the proof of the existence of the suprema will be given in §3.

We now are in a position to state the main theorem for case I.

THEOREM 1. *With the conditions described in the present section,* (2.8) *has, for each* $u > 0$ *and* $\alpha \in [0, \Lambda]$, *solutions* $W^{(j)}(u,\alpha,\xi)$ $(j = 0,1,2)$ *that are holomorphic in* Δ *except at* $\xi = 0$ *and satisfy*

$$W^{(j)}(u,\alpha,\xi) = \mathcal{U}_{u\alpha/2,m}^{(j)}(2u\xi) + \varepsilon^{(j)}(u,\alpha,\xi), \tag{2.62}$$

where

$$\frac{\left|\varepsilon^{(j)}(u,\alpha,\xi)\right|}{M_{u\alpha/2,m}^{(j\pm1)}(2u\xi)}, \frac{\left|\partial\varepsilon^{(j)}(u,\alpha,\xi)/\partial\xi\right|}{2uN_{u\alpha/2,m}^{(j\pm1)}(2u\xi)}$$

$$\le E_{u\alpha/2,m}^{(j)}(2u\xi)^{-1}\left[\exp\left\{\frac{\kappa_m\left(\frac{1}{2}u\alpha+1\right)^{1/3}}{2u}\mathcal{V}_{\mathcal{P}^{(j)}}(F)\right\}-1\right], \tag{2.63}$$

when $\xi \in \Delta^{(j)}$. *In* (2.63) *the suffix on* M *and* N *is* $j+1$ *when* $\xi \in \mathrm{S}_\alpha^{(j)} \cup \mathrm{S}_\alpha^{(j-1)}$ *and is* $j-1$ *when* $\xi \in \mathrm{S}_\alpha^{(j)} \cup \mathrm{S}_\alpha^{(j+1)}$.

The variation of F in (2.63) is given by

$$\mathcal{V}_{\mathcal{P}^{(j)}}(F) = \int_{\mathcal{P}^{(j)}} \frac{|\psi(\alpha,t)|}{|t|\Omega_\delta(2ut)}\,dt. \tag{2.64}$$

Note that the domains $\Delta^{(1,2)}$ depend on the choice of reference points $\tilde{\xi}^{(1,2)}$: these points can be taken at infinity in $\mathrm{S}_\alpha^{(1,2)}$ provided that the variation (2.64) converges at infinity.

It is not difficult to show from (2.64) that $\mathcal{V}_{\mathcal{P}^{(j)}}(F) = O(u^\delta)$ uniformly for $\xi \in \Delta^{(j)}$. For $j = 1,2$ this estimate can be strengthened to $O(1)$ if ξ is bounded away from 0. Consequently, from (2.63) we have the estimates

$$\left|\varepsilon^{(j)}(u,\alpha,\xi)\right| = E_{u\alpha/2,m}^{(j)}(2u\xi)^{-1}M_{u\alpha/2,m}^{(j\pm1)}(2u\xi)O\left((u\alpha+1)^{1/3}u^{-1+\delta}\right) \tag{2.65}$$

as $u \to \infty$ uniformly for $\xi \in \mathrm{S}_\alpha^{(j)} \cup \mathrm{S}_\alpha^{(j\pm1)}$. The number δ can be taken to be zero if $\psi(\alpha,\xi) = O(\xi)$ as $\xi \to 0$; also, for $j = 1,2$ only, δ can be taken to be zero when ξ is bounded away from 0.

3. Existence of the suprema. In this section we establish the existence of the constant κ_m that is defined by (2.56)–(2.61). To do this we require uniform asymptotic approximations for the Whittaker functions $\mathcal{U}_{k,m}^{(j)}(z)$ as $k \to \infty$ and $|z| \to \infty$, which are valid at the turning point $z = 4k$ (in terms of Airy functions) and the pole $z = 0$ (in terms of Bessel functions). In both cases certain Liouville transformations are used, and it is understood that appropriate branches for the branch points are taken so that the transformations are regular at these points.

For the Airy function approximations we use [5, Chap. 11]. The Liouville transformation on (2.26),

$$\frac{2}{3}\zeta^{3/2} = \int_{4k}^{z} \left(\frac{4k-\tau}{\tau}\right)^{1/2} d\tau, \qquad \tilde{W}(\zeta) = \left(\frac{d\zeta}{dz}\right)^{1/2} U(z), \tag{3.1}$$

yields a new differential equation (with $z = 4k, \infty$ corresponding to $\zeta = 0, -\infty$, respectively) of the form

$$\frac{d^2\tilde{W}}{d\zeta^2} = \left\{\zeta + \tilde{\psi}(k,\zeta)\right\}\tilde{W}. \tag{3.2}$$

Here the Schwarzian is given by

$$\tilde{\psi}(k,\zeta) = \frac{\zeta(z^2 + 4k^2 - 4m^2(z-4k)^2)}{z(z-4k)^3} + \frac{5}{16\zeta^2} \tag{3.3}$$

and is analytic at $\zeta = 0$. An application of [5, Chap. 11, Thm. 9.1] (with $u = 1$) yields the solutions

$$\tilde{W}^{(1)}(k,\zeta) = \mathrm{Ai}\left(\zeta e^{2\pi i/3}\right) + \tilde{\varepsilon}^{(1)}(k,\zeta), \tag{3.4}$$

$$\tilde{W}^{(2)}(k,\zeta) = \mathrm{Ai}\left(\zeta e^{-2\pi i/3}\right) + \tilde{\varepsilon}^{(2)}(k,\zeta), \tag{3.5}$$

where the error terms are bounded by [5, Chap. 11, Eq. (9.03)]. In these bounds E and M are auxiliary functions for Airy functions of complex argument satisfying (see [5, Chap. 11, §8.3])

$$\left|\mathrm{Ai}\left(\zeta e^{-2\pi i/3}\right)\right| = \{E_1(\zeta)\}^{-1} M_0(\zeta)\sin\theta_0(\zeta), \tag{3.6a}$$

$$\left|\mathrm{Ai}\left(\zeta e^{2\pi i/3}\right)\right| = \{E_{-1}(\zeta)\}^{-1} M_0(\zeta)\cos\theta_0(\zeta). \tag{3.6b}$$

From the error bounds one can show that

$$\left|\tilde{\varepsilon}^{(1)}(k,\zeta)\right| = \{E_{-1}(\zeta)\}^{-1} M_0(\zeta)O(1), \tag{3.7a}$$

$$\left|\tilde{\varepsilon}^{(2)}(k,\zeta)\right| = \{E_1(\zeta)\}^{-1} M_0(\zeta)O(1), \tag{3.7b}$$

where the $O(1)$ term is uniform for $z \in \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)}$ and $0 \le k < \infty$. Moreover, this $O(1)$ term is $o(1)$ if $k \to \infty$ or $|z| \to \infty$ ($z \in \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)}$).

The asymptotic solution (3.4) is recessive as $\zeta \to \infty e^{-2\pi i/3}$, which corresponds to $z \to i\infty$, and hence we can claim the existence of a constant $\tilde{c}_{k,m}^{(1)}$ such that

$$\mathcal{U}_{k,m}^{(1)}(z) = \tilde{c}_{k,m}^{(1)} \left(\frac{z\zeta}{4k-z}\right)^{1/4} \left[\mathrm{Ai}\left(\zeta e^{2\pi i/3}\right) + \tilde{\varepsilon}^{(1)}(k,\zeta)\right] \tag{3.8}$$

since both solutions are recessive at $z = i\infty$. This constant can be determined by comparing both sides of (3.8) as $\zeta \to -\infty, z \to +\infty$: one finds by this method

$$\tilde{c}_{k,m}^{(1)} = 2\pi^{1/2} e^{\pi i/6} e^{i\theta} \left(\frac{e}{k}\right)^{ik}. \tag{3.9}$$

Similarly, one finds

$$\mathcal{U}_{k,m}^{(2)}(z) = \overline{\tilde{c}_{k,m}^{(1)}} \left(\frac{z\zeta}{4k-z}\right)^{1/4} \left[\mathrm{Ai}\left(\zeta e^{-2\pi i/3}\right) + \tilde{\varepsilon}^{(2)}(k,\zeta)\right]. \tag{3.10}$$

Consider first the supremum (2.57). From (3.1) a straightforward calculation establishes that

$$\lim_{\zeta\to 0, z\to 4k} \left(\frac{z\zeta}{4k-z}\right)^{1/4} = (2k)^{1/6}, \tag{3.11}$$

and hence, by using this result and (2.16), (2.43), (2.44), (3.4), and (3.5), it can be shown that

$$e_{k,m}^{(j)} = O(k^{1/6}) \qquad (k \to \infty). \tag{3.12}$$

Therefore, from the definition (2.42) we see that[2]

$$E_{k,m}^{(j)}(z)^2 \left|\mathcal{U}_{k,m}^{(j)}(z)\right|^2 \le \left\{\frac{e_{k,m}^{(j)}}{\Omega_{1/2}(z)}\right\}^2 \le K \min\{1,|z|\}(1+k)^{1/3} \tag{3.13}$$

for $z \in \mathcal{S}_k^{(0)}, k \ge 0$. From (2.57) and (3.13) it is now clear that $\kappa_{m,1}$ exists as a supremum over $z \in \mathcal{S}_k^{(0)}, k \ge 0$.

To establish the existence of $\kappa_{m,1}$ it remains to consider $z \in \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)}, k \ge 0$. For this range we use (2.16), (3.8), and (3.10) and the following bound ($j = 1, 2$) for Airy functions of complex argument:

$$\left|\mathrm{Ai}(\zeta e^{\pm 2\pi i/3}) + \tilde{\varepsilon}^{(j)}(k,\zeta)\right| \le K \frac{\left|\exp\left\{i\Phi_k^{(j)}(z)\right\}\right|}{1+|\zeta|^{1/4}}. \tag{3.14}$$

To complete the proof of the existence of $\kappa_{m,1}$ we see, from the preceding formulas, that it is sufficient to establish the uniform boundedness of the following function for $z \in \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)}, k \ge 0$:

$$\sigma(k,z) = \sigma_1(k,z)\sigma_2(k,z)\sigma_3(k,z), \tag{3.15}$$

[2] Here and throughout, K is used generically as a positive constant (independent of z and k).

where

$$\sigma_1(k,z) = \left|\tilde{c}_{k,m}^{(j)}\right|^2 \Omega_\delta(z), \tag{3.16}$$

$$\sigma_2(k,z) = \left|\frac{z\zeta}{(1+k)^{2/3}(z-4k)}\right|^{1/2} \frac{1}{\left(1+|\zeta|^{1/4}\right)^2}, \tag{3.17}$$

$$\sigma_3(k,z) = \left|\exp\left\{2i\Phi_k^{(j)}(z) - 2i\Phi_{k+1}^{(j)}(z)\right\}\right|. \tag{3.18}$$

The factor $|\tilde{c}_{k,m}^{(j)}|$ on the right-hand side of (3.16) is clearly bounded for $k \geq 0$. On noting that

$$|z| \geq z_0 > 0 \qquad (z \in S_k^{(1)} \cup S_k^{(2)}, \quad k \geq 0), \tag{3.19}$$

where z_0 is some positive constant, it follows that the balancing function in (3.16) is also uniformly bounded on the range of z and k under consideration. Using (3.1) and (3.11), one can show that $\sigma_2(k,z)$ is also uniformly bounded for this range. To prove the same for $\sigma_3(k,z)$ we consider

$$J(k,z) = \int_{4k}^{z} \left(\frac{t-4k}{t}\right)^{1/2} dt - \int_{4(k+1)}^{z} \left(\frac{t-4(k+1)}{t}\right)^{1/2} dt \tag{3.20}$$

and show that the imaginary part of this is uniformly bounded for $z \in \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)}, k \geq 0$. To do this we consider separately the cases Z bounded and $1/Z$ bounded, where $Z = z/(k+1)$.

Let Z_0 be an arbitrary positive constant. Then, for the case $0 < Z \leq Z_0$ one can show that

$$J(k,z) = 2\ln\left\{\frac{Z-2+(Z^2-4Z)^{1/2}}{2}\right\} + O\left(\frac{1}{k+1}\right), \tag{3.21}$$

and for the case $0 < 1/Z \leq 1/Z_0$ one can show that

$$J(k,z) = 2\ln\left\{\frac{2-4Z^{-1}+2\left(1-4Z^{-1}\right)^{1/2}}{4Z^{-1}}\right\} + O\left(\frac{1}{|z|}\right). \tag{3.22}$$

Both the O terms in (3.21), (3.22) hold uniformly for $z \in \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)}, k \geq 0$, and therefore Im $J(k,z)$ is uniformly bounded, as required. The existence of (2.57) is thus assured, and that of (2.60) is proved in essentially the same manner.

For the second supremum $\kappa_{m,2}$ we consider separately the two cases $0 \leq k \leq 1$ and $1 \leq k \leq \infty$. In the former case Im z is bounded in $\mathcal{S}_k^{(0)}$, and hence one need only consider the asymptotic behavior $E_{k,m}^{(0)}(z)^{-1}E_{k,m}^{(1)}(z)^{-1}$ at $z = 0$ (see (2.16)–(2.21) and (2.46))to establish that this product is uniformly bounded for $z \in \mathcal{S}_k^{(0)}, 0 \leq k \leq 1$. Similarly, for the case $m = 0$ note that

$$\Omega_{\delta'}(z)^{-1}E_{k,0}^{(0)}(z)^{-1}E_{k,0}^{(1)}(z)^{-1} = O\left(z^{\delta'}\ln\left(\frac{1}{z}\right)\right) \qquad (z \to 0). \tag{3.23}$$

For the case $k \geq 1$ we consider two subcases: (i) pairs of z and k such that $\operatorname{Im} \Phi_{k+1}^{(0)}(2k) \leq \operatorname{Im} \Phi_{k+1}^{(0)}(z) \leq \operatorname{Im} \Phi_{k+1}^{(0)}(0)$ and (ii) pairs of z and k such that $0 \leq \operatorname{Im} \Phi_{k+1}^{(0)}(z) < \operatorname{Im} \Phi_{k+1}^{(0)}(2k)$. Let us denote the first domain in the z-plane by $\mathbf{Z}_0$ and the second domain by $\mathbf{Z}_1$ (see Fig. 5). Clearly $\mathcal{S}_k^{(0)} = \mathbf{Z}_0 \cup \mathbf{Z}_1$, the pole $z = 0$ lies in $\mathbf{Z}_0$, and the turning point $z = 4k$ lies in $\mathbf{Z}_1$. Moreover, these critical points are bounded away from the common boundary of $\mathbf{Z}_0$ and $\mathbf{Z}_1$ when $k \geq 1$.

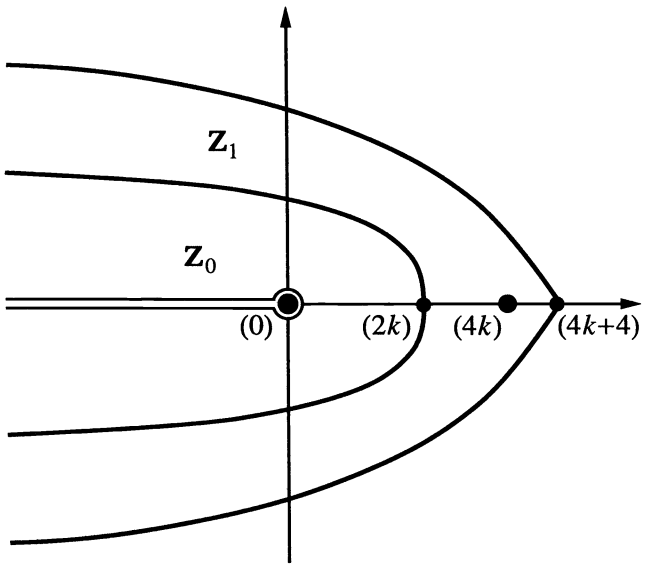


FIG. 5. *z-plane.*

For $z \in \mathbf{Z}_0$ we write $z = 4k\rho$ and consider $\mathcal{U}_{k,m}^{(j)}(4k\rho)$ that satisfy the differential equation

$$\frac{d^2y}{d\rho^2} = \left\{(2k)^2\left(\frac{1-\rho}{\rho}\right) + \frac{(2m)^2-1}{4\rho^2}\right\} y. \tag{3.24}$$

To obtain asymptotic approximations for solutions of (3.24) that are uniformly valid for $z \in \mathbf{Z}_0$ and $k \geq 1$ we use the theory of [5, Chap. 12, §9]. Following this, we make the Liouville transformations

$$\eta^{1/2} = \int_0^\rho \left(\frac{1-t}{t}\right)^{1/2} dt = (\rho-\rho^2)^{1/2} - \frac{i}{2}\ln\left\{1-2\rho+2i(\rho-\rho^2)^{1/2}\right\}, \tag{3.25}$$

$$L = \left(\frac{1-\rho}{\rho}\right)^{1/4} \eta^{1/4} y, \tag{3.26}$$

which transform (3.24) to the form

$$\frac{d^2L}{d\eta^2} = \left\{\frac{(2k)^2}{4\eta} + \frac{(2m)^2-1}{4\eta^2} + \frac{\phi(\eta)}{\eta}\right\} L, \tag{3.27}$$

where

$$\phi(\eta) = \frac{8(1-\rho)}{\rho^4} - \frac{5}{16\rho(1-\rho)^3} - \frac{3}{16\eta} + \frac{\left((2m)^2-1\right)\left(4\rho^2-4\rho+\eta\right)}{16\rho(1-\rho)\eta}. \tag{3.28}$$

The function $\phi(\eta)$ is analytic in the η-domain corresponding to $\mathbf{Z}_0$, in particular at $\eta = 0$ ($z = 0$). Applying [5, Chap. 12, Thm. 9.1] to (3.27) and identifying the solutions that are recessive at $\eta = 0$ ($z = 0$) in the usual manner, we arrive at

$$\begin{aligned} \mathcal{U}_{k,m}^{(0)}(4k\rho) &= l_{k,m}^{(0)} k^{1/2} \left(\frac{\rho\eta}{1-\rho}\right)^{1/4} \\ &\times \left[I_{2m}(2k\eta^{1/2}) + \mathcal{E}_{2m}(2k\eta^{1/2})\mathcal{M}_{2m}(2k\eta^{1/2})O(k^{-1})\right], \end{aligned} \tag{3.29}$$

where

$$l_{k,m}^{(0)} = 2^{1/2}|\Gamma(m + ik + \tfrac{1}{2})|e^{-k\pi/2}k^{-m}. \tag{3.30}$$

Likewise, for solutions that are recessive at $\eta = +\infty$ ($z = i\infty$) we arrive at

$$\begin{aligned} \mathcal{U}_{k,m}^{(1)}(4k\rho) &= l_{k,m}^{(1)} k^{1/2} \left(\frac{\rho\eta}{1-\rho}\right)^{1/4} \\ &\times \left[K_{2m}(2k\eta^{1/2}) + \mathcal{E}_{2m}^{-1}(2k\eta^{1/2})\mathcal{M}_{2m}(2k\eta^{1/2})O(k^{-1})\right], \end{aligned} \tag{3.31}$$

where

$$l_{k,m}^{(1)} = 2\pi^{-1/2}e^{k\pi}e^{i(k+\theta+\pi/4)}k^{-ik}. \tag{3.32}$$

The $O(k^{-1})$ factors in (3.29) and (3.31) hold uniformly for η lying in an unbounded domain whose corresponding z-domain contains all points in the principal z-plane except those on or near the positive real axis from $z = 4k$ to $z = \infty$. Thus, in particular, the identification (3.31) is justified, and (3.29) and (3.31) hold uniformly for $z \in \mathbf{Z}_0$ and $k \geq 1$.

The functions $\mathcal{E}$ and $\mathcal{M}$ are auxiliary functions for Bessel functions of complex argument, defined in [5, Chap. 12, §8]. These have the properties that as $\eta \to 0$

$$\mathcal{M}_{2m}\left(2k\eta^{1/2}\right) \to (2m)^{1/2} \qquad (m > 0), \tag{3.33a}$$

$$\mathcal{E}_{2m}\left(2k\eta^{1/2}\right) \sim \frac{|k\eta^{1/2}|^{2m}}{m^{1/2}\Gamma(2m)} \qquad (m > 0), \tag{3.33b}$$

$$\mathcal{M}_0\left(2k\eta^{1/2}\right) \sim |\ln(1/\eta)|^{1/2}, \tag{3.34a}$$

$$\mathcal{E}_0\left(2k\eta^{1/2}\right) \sim |\tfrac{1}{2}\ln(1/\eta)|^{1/2}, \tag{3.34b}$$

and as $k\eta^{1/2} \to \infty$

$$\mathcal{M}_{2m}\left(2k\eta^{1/2}\right) \sim |k\eta^{1/2}|^{-1/2}, \tag{3.35a}$$

$$\mathcal{E}_{2m}\left(2k\eta^{1/2}\right) \sim \pi^{1/2}\left|\exp\left\{k\eta^{1/2}\right\}\right|, \tag{3.35b}$$

$$\mathcal{E}_{2m}^{-1}\left(2k\eta^{1/2}\right) \leq \pi^{1/2}\left|\exp\left\{-k\eta^{1/2}\right\}\right|\{1=o(1)\}. \tag{3.36}$$

Since $\rho \sim \eta/4$ as $\eta \to 0, \rho \to \infty$ as $\eta \to \infty$, and ρ is bounded away from 1 in the present circumstances, we deduce from [5, Chap. 12, eq. (8.19)] and the preceding results that

$$\begin{aligned}&\Omega_{1/2}(4k\rho)\left|\mathcal{U}_{k,m}^{(0)}(4k\rho)\right| \\ &\leq K l_{k,m}^{(0)} k^{1/2} \frac{|\eta|^{1/2}}{1+|\eta|^{1/4}} \Omega_{1/2}(k\eta)\mathcal{E}_{2m}(2k\eta^{1/2})\mathcal{M}_{2m}(2k\eta^{1/2}),\end{aligned} \tag{3.37}$$

$$\begin{aligned}&\Omega_{1/2}(4k\rho)\left|\mathcal{U}_{k,m}^{(1)}(4k\rho)\right| \\ &\leq K l_{k,m}^{(1)} k^{1/2} \frac{|\eta|^{1/2}}{1+|\eta|^{1/4}} \Omega_{1/2}(k\eta)\mathcal{E}_{2m}^{-1}(2k\eta^{1/2})\mathcal{M}_{2m}(2k\eta^{1/2})\end{aligned} \tag{3.38}$$

for $z \in \mathbf{Z}_0$ and $k \geq 1$.

By virtue of the maximum modulus theorem, it follows from (2.42) that for $k \geq 1$ and $|\rho| > 0$

$$E_{k,m}^{(1)}(4k\rho)^{-1} = \left\{e_{k,m}^{(1)}\right\}^{-1} \Omega_{1/2}(4k\rho_1)\left|\mathcal{U}_{k,m}^{(1)}(4k\rho_1)\right|, \tag{3.39}$$

where ρ_1 lies either on the boundary of $\mathcal{D}_k^{(1)}(4k\rho)$ or on the circle $|\rho| = 1/(4k)$ (on any part of this circle that may lie inside $\mathcal{D}_k^{(1)}(4k\rho)$). Likewise,

$$E_{k,m}^{(0)}(4k\rho)^{-1} = \left\{e_{k,m}^{(0)}\right\}^{-1} \Omega_{1/2}(4k\rho_0)\left|\mathcal{U}_{k,m}^{(0)}(4k\rho_0)\right|, \tag{3.40}$$

where ρ_0 lies either on the boundary of $\mathcal{D}_k^{(0)}(4k\rho)$ or on the part (if any) of the circle $|\rho| = 1/(4k)$ that lies inside $\mathcal{D}_k^{(0)}(4k\rho)$.

Note that $|\rho_0| \leq |\rho_1|$ and that $\rho_j \to 0$ if and only if $\rho \to 0$ $(j = 0, 1)$. Let η_j correspond to ρ_j . Then from (3.37)–(3.40) we can prove that $E_{k,m}^{(0)}(z)^{-1}E_{k,m}^{(1)}(z)^{-1}$ is uniformly bounded for $z \in \mathbf{Z}_0$ and $k \geq 1$ by showing that the following function is uniformly bounded:

$$\left|l_{k,m}^{(0)} l_{k,m}^{(1)}\right| \left\{e_{k,m}^{(0)} e_{k,m}^{(1)}\right\}^{-1} \mathcal{E}_{2m}(2k\eta_0^{1/2})\mathcal{E}_{2m}^{-1}(2k\eta_1^{1/2})\gamma_m(k,\rho_0)\gamma_m(k,\rho_1), \tag{3.41}$$

where

$$\gamma_m(k,\rho) = k^{1/2}\frac{|\eta|^{1/2}\Omega_{1/2}(k\eta)}{\left(1+|\eta|^{1/4}\right)}\mathcal{M}_{2m}(2k\eta^{1/2}). \tag{3.42}$$

Consider first the coefficients in (3.41). By considering the asymptotic behavior of the Gamma function of large complex argument (e.g., see [5, p. 294]), it is straightforward to show that $|l_{k,m}^{(0)} l_{k,m}^{(1)}|$ is bounded as $k \to \infty$. Also, note that $\{e_{k,m}^{(0)} e_{k,m}^{(1)}\}^{-1}$ is bounded for all nonnegative k.

Next, from the definitions of $\mathcal{D}_k^{(0)}(z)$ and $\mathcal{D}_k^{(1)}(z)$ we observe that Re $\{k\eta_1^{1/2}\} \geq$ Re $\{k\eta_0^{1/2}\}$. Therefore, as $k\eta_0^{1/2} \to \infty$ it is clear from (3.35b) that

$$\mathcal{E}_{2m}(2k\eta_0^{1/2})\mathcal{E}_{2m}^{-1}(2k\eta_1^{1/2}) = O(1). \tag{3.43}$$

Furthermore, on recalling that $|\eta_1| \geq |\eta_0|$, we see from (3.33b) that (3.43) is also true when $k\eta_1^{1/2} \to 0$.

It remains to establish that $\gamma_m(k,\rho)$ is uniformly bounded. Essentially, there are three cases to consider, namely,

(i) $k\rho \to \infty$ $(k\eta^{1/2} \to \infty)$ such that $k\eta \to \infty$,
(ii) $k\rho \to \infty$ $(k\eta^{1/2} \to \infty)$ such that $k\eta \to 0$,
(iii) $k\rho \to 0$ $(k\eta^{1/2} \to 0)$.

For case (i) we note that $\Omega_{1/2}(k\eta) \to 1$, and so from (3.35a) we deduce that

$$\gamma_m(k,\rho) = O(1) \qquad \text{(case (i))}. \tag{3.44}$$

For case (ii) we have $\Omega_{1/2}(k\eta) \sim \{k\eta\}^{-1/2}$, and so, again using (3.35a), we find that

$$\gamma_m(k,\rho) = O\left(\frac{1}{k^{1/2}\eta^{1/4}}\right\} \qquad \text{(case (ii))}. \tag{3.45}$$

For case (iii) we use (3.33a) and find (for $m > 0$)

$$\gamma_m(k,\rho) \to (2m)^{1/2} \qquad \text{(case (iii))}. \tag{3.46}$$

To complete the proof of the existence of $\kappa_{m,2}$ one needs to show that $E_{k,m}^{(0)}(z)^{-1}$ $E_{k,m}^{(1)}(z)^{-1}$ is uniformly bounded for $z \in \mathbf{Z}_1$ and $1 \leq k < \infty$. This can be done in a manner similar to that for the case $z \in \mathbf{Z}_0$, by using the uniform asymptotic approximations (3.8) and (3.10) (and the connection formula (2.16)), which are valid in $\mathbf{Z}_1$. Details need not be recorded here. The proof of the existence of the supremum (2.61) (for the case $m = 0$) follows similarly to the preceding.

4. Case II: $(z - z_t(a))^{-1} f(a,z) > 0$ on the positive real z-axis. In this section we construct asymptotic solutions for case II, where $f(a,z)$ is negative on the real axis between the turning point and the pole (when $a > 0$). The appropriate Liouville transformation in this case is given by

$$f(a,z)\left(\frac{dz}{d\xi}\right)^2 = \frac{\xi - \alpha}{\xi}, \tag{4.1a}$$

$$W(\xi) = \left(\frac{d\xi}{dz}\right)^{1/2} w(z), \tag{4.1b}$$

which transforms (1.1) to the form

$$\frac{d^2W}{d\xi^2} = \left\{u^2\left(\frac{\xi - \alpha}{\xi}\right) + \frac{m^2 - \frac{1}{4}}{\xi^2} + \frac{\hat{\psi}(\alpha,\xi)}{\xi}\right\} W, \tag{4.2}$$

where

$$\hat{\psi}(\alpha,\xi) = \frac{\alpha^2 + 4\xi^2 - 16m^2(\xi - \alpha)^2}{16\xi(\xi - \alpha)^2} + \frac{(\xi - \alpha)(4ff'' + 16f^2g - 5f'^2)}{16f^3}. \tag{4.3}$$

We integrate (4.1a) to give

$$\int_\alpha^\xi \left\{\frac{\tau-\alpha}{\tau}\right\}^{1/2} d\tau = \int_{z_t}^z \{f(a,t)\}^{1/2}\, dt, \tag{4.4}$$

which yields the relationship

$$\int_{z_t}^z \{f(a,t)\}^{1/2}\, dt = \xi^{1/2}(\xi-\alpha)^{1/2} - \frac{\alpha}{2}\ln\left\{\frac{2\xi-\alpha+2\xi^{1/2}(\xi-\alpha)^{1/2}}{\alpha}\right\}. \tag{4.5}$$

The choice of branches is such that both sides are real and positive for $z \in (z_t, \infty)$ and $\xi \in (\alpha, \infty)$, with $\xi(z)$ an analytic function of z in $\mathbf{D}$. We denote by $\hat{\Delta}$ the ξ domain corresponding to the z domain $\mathbf{D}$.

Again, we specify α so that the poles $z = 0$ and $\xi = 0$ correspond, which gives the formula

$$\alpha = \frac{2}{\pi}\int_0^{z_t} \{-f(a,t)\}^{1/2}\, dt. \tag{4.6}$$

We assume that α is a strictly increasing function of a, with $\alpha \in [0, \Lambda]$.

The comparison equation this time has as its solutions the Whittaker functions $M_{u\alpha/2,m}(2u\xi)$, $W_{u\alpha/2,m}(2u\xi)$, and $W_{-u\alpha/2,m}(2u\xi e^{\pm\pi i})$. We thus seek asymptotic solutions of (4.2) of the form

$$\hat{W}^{(j)}(u,\alpha,\xi) = \hat{\mathcal{U}}^{(j)}_{u\alpha/2,m}(2u\xi) + \hat{\varepsilon}^{(j)}(u,\alpha,\xi) \qquad (j = 0,1,2,3,4), \tag{4.7}$$

where we define

$$\hat{\mathcal{U}}^{(0)}_{k,m}(z) = \frac{\Gamma(k+m+1/2)}{\gamma(k)\Gamma(1+2m)} M_{k,m}(z), \tag{4.8}$$

$$\hat{\mathcal{U}}^{(1)}_{k,m}(z) = e^{-m\pi i}\gamma(k) W_{-k,m}(ze^{-\pi i}), \tag{4.9}$$

$$\hat{\mathcal{U}}^{(2)}_{k,m}(z) = e^{m\pi i}\gamma(k) W_{-k,m}(ze^{\pi i}), \tag{4.10}$$

$$\hat{\mathcal{U}}^{(3)}_{k,m}(z) = \frac{\gamma(k)\Gamma(m-k+1/2)}{\Gamma(k+m+1/2)} W_{k,m}(z), \tag{4.11}$$

$$\hat{\mathcal{U}}^{(4)}_{k,m}(z) = \frac{e^{(k-m+1/2)\pi i}}{\gamma(k)} W_{k,m}(z), \tag{4.12}$$

introducing, for convenience, the parameter

$$\gamma(k) = k^k e^{-k}. \tag{4.13}$$

The reason that we have introduced $\hat{\mathcal{U}}^{(4)}_{k,m}(z)$, which will become clearer later, is due to the complication that $M_{k,m}(z)$ and $W_{k,m}(z)$ are linearly dependent when $k-m-1/2 \in \mathbf{N}$. When $k-m-1/2 \in \mathbf{N}$ it can be shown from (4.8) and (4.12), by using well-known connection formulas for Whittaker functions, that

$$\hat{\mathcal{U}}^{(4)}_{k,m}(z) = \hat{\mathcal{U}}^{(0)}_{k,m}(z). \tag{4.14}$$

For the time being we assume that $k-m-1/2 \notin \mathbf{N}$.

Further connection formulas for the functions $\hat{\mathcal{U}}_{k,m}^{(j)}(z)$ are given by

$$\hat{\mathcal{U}}_{k,m}^{(0)}(z) = \frac{\Gamma(k-m+\frac{1}{2})\Gamma(k+m+\frac{1}{2})}{2\pi\gamma^2(k)}\left[\hat{\mathcal{U}}_{k,m}^{(1)}(z)+\hat{\mathcal{U}}_{k,m}^{(2)}(z)\right], \tag{4.15}$$

$$\hat{\mathcal{U}}_{k,m}^{(4)}(z) = \frac{\Gamma(k-m+\frac{1}{2})\Gamma(k+m+\frac{1}{2})}{2\pi\gamma^2(k)}\left[\hat{\mathcal{U}}_{k,m}^{(1)}(z)-e^{2(k-m)\pi i}\hat{\mathcal{U}}_{k,m}^{(2)}(z)\right], \tag{4.16}$$

and

$$\hat{\mathcal{U}}_{k,m}^{(4)}(z) = \pm\frac{e^{(k-m)\pi i}\Gamma(k+m+\frac{1}{2})}{\gamma^2(k)\Gamma(m-k+\frac{1}{2})}\hat{\mathcal{U}}_{k,m}^{(1,2)}(z) \mp e^{(k-m)\pi i}e^{\pm(k-m)\pi i}\hat{\mathcal{U}}_{k,m}^{(0)}(z). \tag{4.17}$$

The motivating reason for using the particular Whittaker functions (4.8)–(4.12) is their asymptotic behavior at the singularities, which is given as follows:

$$\hat{\mathcal{U}}_{k,m}^{(0)}(z) \sim \frac{\Gamma(k+m+\frac{1}{2})}{\gamma(k)\Gamma(m-k+\frac{1}{2})}z^{-k}e^{z/2} \qquad \left(z\to\infty, \quad -\frac{\pi}{2}<\arg z<\frac{\pi}{2}\right), \tag{4.18}$$

$$\hat{\mathcal{U}}_{k,m}^{(0)}(z) \sim \frac{ie^{-(k-m)\pi i}}{\gamma(k)}z^{k}e^{-z/2} \qquad \left(z\to\infty, \quad \frac{\pi}{2}<\arg z<\frac{3\pi}{2}\right), \tag{4.19}$$

$$\hat{\mathcal{U}}_{k,m}^{(0)}(z) \sim -\frac{ie^{(k-m)\pi i}}{\gamma(k)}z^{k}e^{-z/2} \qquad \left(z\to\infty, \quad -\pi<\arg z<-\frac{3\pi}{2}\right), \tag{4.20}$$

$$\hat{\mathcal{U}}_{k,m}^{(1)}(z) \sim \gamma(k)e^{(k-m)\pi i}z^{-k}e^{z/2} \qquad \left(z\to\infty, \quad -\frac{\pi}{2}<\arg z<\frac{5\pi}{2}\right), \tag{4.21}$$

$$\hat{\mathcal{U}}_{k,m}^{(2)}(z) \sim \gamma(k)e^{-(k-m)\pi i}z^{-k}e^{z/2} \qquad \left(z\to\infty, \quad -\frac{5\pi}{2}<\arg z<\frac{\pi}{2}\right), \tag{4.22}$$

$$\hat{\mathcal{U}}_{k,m}^{(3)}(z) \sim \frac{\gamma(k)\Gamma(m-k+\frac{1}{2})}{\Gamma(k+m+\frac{1}{2})}z^{k}e^{-z/2} \qquad \left(z\to\infty, \quad -\frac{3\pi}{2}<\arg z<\frac{3\pi}{2}\right), \tag{4.23}$$

$$\hat{\mathcal{U}}_{k,m}^{(4)}(z) \sim -\frac{e^{(k-m-1/2)\pi i}}{\gamma(k)}z^{k}e^{-z/2} \qquad \left(z\to\infty, \quad -\frac{3\pi}{2}<\arg z<\frac{3\pi}{2}\right), \tag{4.24}$$

$$\hat{\mathcal{U}}_{k,m}^{(0)}(z) \sim \frac{\Gamma(k+m+\frac{1}{2})}{\gamma(k)\Gamma(1+2m)}z^{m+1/2} \qquad (z\to 0), \tag{4.25}$$

$$\hat{\mathcal{U}}_{k,m}^{(3)}(z) \sim \frac{\gamma(k)\Gamma(2m)}{\Gamma(k+m+\frac{1}{2})}z^{1/2-m} \qquad (z\to 0, \quad m>0), \tag{4.26}$$

$$\hat{U}_{k,0}^{(3)}(z) \sim \frac{\gamma(k)}{\Gamma(k+\frac{1}{2})} z^{1/2} \ln(1/z) \qquad (z \to 0), \tag{4.27}$$

$$\hat{U}_{k,m}^{(1,2)}(z) \sim \mp i \frac{\gamma(k)\Gamma(2m)}{\Gamma(k+m+\frac{1}{2})} z^{1/2-m} \qquad (z \to 0, \quad m > 0), \tag{4.28}$$

$$\hat{U}_{k,0}^{(1,2)}(z) \sim \mp i \frac{\gamma(k)}{\Gamma(k+\frac{1}{2})} z^{1/2} \ln(1/z) \qquad (z \to 0). \tag{4.29}$$

And so in particular, $\hat{U}_{k,m}^{(0)}(z)$ is recessive at $z = 0, \hat{U}_{k,m}^{(1)}(z)$ is recessive at $z = \infty$ $(\pi/2 < \arg z < 3\pi/2), \hat{U}_{k,m}^{(2)}(z)$ is recessive at $z = \infty$ $(-3\pi/2 < \arg z < -\pi/2)$, and $\hat{U}_{k,m}^{(3)}(z)$ is recessive at $z = \infty$ $(-\pi/2 < \arg z < \pi/2)$. Also, $\hat{U}_{k,m}^{(4)}(z)$ is recessive at $z = \infty$ $(-\pi/2 < \arg z < \pi/2)$ and, in addition, is recessive at $z = 0$ when $k - m - 1/2 \in \mathbf{N}$ (as a result of (4.14)).

Corresponding to (2.29), let us define the functions

$$\hat{\Phi}_k^{(j)}(z) = \frac{1}{2} \int_{4k+1}^{z} \left(\frac{t-4k}{t} \right)^{1/2} dt \qquad (k > 0, \quad j = 1, 2, 3), \tag{4.30}$$

with branches defined as follows. With respect to the branch point $z = 4k$ of $\hat{\Phi}_k^{(j)}(z)$ we introduce a cut along a certain curve $\hat{C}_j$. The curve $\hat{C}_2$ consists of the union of the real segment from $z = 4k$ to $z = 4k+1$, with the curve in the upper half plane on which $\operatorname{Re} \Phi_k^{(j)}(z) = 0$. The curve $\hat{C}_1$ is defined as the conjugate of $\hat{C}_2$, and $\hat{C}_3$ is the finite segment of the positive real axis from $z = 4k+1$ to $z = 0$. The nonreal parts of the curves $\hat{C}_1$ and $\hat{C}_2$ emanate from $z = 4k+1$ at an angle of $\mp\pi/2$ with the positive real axis, respectively, and at infinity are asymptotic to the curves

(4.31a)
$$y = \mp \left[\frac{1}{4}(1+\beta(k))^2 e^{(x-\beta(k))/k} - x^2 \right]^{1/2} \qquad (x = \operatorname{Re} z, \quad y = \operatorname{Im} z, \quad k > 0),$$

where

$$\beta(k) = 2k + (4k+1)^{1/2}. \tag{4.31b}$$

When $k = 0$ the curves $\hat{C}_1$ and $\hat{C}_2$ are the vertical lines $\operatorname{Re} z = 1$.

The three curves $\hat{C}_j$ divide the principle z-plane into three domains, which we call $\mathcal{S}_k^{(j)} (j = 1, 2, 3)$. The regions in the ξ-plane corresponding to these will be labeled $\hat{\mathrm{S}}_\alpha^{(j)}$, respectively; see Fig. 6. In this figure pairs of numerically satisfactory solutions in $\hat{\mathrm{S}}_\alpha^{(j)}$ are indicated.

We define $\hat{\Phi}_k^{(j)}(z)$ $(j = 1, 2, 3)$ to have the branch cut $\hat{C}_j$, and with respect to the branch point at $z = 0$ we introduce, for all three functions, a cut along the negative real z-axis from $z = 0$ to $z = -\infty$. The branches in (2.29) are now selected so that the three functions are continuous in their respective cut planes, such that $\operatorname{Re} \hat{\Phi}_k^{(j)}(z) \to \infty$ for $z \in \hat{\mathcal{S}}_k^{(j)}$ as $|z| \to \infty$.

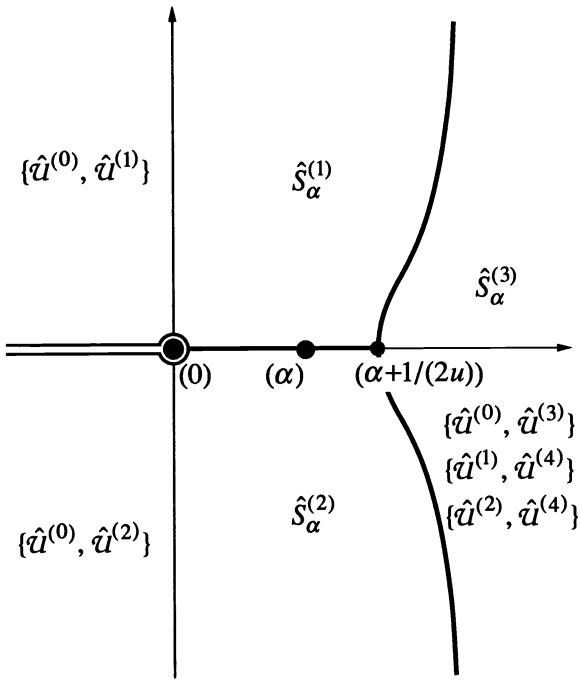


FIG. 6. *ξ-plane.*

We next define

$$\hat{\Phi}_k^{(4)}(z) = \hat{\Phi}_k^{(3)}(z) \qquad (k > 0) \tag{4.32}$$

$$\hat{\Phi}_k^{(0)}(z) = -\hat{\Phi}_k^{(j)}(z) \qquad \left(k > 0, z \in \hat{\mathcal{S}}_k^{(j)}\right), \tag{4.33}$$

so that $\operatorname{Re} \Phi_k^{(0)}(z) \to -\infty$ as $|z| \to \infty$. For the case $k = 0$ we define $\Phi_0^{(j)}(z) = \lim_{k\to 0} \Phi_k^{(j)}(z)$. Thus

$$\hat{\Phi}_0^{(1,2)}(z) = \begin{cases} \frac{1}{2}(z-1) & \left(z \in \hat{\mathcal{S}}_0^{(2,1)}\right), \\ \frac{1}{2}(1-z) & \left(z \in \hat{\mathcal{S}}_0^{(1,2)} \cup \hat{\mathcal{S}}_0^{(3)}\right), \end{cases} \tag{4.34}$$

$$\hat{\Phi}_0^{(4)}(z) = \hat{\Phi}_0^{(3)}(z) = \tfrac{1}{2}(z-1), \tag{4.35}$$

$$\hat{\Phi}_0^{(0)}(z) = -\hat{\Phi}_0^{(j)}(z) \left(z \in \hat{\mathcal{S}}_0^{(j)}\right). \tag{4.36}$$

The level curves in the ξ-plane for case II are defined by $\operatorname{Re} \hat{\Phi}_{u\alpha/2}^{(j)}(2u\xi) =$ constant, some of which are indicated in Fig. 7. In this figure the heavy lines emanating from the point $\xi = \alpha + 1/(2u)$ are the curves $\operatorname{Re} \hat{\Phi}_{u\alpha/2}^{(j)}(2u\xi) = 0$, and those emanating from the turning point $\xi = \alpha$ $(\alpha > 0)$ make an angle $\pm\pi/3$ with the positive

real ξ-axis and satisfy

$$\text{Re } \hat{\Phi}^{(0)}_{u\alpha/2}(2u\xi) = \frac{1}{2}(u\alpha - \beta(u\alpha/2)) + \frac{u\alpha}{2}\ln\left(\frac{1+\beta(u\alpha/2)}{u\alpha}\right). \tag{4.37}$$

When $\alpha = 0$ the level curves are the vertical lines Re ξ = constant.

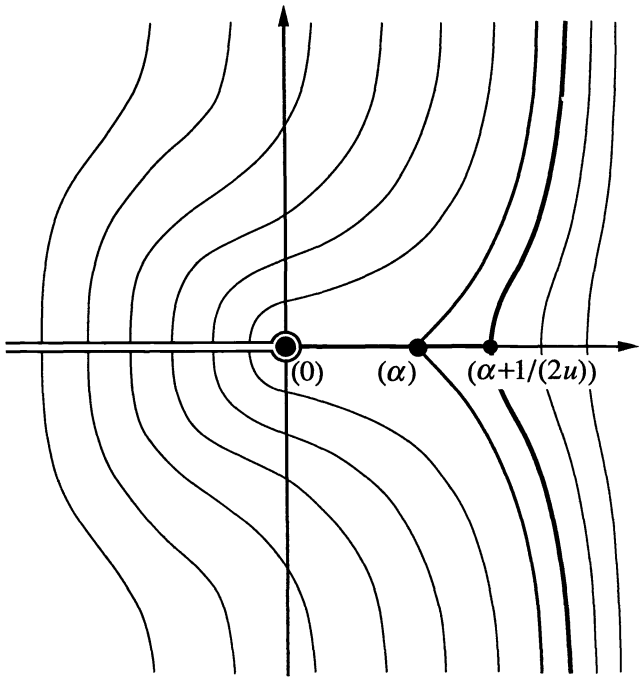


FIG. 7. *ξ-plane.*

We now begin our definitions of the weight functions for $\hat{\mathcal{U}}^{(j)}_{k,m}(z)$. For $z \in \hat{\mathcal{S}}^{(3)}_k$ we define

$$\hat{E}^{(j)}_{k,m}(z) = \left|\exp\left\{\hat{\Phi}^{(j)}_k(z)\right\}\right| \qquad (j = 0, 1, 2, 3, 4), \tag{4.38}$$

and for $j = 4$ only we use the definition (4.38) also for $z \in \hat{\mathcal{S}}^{(1)}_k \cup \hat{\mathcal{S}}^{(2)}_k$.

The definitions of the weight functions for $z \in \hat{\mathcal{S}}^{(1)}_k \cup \hat{\mathcal{S}}^{(2)}_k$ are similar to the corresponding ones for $z \in \hat{\mathcal{S}}^{(0)}_k$ in case I. Given any point $z \in \hat{\mathcal{S}}^{(1)}_k$, we define $\hat{\mathcal{D}}^{(1)}_k(z)(k \geq 0)$ to be the set of points in the t-plane satisfying

$$\text{Re } \hat{\Phi}^{(1)}_k(t) \geq \text{Re } \hat{\Phi}^{(1)}_k(z), \tag{4.39}$$

$$|t| \geq \min\{4k+1, |z|\}, \tag{4.40}$$

$$0 \leq \arg t < \pi; \tag{4.41}$$

see Fig. 8. Given any point $z \in \hat{\mathcal{S}}^{(2)}_k$, we define $\hat{\mathcal{D}}^{(2)}_k(z)$ to be conjugate of $\hat{\mathcal{D}}^{(1)}_k(z)(k \geq 0)$.

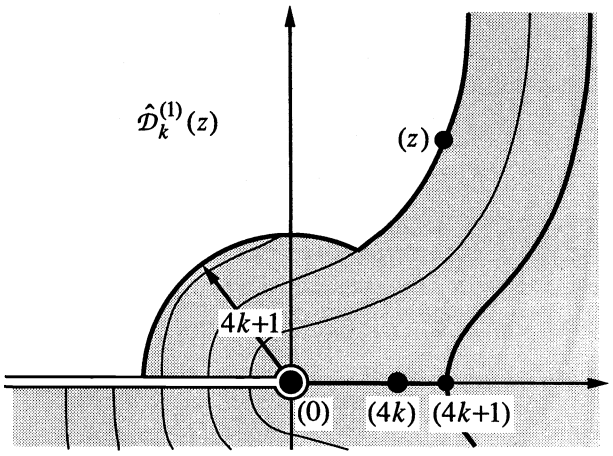


FIG. 8. *t-plane.*

For $k > 0$ choose any point $z \in \hat{\mathcal{S}}_k^{(1)} \cup \hat{\mathcal{S}}_k^{(2)}$ and define $\hat{\mathcal{D}}_k^{(0)}(z)$ to be the set of points in the t-plane satisfying

$$\text{Re } \hat{\Phi}_k^{(0)}(t) \geq \text{Re } \hat{\Phi}_k^{(0)}(z), \tag{4.42}$$

$$|t| \leq |z|, \tag{4.43}$$

$$-\infty < \text{Re } t \leq 4k + 1, \tag{4.44}$$

$$-\pi < \arg t < \pi; \tag{4.45}$$

see Fig. 9(a).

For $k = 0$ we introduce the radial function

$$\hat{R}_0(z) = \frac{|z|}{\min\{1, |1 - \text{Re}(z)|\}}, \tag{4.46}$$

which has the properties

$$\hat{R}_0(z) \sim |z| \qquad (z \to 0), \tag{4.47}$$

$$\hat{R}_0(z) = |z| \qquad (\text{Re}(z) \leq 0), \tag{4.48}$$

$$\hat{R}_0(z) \to \infty \qquad (\text{Re}(z) \to 1). \tag{4.49}$$

Then, given any point $z \in \hat{\mathcal{S}}_0^{(1)} \cup \hat{\mathcal{S}}_0^{(2)}$, we define $\hat{\mathcal{D}}_0^{(0)}(z)$ to be the set of points in the t-plane satisfying

$$-\infty < \text{Re}(t) \leq 1, \tag{4.50}$$

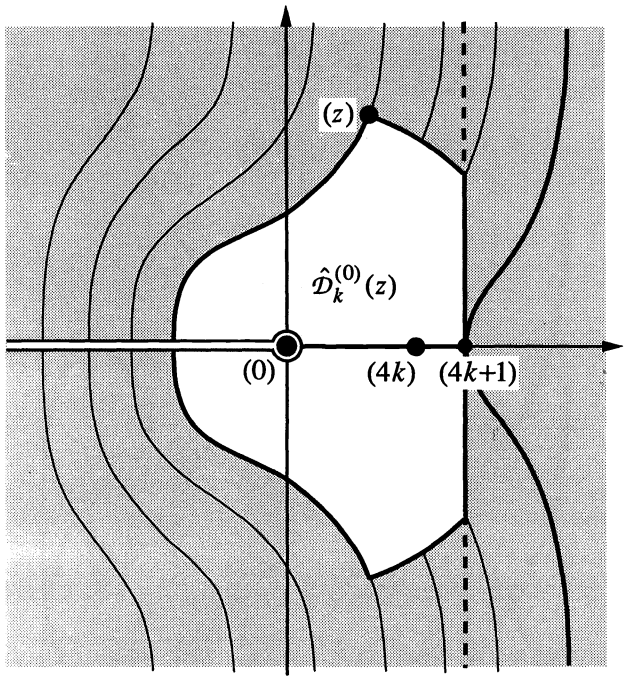


FIG. 9(a). *t-plane.*

$$|\mathrm{Re}(t)| \leq |\mathrm{Re}(z)|, \tag{4.51}$$

$$|t| < \hat{R}_0(z), \tag{4.52}$$

$$-\pi < \arg t < \pi; \tag{4.53}$$

see Fig. 9(b).

For $(j = 0, 1, 2)$ we then define

$$\hat{e}_{k,m}^{(j)} \hat{E}_{k,m}^{(j)}(z)^{-1} = \sup_{t \in \hat{\mathcal{D}}_k^{(j)}(z)} \left\{ \Omega_{1/2}(t) \left| \hat{\mathcal{U}}_{k,m}^{(j)}(t) \right| \right\}, \tag{4.54}$$

with

$$\hat{e}_{k,m}^{(j)} = \sup_{t \in \hat{\mathcal{D}}_k^{(j)}(4k+1)} \left\{ \Omega_{1/2}(t) \left| \hat{\mathcal{U}}_{k,m}^{(j)}(t) \right| \right\}, \tag{4.55}$$

where in (4.54) it is understood that $z \in \hat{\mathcal{S}}_k^{(j)}$ when $j = 1, 2$ and $z \in \hat{\mathcal{S}}_k^{(1)} \cup \hat{\mathcal{S}}_k^{(2)}$ when $j = 0$. By using the uniform asymptotic approximations for Whittaker functions in [2] it can be shown that for all j $\hat{e}_{k,m}^{(j)} = O\left(k^{1/6}\right)$ as $k \to \infty$.[3]

[3] In [2] there are two errors: in equation (4.7) the term -1 should be included in the right-hand side, and in equations (6.17) and (6.20) the term 4 should be replaced by the term e^{-2}.

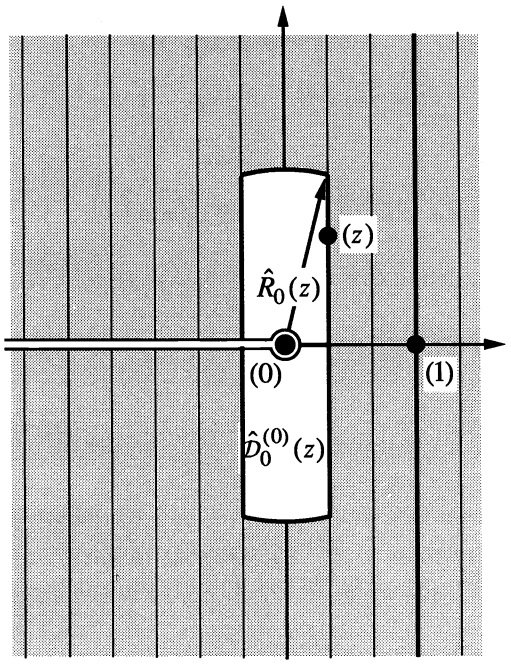


FIG. 9(b). *z-plane.*

The only points in $\hat{\mathcal{S}}_k^{(1)} \cup \hat{\mathcal{S}}_k^{(2)}$ for which we define a weight function for $\hat{\mathcal{U}}_{k,m}^{(3)}(z)$ are those lying in the real interval $0 < z \leq 4k+1$. For these values of z we prescribe

$$\hat{e}_{k,m}^{(3)} \hat{E}_{k,m}^{(3)}(z)^{-1} = \sup_{\substack{\mathrm{Re}(t) \geq z \\ t \in \hat{C}_1 \cup \hat{C}_2 \cup \hat{C}_3}} \left\{ \Omega_{1/2}(t) \left| \hat{\mathcal{U}}_{k,m}^{(3)}(t) \right| \right\}, \tag{4.56}$$

$$\hat{e}_{k,m}^{(3)} = \sup_{\substack{\mathrm{Re}(t) \geq 4k+1 \\ t \in \hat{C}_1 \cup \hat{C}_2}} \left\{ \Omega_{1/2}(t) \left| \hat{\mathcal{U}}_{k,m}^{(3)}(t) \right| \right\}. \tag{4.57}$$

Modulus and phase functions are then defined to satisfy

$$\left| \hat{\mathcal{U}}_{k,m}^{(j)}(z) \right| = \hat{E}_{k,m}^{(j)}(z)^{-1} \hat{M}_{k,m}^{(j,l)}(z) \sin \hat{\theta}_{k,m}^{(j,l)}(z), \tag{4.58}$$

$$\left| \hat{\mathcal{U}}_{k,m}^{(l)}(z) \right| = \hat{E}_{k,m}^{(l)}(z)^{-1} \hat{M}_{k,m}^{(j,l)}(z) \cos \hat{\theta}_{k,m}^{(j,l)}(z), \tag{4.59}$$

where j, l run through the integer values $0 \leq j, l \leq 4, (l \neq j)$, yielding

$$\hat{M}_{k,m}^{(j,l)}(z) = \left\{ \hat{E}_{k,m}^{(j)}(z)^2 \left| \hat{\mathcal{U}}_{k,m}^{(j)}(z) \right|^2 + \hat{E}_{k,m}^{(l)}(z)^2 \left| \hat{\mathcal{U}}_{k,m}^{(l)}(z) \right|^2 \right\}^{1/2}, \tag{4.60}$$

$$\hat{\theta}_{k,m}^{(j,l)}(z) = \tan^{-1} \left\{ \frac{\hat{E}_{k,m}^{(j)}(z) \left| \hat{\mathcal{U}}_{k,m}^{(j)}(z) \right|}{\hat{E}_{k,m}^{(l)}(z) \left| \hat{\mathcal{U}}_{k,m}^{(l)}(z) \right|} \right\}. \tag{4.61}$$

The following asymptotic behavior as $z \to 0$ should be noted (for $0 \le j \le 3$):

$$\hat{e}_{k,m}^{(j)} \hat{E}_{k,m}^{(j)}(z)^{-1} \sim |z|^{-1/2} \left| \hat{\mathcal{U}}_{k,m}^{(j)}(z) \right|, \tag{4.62}$$

$$\hat{M}_{k,m}^{(j,l)}(z) \sim \left\{ \left(\hat{e}_{k,m}^{(j)} \right)^2 + \left(\hat{e}_{k,m}^{(l)} \right)^2 \right\}^{1/2} |z|^{1/2}. \tag{4.63}$$

For the derivatives we define (for $0 \le j, l \le 4, (l \ne j)$)

$$\left| \hat{\mathcal{U}}_{k,m}^{(j)'}(z) \right| = \hat{E}_{k,m}^{(j)}(z)^{-1} \hat{N}_{k,m}^{(j,l)}(z) \sin \hat{\omega}_{k,m}^{(j,l)}(z), \tag{4.64}$$

$$\left| \hat{\mathcal{U}}_{k,m}^{(l)'}(z) \right| = \hat{E}_{k,m}^{(l)}(z)^{-1} \hat{N}_{k,m}^{(j,l)}(z) \cos \hat{\omega}_{k,m}^{(j,l)}(z). \tag{4.65}$$

We next choose reference points $\hat{\xi}^{(j)} (1 \le j \le 4)$, where $\hat{\xi}^{(0)} = 0, \hat{\xi}^{(j)} (j = 1, 2, 3)$ is some suitably chosen point in $\hat{\mathrm{S}}_\alpha^{(j)}$ (possibly at infinity) and $\hat{\xi}^{(4)} = \hat{\xi}^{(3)}$.

We define paths of integration $\hat{\mathcal{P}}^{(j)} (1 \le j \le 4)$ in the v-plane to run from $v = \hat{\xi}^{(j)}$ to $v = \xi$, such that

(i) $\hat{\mathcal{P}}^{(j)}$ consists of a finite chain of R_2 arcs;

(ii) $\operatorname{Re} \hat{\Phi}_{u\alpha/2}^{(j)}(2uv)$ is nonincreasing as v passes along $\hat{\mathcal{P}}^{(j)}$ from $\hat{\xi}^{(j)}$ to ξ, except when $\alpha = 0, j = 0$, in which case $|\operatorname{Re}(v)|$ must be nondecreasing as v passes along $\hat{\mathcal{P}}^{(0)}$;

(iiia) for the segment of $\hat{\mathcal{P}}^{(0)}$ in $\hat{\mathrm{S}}_\alpha^{(1)} \cup \hat{\mathrm{S}}_\alpha^{(2)}$, $|v|$ is nondecreasing as v passes along $\hat{\mathcal{P}}^{(0)}$ from 0;

(iiib) if $\xi \in \hat{\mathrm{S}}_\alpha^{(1,2)}$, $|v|$ is nonincreasing as v passes to ξ along the segment of $\hat{\mathcal{P}}^{(1,2)}$ lying in $\hat{\mathrm{S}}_\alpha^{(1,2)}$;

(iv) the only segment of $\hat{\mathcal{P}}^{(3)}$ that can lie in $\hat{\mathrm{S}}_\alpha^{(1)} \cup \hat{\mathrm{S}}_\alpha^{(2)}$ is the one that consists of the real interval $\xi \le v \le \alpha + 1/(2u)$.

The subsequent error bounds will be uniformly valid in subdomains $\hat{\Delta}_{u\alpha/2,m}^{(j)} (1 \le j \le 4)$, defined as the set of points in $\Xi_{u\alpha/2,m}^{(j)}$ that can be linked to $\tilde{\xi}^{(j)}$ by a path $\mathcal{P}^{(j)}$ in $\Xi_{u\alpha/2,m}^{(j)}$ satisfying the conditions (i)–(iv), where

$$\Xi_{u\alpha/2,m}^{(0)} = \begin{cases} \hat{\Delta} & \left(\frac{1}{2}u\alpha - m - \frac{1}{2} \notin \mathbf{N} \right), \\ \hat{\mathrm{S}}_\alpha^{(1)} \cup \hat{\mathrm{S}}_\alpha^{(2)} & \left(\frac{1}{2}u\alpha - m - \frac{1}{2} \in \mathbf{N} \right), \end{cases} \tag{4.66}$$

$$\Xi_{u\alpha/2,m}^{(1)} = \hat{\mathrm{S}}_\alpha^{(1)} \cup \hat{\mathrm{S}}_\alpha^{(3)}, \tag{4.67}$$

$$\Xi_{u\alpha/2,m}^{(2)} = \hat{\mathrm{S}}_\alpha^{(2)} \cup \hat{\mathrm{S}}_\alpha^{(3)}, \tag{4.68}$$

$$\Xi_{u\alpha/2,m}^{(3)} = \begin{cases} \hat{\mathrm{S}}_\alpha^{(3)} \cup \{\xi : 0 < \xi \le \alpha + 1/(2u)\} & \left(\frac{1}{2}u\alpha - m - \frac{1}{2} \notin \mathbf{N} \right), \\ \emptyset & \left(\frac{1}{2}u\alpha - m - \frac{1}{2} \in \mathbf{N} \right), \end{cases} \tag{4.69}$$

$$\Xi^{(4)}_{u\alpha/2,m} = \hat{S}^{(3)}_{\alpha}. \tag{4.70}$$

Here $\emptyset$ denotes the empty set, and so, for instance, there will be no bounds for $\hat{\varepsilon}^{(3)}(u,\alpha,\xi)$ when $\frac{1}{2}u\alpha - m - \frac{1}{2} \in \mathbf{N}$. We denote the z-domains corresponding to $\Xi^{(j)}_{u\alpha/2,m}$ by $\mathbf{X}^{(j)}_{k,m}$.

As in case I, we choose a δ in the interval

$$0 \le \delta < \frac{2}{3}, \tag{4.71}$$

such that the following integral converges uniformly with respect to α at $\xi = 0$:

$$\hat{F}(u,\alpha,\xi) = \int \frac{\hat{\psi}(\alpha,\xi)}{\xi\Omega_{\delta}(2u\xi)}d\xi. \tag{4.72}$$

Analogously to (2.56), we introduce the following constant:

$$\hat{\kappa}_m = \hat{\kappa}_{m,1}\hat{\kappa}_{m,2}, \tag{4.73}$$

where

$$\hat{\kappa}_{m,1} = \sup\left\{\Omega_{\delta}(z)(1+k)^{-1/3}\sum_{j=0}^{4}\hat{E}^{(j)}_{k,m}(z)^2\left|\hat{\mathcal{U}}^{(j)}_{k,m}(z)\right|^2\right\} \qquad (m>0), \tag{4.74}$$

$$\hat{\kappa}_{m,2} = \sup\left\{\hat{E}^{(0)}_{k,m}(z)^{-1}\hat{E}^{(1)}_{k,m}(z)^{-1}\right\} \qquad (m>0), \tag{4.75}$$

the supremum being taken over all $k \ge 0$ and all $z \in \mathbf{X}^{(j)}_{k,m}$ for (4.74) and the supremum being taken over all $k \ge 0$ and all $z \in \hat{S}^{(1)}_k \cup \hat{S}^{(2)}_k \backslash \{0\}$ for (4.75).

For the case $m = 0$ we select any δ' satisfying

$$0 < \delta' < \frac{1}{3} \tag{4.76}$$

and then define

$$\hat{\kappa}_{0,1} = \sup\left\{\Omega_{\delta}(z)\Omega_{\delta'}(z)(1+k)^{-1/3}\sum_{j=0}^{2}\hat{E}^{(j)}_{k,0}(z)^2\left|\hat{\mathcal{U}}^{(j)}_{k,0}(z)\right|^2\right\}, \tag{4.77}$$

$$\hat{\kappa}_{0,2} = \sup\left\{\Omega_{\delta'}(z)^{-1}\hat{E}^{(0)}_{k,0}(z)^{-1}\hat{E}^{(1)}_{k,0}(z)^{-1}\right\}, \tag{4.78}$$

where the suprema are taken over the same ranges as for (4.74) and (4.75), respectively. The existence of these suprema can be established in a similar manner to the proofs in §3, by using, for example, the uniform asymptotic approximations for Whittaker functions in [2].

We now state our theorem on error bounds for case II.

THEOREM 2. *With the conditions described in the present section,* (4.2) *has, for each* $u > 0$ *and* $\alpha \in [0,\Lambda]$, *solutions* $\hat{W}^{(j)}(u,\alpha,\xi)(j = 0,1,2,3,4)$ *that are holomorphic in* $\hat{\Delta}$ *except at* $\xi = 0$ *and satisfy*

$$\hat{W}^{(j)}(u,\alpha,\xi) = \hat{\mathcal{U}}^{(j)}_{u\alpha/2,m}(2u\xi) + \hat{\varepsilon}^{(j)}(u,\alpha,\xi), \tag{4.79}$$

where

$$\frac{\left|\hat{\varepsilon}^{(j)}(u,\alpha,\xi)\right|}{\hat{M}^{(j,l)}_{u\alpha/2,m}(2u\xi)}, \qquad \frac{\left|\partial\hat{\varepsilon}^{(j)}(u,\alpha,\xi)/\partial\xi\right|}{2u\hat{N}^{(j,l)}_{u\alpha/2,m}(2u\xi)}$$

(4.80)

$$\leq \hat{E}^{(j)}_{u\alpha/2,m}(2u\xi)^{-1}\left[\exp\left\{\frac{\hat{\kappa}_m(\frac{1}{2}u\alpha+1)^{1/3}}{2u}\mathcal{V}_{\hat{\mathcal{P}}^{(j)}}(\hat{F})\right\}-1\right],$$

when $\xi \in \hat{\Delta}^{(j)}_{u\alpha/2,m}$. *In* (4.80) *the suffix* l *accompanying the* j *in* $\hat{M}$ *and* $\hat{N}$ *depends on which region* ξ *lies in, according to Table* 1.

TABLE 1

	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$
$l=0$	—	$\hat{S}^{(1)}_\alpha$	$\hat{S}^{(2)}_\alpha$	$\Xi^{(3)}_{u\alpha/2,m}$	–
$l=1$	$\hat{S}^{(1)}_\alpha$	–	–	–	$\Xi^{(4)}_{u\alpha/2,m}$
$l=2$	$\hat{S}^{(2)}_\alpha$	–	–	–	–
$l=3$	$\hat{S}^{(3)}_\alpha \cap \Xi^{(0)}_{u\alpha/2,m}$	–	–	–	–
$l=4$	–	$\hat{S}^{(3)}_\alpha$	$\hat{S}^{(3)}_\alpha$	–	–

Discussions similar to those for (2.63) lead to the estimates

$$\left|\hat{\varepsilon}^{(j)}(u,\alpha,\xi)\right| = \hat{E}^{(j)}_{u\alpha/2,m}(2u\xi)^{-1}\hat{M}^{(j,l)}_{u\alpha/2,m}(2u\xi)O\Big((u\alpha+1)^{1/3}u^{-1+\delta}\Big) \tag{4.81}$$

as $u \to \infty$ uniformly for $\xi \in \hat{\Delta}^{(j)}_{u\alpha/2,m}$ (with l given by Table 1). As in case I, the constant δ can be taken to be zero if $\hat{\psi}(\alpha,\xi) = O(\xi)$ as $\xi \to 0$; also, for $j \neq 0, \delta$ can be taken to be zero when ξ is bounded away from 0 (which is always true for $j = 4$).

Finally, we remark that the solutions given by Theorem 2 form a numerically satisfactory set in the principal ξ-plane for all ranges of u and α under consideration. The pairs $\{\hat{W}^{(0)}, \hat{W}^{(1)}\}$, $\{\hat{W}^{(0)}, \hat{W}^{(2)}\}$, and $\{\hat{W}^{(1)}, \hat{W}^{(4)}\}$ form sets of numerically satisfactory solutions in $\hat{\mathrm{S}}^{(1)}_\alpha$, $\hat{\mathrm{S}}^{(2)}_\alpha$, and $\hat{\mathrm{S}}^{(3)}_\alpha$, respectively. Moreover, when $u\alpha/2-m-1/2 \notin \mathbf{N}$, the pair $\{\hat{W}^{(0)}, \hat{W}^{(3)}\}$ form a numerically satisfactory set of solutions on the positive real ξ-axis.

REFERENCES

[1] 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 regular singularity, with an application to Legendre functions*, SIAM J. Math. Anal., 17 (1986), pp. 422–450.

[2] T. M. DUNSTER, *Uniform asymptotic expansions for Whittaker's confluent hypergeometric functions*, SIAM J. Math. Anal., 20 (1989), pp. 744–760.

[3] ———, *Uniform asymptotic solutions of second-order linear differential equations having a double pole with complex exponent and a coalescing turning point*, SIAM J. Math. Anal., 21 (1990), pp. 594–618.

[4] J. J. NESTOR, *Uniform Asymptotic Approximations of Solutions of Second-Order Linear Differential Equations, with a Coalescing Simple Turning Point and Simple Pole*, Ph.D. thesis, University of Maryland, 1984 College Park, MD.

[5] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[6] 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.

[7] ———, *Legendre functions with both parameters large*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 175–185.

[8] ———, *Unsolved problems in the asymptotic estimation of special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975 pp. 585–596.

[9] ———, *Whittaker functions with both parameters large: Uniform approximations in terms of parabolic cylinder functions*, Proc. Roy. Soc. Edinburgh Sec. A, 86 (1980) pp. 213–234.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 354–364, March 1994

007

ON THE DERIVATIVE WITH RESPECT TO A PARAMETER OF A ZERO OF A STURM–LIOUVILLE FUNCTION*

ÁRPÁD ELBERT† AND MARTIN E. MULDOON‡

Abstract. A formula for the derivative with respect to a parameter ν of a zero of a suitable solution $z(t,\nu)$ of the differential equation $(p(t,\nu)z')' + r(t,\nu)z = 0$ is derived. This provides a kind of quantitative confirmation of the Sturm comparison theorem in that it shows that the zeros are monotonic in ν in the same direction as $p(t,\nu)$ and in a direction opposite to that of $r(t,\nu)$. It is also shown that the derivative of a zero with respect to a parameter is equal to the value at the zero of a function that satisfies a third-order linear nonhomogeneous differential equation. Different methods of solving this equation lead to different formulas for the derivative. The results are applied to get formulas for the derivative with respect to λ of the zeros of the Gegenbauer polynomial $P_n^{(\lambda)}(x)$. The method also yields a generalization to arbitrary cylinder functions of formulas due, respectively, to Schläfli and Schafheitlin for the derivative with respect to order of the zeros of the Bessel functions $J_\nu(x)$ and $Y_\nu(x)$.

Key words. Sturm–Liouville functions, zeros, monotonicity, Bessel function, ultraspherical polynomials

AMS subject classifications. primary 34C10; secondary 33A65, 33A40

1. Introduction.

The Sturm comparison theorem [21], [22, p. 19] says essentially that if y and Y vanish at a and satisfy

$$(py')' + ry = 0,$$

$$(PY')' + RY = 0,$$

where $0 < P(t) < p(t)$, $R(t) > r(t)$, $a < t < b$, then the next zero of Y to the right of a on (a,b) occurs before the next zero of y. Thus if we have a *family* of functions $y(\nu,t)$ satisfying

$$(p(t,\nu)y')' + r(t,\nu)y = 0, \tag{1.1}$$

where $p(t,\nu)$ increases and $r(t,\nu)$ decreases in ν for each t, and $y(a,\nu) = 0$ for each ν, then the next zero $c = c(\nu)$ of $y(t,\nu)$ to the right of a increases with ν or $dc/d\nu > 0$. This raises the question of whether we can find a formula for $dc/d\nu$ from which its positivity would be obvious. We prove such a formula for suitable functions $p(t,\nu)$ and $r(t,\nu)$:

$$p(c,\nu)[y'(c,\nu)]^2\frac{dc}{d\nu} = \int_a^c \left[-r_\nu(s,\nu)y^2(s,\nu) + p_\nu(s,\nu)[y'(s,\nu)]^2\right] ds. \tag{1.2}$$

(Here and in what follows, the prime indicates partial differentiation with respect to the first place variable, and subscripts denote partial differentiation with respect to the indicated variable.)

* Received by the editors April 6, 1992; accepted for publication (in revised form) April 13, 1993. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

† Mathematical Institute, Hungarian Academy of Sciences, P.O. Box 127, Budapest H–1364, Hungary.

‡ Department of Mathematics and Statistics, York University, North York, Ontario M3J 1P3, Canada.

In §5 we show also that the derivative of a zero with respect to a parameter is equal to the value at the zero of a function that satisfies a third-order linear nonhomogeneous differential equation.

We apply our results to get formulas for the derivative with respect to λ of the zeros of the Gegenbauer polynomial $P_n^{(\lambda)}(x)$. Our method also yields a generalization to arbitrary cylinder functions $C_\nu(x) = \cos\alpha J_\nu(x) - \sin\alpha Y_\nu(x)$ of formulas due to Schläfli [20], [23, p. 508] and Schafheitlin [19], respectively, for the derivatives with respect to order of the zeros of the Bessel functions $J_\nu(t)$ and $Y_\nu(t)$.

In a later paper [6] we apply some of these results to zeros of Hermite functions.

2. The Richardson formula. We begin by considering the differential equation

$$(p(t,\nu)y')' + r(t,\nu)y = 0, \qquad t \in I, \tag{2.1}$$

for each ν in some interval J. We suppose also that $p^{-1}(t,\nu)$ and $r(t,\nu)$ are of class C^1 in a domain of (t,ν)-space that includes $I \times J$. This implies, in view of [10, Cor. 4.1, p. 101], that the solutions of (2.1) are of class C^1 throughout their domains of existence in the (t,ν) plane. Furthermore, [10, Thm. 3.1, p. 95] shows that $y_\nu(t,\nu)$ satisfies the equation obtained by formally differentiating equation (2.1) with respect to ν, with the implied interchange of orders of differentiation. Thus we get

$$(p_\nu y_t)_t + (py_{\nu t})_t + r_\nu y + ry_\nu = 0. \tag{2.2}$$

Multiplying (2.1) by y_ν and (2.2) by $-y$ and adding, we get, as in [18, p. 293], that

$$D_t\left[py_t y_\nu - pyy_{\nu t} - p_\nu y_t y\right] = r_\nu y^2 - p_\nu y_t^2. \tag{2.3}$$

Now, if we are in a situation where either $y(a,\nu) = 0$ or $y_t(a,\nu) = 0$ for each ν in J, and if $c = c(\nu)$ is another zero of $y(t,\nu)$, we see by integrating (2.3) from a to c that

$$p(c)y_t(c,\nu)y_\nu(c,\nu) = \int_a^c \left[r_\nu y^2 - p_\nu y_t^2\right] ds. \tag{2.4}$$

Also, by differentiating $y(c,\nu) = 0$ we have

$$y_t(c,\nu)\frac{dc}{d\nu} + y_\nu(c,\nu) = 0, \tag{2.5}$$

so that we finally get

$$p(c,\nu)[y'(c,\nu)]^2\frac{dc}{d\nu} = \int_a^c \left[-r_\nu(s,\nu)y^2(s,\nu) + p_\nu(s,\nu)[y'(s,\nu)]^2\right] ds. \tag{2.6}$$

Actually, by a well-known transformation we can reduce the differential equation (2.1) to the case where $p(t,\nu) \equiv 1$. In that case, (2.6) takes the simpler form

$$[y'(c,\nu)]^2\frac{dc}{d\nu} = \int_a^c \left[-r_\nu(s,\nu)y^2(s,\nu)\right] ds. \tag{2.7}$$

We consider now the frequently occurring case where the point a is a singular point of the differential equation. We assume again that $p(t,\nu) \equiv 1$. Then if a is the left-hand endpoint of the interval I, for $c > a$ we get, as before,

$$y_t(c,\nu)y_\nu(c,\nu) + \left[yy_{\nu t} - y_t y_\nu\right]_{t=a+\epsilon} = \int_{a+\epsilon}^c \left[r_\nu y^2\right] ds \tag{2.8}$$

for each $\epsilon > 0$. Thus we get the result (2.7) again, provided that the integral on its right-hand side exists and provided that

$$\lim_{\epsilon \to 0^+} [y(a+\epsilon)y_{\nu t}(a+\epsilon) - y_t(a+\epsilon)y_\nu(a+\epsilon)] = 0. \tag{2.9}$$

The result is modified in an obvious way when $c < a$.

3. Applications of the Richardson formula.

3.1. The Bessel equation. As a first application we consider the equation

$$y'' + \left[1 + \frac{1/4 - \nu^2}{t^2}\right] y = 0, \tag{3.1}$$

with a solution $y(t,\nu) = t^{1/2}J_\nu(t)$ vanishing at $t = 0$. The method of §2 is applicable, with $a = 0$, since $J_\nu(0) = 0$, $\nu > 0$, and we recover Schläfli's formula [20], [23, p. 508]

$$\frac{dj}{d\nu} = \frac{2\nu}{jJ_{\nu+1}^2(j)} \int_0^j s^{-1}J_\nu^2(s)ds, \qquad \nu > 0, \tag{3.2}$$

where $j = j(\nu, k)$ is a positive zero of $J_\nu(x)$. The verification of condition (2.9) for $\nu > 0$ is based on the series expansion of $J_\nu(x)$.

3.2. The generalized Airy equation. Here we consider the differential equation

$$y'' + t^\alpha y = 0, \qquad t \geq 0, \tag{3.3}$$

with the initial conditions

$$y(0) = 0, \qquad y'(0) = 1. \tag{3.4}$$

α is supposed to be a fixed positive number. Using the comparison equation $z''+z = 0$, we see that all the positive zeros of y exceed 1. (Suppose that y has a zero on $(0,1)$. The equation $z'' + z = 0$ is a Sturm majorant of (3.3) on $(0,1)$. Thus its solution $\sin t$, which satisfies $z(0) = 0$, $z'(0) = 1$, would have to have a zero there too, which is impossible.) A well-known transformation [23, p. 96] shows that the solution of (3.3) is given by $y(t,\alpha) = \nu^{-\nu}\Gamma(\nu+1)t^{1/2}J_\nu(2\nu t^{1/(2\nu)})$, where $\nu = 1/(\alpha+2)$. Since $y(t,\alpha)$ vanishes at 0 for each $\alpha > 0$, we find, as in §2, that for every positive zero $a_{\alpha k}$ of $y(t,\alpha)$

$$[y'(a,\alpha)]^2 \frac{da}{d\alpha} = -\int_0^a s^\alpha \log s [y(s,\alpha)]^2\, ds. \tag{3.5}$$

Since the integrand changes sign at $s = 1$, it does not seem to be easy to deduce from (3.5) the fact [9] that $da/d\alpha < 0$, $\alpha > 0$.

3.3. Ultraspherical polynomials. With a usual notation ([22], not [7]) for the ultraspherical or Gegenbauer polynomials, we recall [22] that $y_n(t,\lambda) = (1 - t^2)^{\lambda/2+1/4}P_n^{(\lambda)}(t)$ is a solution of

$$\frac{d^2y}{dx^2} + \left\{\frac{(n+\lambda)^2}{1-x^2} + \frac{1/2 + \lambda - \lambda^2 + x^2/4}{(1-x^2)^2}\right\} y = 0. \tag{3.6}$$

For $\lambda > -\frac{1}{2}$, $y_n(1,\lambda) = 0$ for every n, while for every λ we have $y_n(0,\lambda) = 0$ for n odd and $y_n'(0,\lambda) = 0$ for n even. Applying the results of §2, with $a = 0$, we get for a zero c of $P_n^{(\lambda)}(x)$ the formula

$$\frac{dc}{d\lambda} = -(1-c^2)^{-\lambda-1/2}[P_n^{(\lambda)'}(c)]^{-2}\int_0^c \frac{2n+1-2(n+\lambda)s^2}{(1-s^2)^{-\lambda+3/2}}\left[P_n^{(\lambda)}(s)\right]^2 ds. \tag{3.7}$$

Formula (3.7) shows that a positive zero of $P_n^{(\lambda)}(x)$ is a decreasing function of λ for those values of λ that satisfy $\lambda < (2n+1)/(2c^2) - n$. This covers all the zeros on $(0,1)$ in case $-\frac{1}{2} < \lambda \leq \frac{1}{2}$. On the other hand, when $\lambda > \frac{1}{2}$, the integrand in (3.7) may change sign. But in this situation we can use the formula

$$\frac{dc}{d\lambda} = (1-c^2)^{-\lambda-1/2}[P_n^{(\lambda)'}(c)]^{-2}\int_c^1 \frac{2n+1-2(n+\lambda)s^2}{(1-s^2)^{-\lambda+3/2}}\left[P_n^{(\lambda)}(s)\right]^2 ds \tag{3.8}$$

obtained by using the method described at the end of §2 with $a = 1$ or by using the identity

$$\begin{aligned}(2n+1)&\int_{-1}^1 (1-s^2)^{\lambda-3/2}\left[P_n^{(\lambda)}(s)\right]^2 ds \\ &= 2(n+\lambda)\int_{-1}^1 (1-s^2)^{\lambda-3/2}s^2\left[P_n^{(\lambda)}(s)\right]^2 ds.\end{aligned} \tag{3.9}$$

To prove (3.9) we let A denote the integral on its left-hand side and B the integral on its right-hand side. We have, using [22, eq. (4.7.28)],

$$nP_n^{(\lambda)}(x) = x\frac{d}{dx}\{P_n^{(\lambda)}(x)\} - \frac{d}{dx}\{P_{n-1}^{(\lambda)}(x)\} = x\frac{d}{dx}\{P_n^{(\lambda)}(x)\} - \sum_{k=0}^{n-2} c_k P_k^{(\lambda)}(x),$$

where the c_k are constants. Thus, using the orthogonality property of the ultraspherical polynomials, we get

$$\begin{aligned}&\int_{-1}^1 (1-s^2)^{\lambda-1/2}sP_n^{(\lambda)}(s)\frac{d}{ds}P_n^{(\lambda)}(s)ds \\ &\quad = n\int_{-1}^1 (1-s^2)^{\lambda-1/2}\left[P_n^{(\lambda)}(s)\right]^2 ds = n(A-B).\end{aligned} \tag{3.10}$$

Now

$$(2\lambda-1)B = -\int_{-1}^1 s\left[P_n^{(\lambda)}(s)\right]^2\frac{d}{ds}(1-s^2)^{\lambda-1/2}ds,$$

and so, using integration by parts, we obtain

$$\begin{aligned}(2\lambda-1)B &= \int_{-1}^1 \left[P_n^{(\lambda)}(s)\right]^2(1-s^2)^{\lambda-1/2}ds + \int_{-1}^1 2sP_n^{(\lambda)}(s)\frac{d}{ds}P_n^{(\lambda)}(s)(1-s^2)^{\lambda-1/2}ds \\ &= A - B + 2n(A-B).\end{aligned}$$

From this we get $(2n+1)A = 2(n+\lambda)B$, which is the identity (3.9). To recapitulate, formula (3.7) shows that a positive zero of $P_n^{(\lambda)}(x)$ is a decreasing function of λ for those values of λ that satisfy $\lambda < (2n+1)/(2c^2) - n$, and (3.8) shows it for $\lambda > (2n+1)/(2c^2) - n$. Together, these results recover the known result [22] that all the positive zeros are decreasing functions of λ. We remark that the integral in (3.8) exists only for $\lambda > \frac{1}{2}$. But on considering that $0 < c^2 < 1$, this follows from the condition $\lambda > (2n+1)/(2c^2) - n$.

4. The determinant $Q(t,\nu)$. Here we consider a pair of linearly independent solutions $x(t,\nu)$, $y(t,\nu)$ of the differential equation

$$z'' + q(t,\nu)z = 0, \qquad t \in I, \tag{4.1}$$

satisfying the initial conditions

$$x(a,\nu) = \phi(\nu), \qquad y(a,\nu) = 0, \tag{4.2}$$

$$x_t(a,\nu) = 0, \qquad y_t(a,\nu) = 1/\phi(\nu) \tag{4.3}$$

for each ν in some interval J. The function $\phi(\nu)$ is supposed to be differentiable on J. The Wronskian

$$W = x(t,\nu)y_t(t,\nu) - x_t(t,\nu)y(t,\nu) \equiv 1, \qquad t \in I, \tag{4.4}$$

for each $\nu \in J$. We suppose, as in §2, that $q(t,\nu)$ is of class C^1 in a domain of (t,ν)-space that includes $I \times J$. This implies, as pointed out in §2, that the solutions $x(t,\nu)$, $y(t,\nu)$ and, indeed, any linear combination

$$z(t,\nu) = \cos\alpha\, x(t,\nu) - \sin\alpha\, y(t,\nu) \tag{4.5}$$

are of class C^1 throughout their domains of existence in the (t,ν)-plane. Let $c = c(\nu,\alpha)$ be a zero of $z(t,\nu)$ for some fixed α. Thus

$$\cos\alpha\, x(c,\nu) - \sin\alpha\, y(c,\nu) = 0. \tag{4.6}$$

Proceeding somewhat as in [23, p. 508], we differentiate (4.6) with respect to ν to get

$$\cos\alpha[x_\nu(c,\nu) + x_t(c,\nu)c_\nu] - \sin\alpha[y_\nu(c,\nu) + y_t(c,\nu)c_\nu] = 0. \tag{4.7}$$

In order for (4.6) and (4.7) to hold simultaneously (since $\cos^2\alpha + \sin^2\alpha \neq 0$) we must have (abbreviating the notation)

$$x(y_\nu + y_t c_\nu) - y(x_\nu + x_t c_\nu) = 0, \tag{4.8}$$

or

$$\begin{vmatrix} x & y \\ x_\nu & y_\nu \end{vmatrix} + \begin{vmatrix} x & y \\ x_t & y_t \end{vmatrix} c_\nu = 0. \tag{4.9}$$

Hence, using (4.4), we have

$$\frac{dc}{d\nu} = -Q(t,\nu), \tag{4.10}$$

where

$$Q(t,\nu) = \begin{vmatrix} x(t,\nu) & y(t,\nu) \\ x_\nu(t,\nu) & y_\nu(t,\nu) \end{vmatrix}. \tag{4.11}$$

We will show in §5 that $Q(t,\nu)$ satisfies the third-order nonhomogeneous linear differential equation

$$w''' + 4q(t,\nu)w' + 2q_t(t,\nu)w = 2q_\nu(t,\nu) \tag{4.12}$$

with initial conditions

$$w(a,\nu) = w_{tt}(a,\nu) = 0, \qquad w_t(a,\nu) = -2\phi'(\nu)/\phi(\nu). \tag{4.13}$$

It will then follow that

$$Q(t,\nu) = [-2\phi'(\nu)/\phi(\nu)]x(t,\nu)y(t,\nu) + \int_a^t q_\nu(s,\nu)[z(s,t;\nu)]^2 ds, \tag{4.14}$$

where

$$z(t,s;\nu) = \begin{vmatrix} x(t,\nu) & y(t,\nu) \\ x(s,\nu) & y(s,\nu) \end{vmatrix} \tag{4.15}$$

is the solution of

$$\frac{d^2z}{ds^2} + q(s,\nu)z = 0, \qquad z(t,t;\nu) = 0,\ z_s(t,t;\nu) = 1. \tag{4.16}$$

From (4.14) we get

$$\frac{dc}{d\nu} = [2\phi'(\nu)/\phi(\nu)]x(c,\nu)y(c,\nu) - \int_a^c q_\nu(s,\nu)[z(s,c;\nu)]^2 ds. \tag{4.17}$$

In case $\alpha = 0$ or $\alpha = \pi/2$ the first term on the right-hand side vanishes, and so we see that (4.17) gives a generalization of the Richardson formula (2.7).

5. A third-order differential equation. Our approach here is motivated by what was done in [15] in the case of Bessel functions. There a formula of Watson [23, p. 444],

$$J_\nu(z)\partial Y_\nu(z)/\partial\nu - Y_\nu(z)\partial J_\nu(z)/\partial\nu = -4\pi^{-1}\int_0^\infty K_0(2z\sinh t)e^{-2\nu t}dt, \tag{5.1}$$

and its almost immediate consequence,

$$\frac{dc}{d\nu} = 2c\int_0^\infty K_0(2c\sinh t)e^{-2\nu t}dt, \tag{5.2}$$

were proved by showing that both sides of (5.1) satisfy the same third-order differential equation and have the same asymptotic behavior. Here we prove (4.14) by showing that both sides satisfy (4.12) with initial conditions (4.13). We remark that the product xy of any two solutions x and y of (4.1) satisfies the corresponding homogeneous equation

$$w''' + 4q(t,\nu)w' + 2q_t(t,\nu)w = 0; \tag{5.3}$$

see [1]. Thus the general solution of (4.12) is

$$Q(t,\nu) = k_1(\nu)x^2(t,\nu) + k_2(\nu)x(t,\nu)y(t,\nu) + k_3(\nu)y^2(t,\nu) + S(t,\nu), \tag{5.4}$$

where

$$S(t,\nu) = \int_a^t q_\nu(s,\nu)[z(s,t;\nu)]^2 ds \tag{5.5}$$

is the integral on the right-hand side of (4.14). To show that $Q(t,\nu)$ satisfies (4.12) we find it convenient to introduce also the determinant

$$R(t,\nu) = \begin{vmatrix} x_t(t,\nu) & y_t(t,\nu) \\ x_{\nu t}(t,\nu) & y_{\nu t}(t,\nu) \end{vmatrix}. \tag{5.6}$$

Using the well-known rules for differentiation of determinants, we have

$$R_t(t,\nu) = \begin{vmatrix} x_{tt} & y_{tt} \\ x_{\nu t} & y_{\nu t} \end{vmatrix} + \begin{vmatrix} x_t & y_t \\ x_{\nu tt} & y_{\nu tt} \end{vmatrix}. \tag{5.7}$$

Now, using (4.1) and [10, Thm. 3.1, p. 95], which implies that x_ν and y_ν satisfy the differential equation obtained by formally differentiating (4.1) (with the implied interchange of orders of integration), we get

$$\begin{aligned} R_t(t,\nu) &= -q \begin{vmatrix} x & y \\ x_{\nu t} & y_{\nu t} \end{vmatrix} - q_\nu \begin{vmatrix} x_t & y_t \\ x & y \end{vmatrix} - q \begin{vmatrix} x_t & y_t \\ x_\nu & y_\nu \end{vmatrix} \\ &= -qQ_t + q_\nu. \end{aligned}$$

Similarly,

$$\begin{aligned} Q_{tt}(t,\nu) &= \begin{vmatrix} x_{tt} & y_{tt} \\ x_\nu & y_\nu \end{vmatrix} + 2 \begin{vmatrix} x_t & y_t \\ x_{\nu t} & y_{\nu t} \end{vmatrix} + \begin{vmatrix} x & y \\ x_{\nu tt} & y_{\nu tt} \end{vmatrix} \\ &= -q \begin{vmatrix} x & y \\ x_\nu & y_\nu \end{vmatrix} + 2R + \begin{vmatrix} x & y \\ -q_\nu x - q x_\nu & -q_\nu y - q y_\nu \end{vmatrix} \\ &= -2qQ + 2R. \end{aligned} \tag{5.8}$$

Differentiating this last equation with respect to t and using (5.8), we see that $Q(t,\nu)$ satisfies (4.12). As already remarked, the first term on the right-hand side of (4.14) satisfies the homogeneous equation (5.3). Hence in order to show that $Q(t,\nu)$ satisfies (4.12) it will be enough to show that $S(t,\nu)$, as given by (5.5), satisfies (4.12). We remark, first of all, that

$$z(t,s;\nu) = \begin{vmatrix} x(t,\nu) & y(t,\nu) \\ x(s,\nu) & y(s,\nu) \end{vmatrix} \tag{5.9}$$

since both sides satisfy (4.16). We also have $z(t,s;\nu) = -z(s,t;\nu)$, and hence $z(s,t,\nu)$ satisfies

$$\frac{d^2 z}{dt^2} + q(t,\nu)z = 0, \quad z(s,s;\nu) = 0, \quad z_t(s,s;\nu) = 1. \tag{5.10}$$

Now

$$S_t(t,\nu) = 2\int_a^t q_\nu(s,\nu) z(s,t;\nu) z_t(s,t,\nu)\,ds. \tag{5.11}$$

Differentiating again and using (5.10), we get

$$S_{tt}(t,\nu) = -2q(t,\nu)S(t,\nu) + 2\int_a^t q_\nu(s,\nu)[z_t(s,t;\nu)]^2\,ds. \tag{5.12}$$

A final differentiation and use of (5.10) and (5.11) show that $S(t,\nu)$ satisfies (4.12). Though the verification is easy, we remark that in order to *discover* the form of the

right-hand side of (4.14) we used the method of variation of parameters to solve (4.12) by using the linearly independent solutions x^2, xy, and y^2 of (5.3).

It is clear on using the initial conditions (4.2) and (4.3) that $Q(a,\nu) = 0$ and $R(a,\nu) = 0$. We have

$$Q_t(t,\nu) = \begin{vmatrix} x_t & y_t \\ x_\nu & y_\nu \end{vmatrix} + \begin{vmatrix} x & y \\ x_{\nu t} & y_{\nu t} \end{vmatrix}, \tag{5.13}$$

so that

$$Q_t(a,\nu) = -2\phi'(\nu)/\phi(\nu), \tag{5.14}$$

and from (5.8), $Q_{tt}(a,\nu) = 0$. On the other hand, from (5.10), (5.5), and (5.12), $S(a,\nu) = S_t(a,\nu) = S_{tt}(a,\nu) = 0$. It is clear that choosing $k_1(\nu) = k_3(\nu) = 0, k_2(\nu) = -2\phi'(\nu)/\phi(\nu)$ in (5.4) will make $Q(t,\nu)$ satisfy the initial conditions $Q(a,\nu) = Q_{tt}(a,\nu) = 0$, $Q_t(a,\nu) = -2\phi'(\nu)/\phi(\nu)$. Thus the proof of (4.14) is complete.

6. Application of third-order-equation method to cylinder functions. In the case of cylinder functions,

$$C_\nu(t) = \cos\alpha J_\nu(t) - \sin\alpha Y_\nu(t),$$

the relevant differential equation satisfied by $t^{1/2}C_\nu(t)$ is

$$y'' + \left[1 + \frac{1/4 - \nu^2}{t^2}\right] y = 0. \tag{6.1}$$

For the Bessel function of the first kind, $J_\nu(t)$, we get the Schläfli formula of §3. The discussion of §4, however, enables us to extend this result to all cylinder functions, rather than just $J_\nu(t)$. The approach of §3 does not work because the functions no longer vanish at 0, nor do they have a common finite zero (for all ν). We avoid this problem by choosing, in effect, $a = \infty$.

We suppose that $Q(t,\nu)$ is given by (4.11), where

$$x(t,\nu) = -(t\pi/2)^{1/2}Y_\nu(t), \qquad y(t,\nu) = (t\pi/2)^{1/2}J_\nu(t).$$

We can verify easily that in the present situation (4.12) has a particular solution

$$S(t,\nu) = -\int_t^\infty q_\nu(s,\nu) \begin{vmatrix} x(t,\nu) & y(t,\nu) \\ x(s,\nu) & y(s,\nu) \end{vmatrix}^2 ds, \tag{6.2}$$

so that if we use the considerations leading to (5.4) of §5, its general solution is given by (5.4). With the standard notation [23] for Bessel functions this leads to

$$\begin{aligned} Q(t,\nu) &= k_4(\nu)t[J_\nu^2(t) + Y_\nu^2(t)] + k_5(\nu)t[H_\nu^{(1)}(t)]^2 \\ &\quad + k_6(\nu)t[H_\nu^{(2)}(t)]^2 + (\nu\pi^2/2)t\int_t^\infty s^{-1} \begin{vmatrix} J_\nu(t) & Y_\nu(t) \\ J_\nu(s) & Y_\nu(s) \end{vmatrix}^2 ds. \end{aligned} \tag{6.3}$$

Now from standard asymptotic expansions for the Bessel functions ([23, Chap. 7] or [8, eq. 7.13]) we have (see [17, p. 341] and [15])

$$Q(t,\nu) = -(\pi/2) + O(t^{-1}), \tag{6.4}$$

$$J_\nu^2(t) + Y_\nu^2(t) = (2/\pi)t^{-1} + O(t^{-2}), \tag{6.5}$$

$$t[H_\nu^{(1,2)}(t)]^2 \sim (2/\pi)e^{\pm 2i(t-\nu\pi/2-\pi/4)} \tag{6.6}$$

as $t \to \infty$. It is also clear that

$$\int_t^\infty s^{-1} \begin{vmatrix} J_\nu(t) & Y_\nu(t) \\ J_\nu(s) & Y_\nu(s) \end{vmatrix}^2 ds = O(t^{-2}), \qquad t \to \infty, \tag{6.7}$$

since [23, Chap. 7]

$$\left|t^{1/2}J_\nu(t)\right|, \quad \left|t^{1/2}Y_\nu(t)\right| = O(1), \quad t \to \infty. \tag{6.8}$$

Using these asymptotic estimates in (6.3), we conclude that $k_4(\nu) = -\pi^2/4$, $k_5(\nu) = k_6(\nu) = 0$. Thus we have

$$Q(t,\nu) = -\frac{\pi^2 t}{4}[J_\nu^2(t) + Y_\nu^2(t)] + \frac{\nu\pi^2 t}{2}\int_t^\infty s^{-1} \begin{vmatrix} J_\nu(t) & Y_\nu(t) \\ J_\nu(s) & Y_\nu(s) \end{vmatrix}^2 ds. \tag{6.9}$$

If $c = c(\nu, k, \alpha)$ is a zero of any cylinder function $C_\nu(t) = \cos\alpha J_\nu(t) - \sin\alpha Y_\nu(t)$, we get from (4.10)

$$\frac{dc}{d\nu} = \frac{\pi^2 c}{4}\left[J_\nu^2(c) + Y_\nu^2(c) - 2\nu\int_c^\infty s^{-1} \begin{vmatrix} J_\nu(c) & Y_\nu(c) \\ J_\nu(s) & Y_\nu(s) \end{vmatrix}^2 ds\right]. \tag{6.10}$$

But the determinant in the integrand here is a solution of the Bessel equation (in the variable s), which vanishes at c. Hence it is a constant multiple of $C_\nu(s)$. Differentiating with respect to s, setting $s = c$, and using the Wronskian formula [23, p. 76], we find that this constant is $2/[\pi c C_\nu'(c)]$. Hence (6.10) can be written

$$\frac{dc}{d\nu} = \frac{\pi^2 c}{4}\left[J_\nu^2(c) + Y_\nu^2(c)\right] - \frac{2\nu}{cC_\nu'^2(c)}\int_c^\infty s^{-1}C_\nu^2(s)\,ds. \tag{6.11}$$

In the special case where $\alpha = 0$ and $c = j = j_{\nu k}$, a positive zero of $J_\nu(t)$, this becomes

$$\frac{dj}{d\nu} = \frac{\pi^2 j Y_\nu^2(j)}{4}\left[1 - 2\nu\int_j^\infty s^{-1}J_\nu^2(s)\,ds\right]. \tag{6.12}$$

Using the formula [23]

$$\int_0^\infty s^{-1}J_\nu^2(s)\,ds = 1/(2\nu), \tag{6.13}$$

as well as the Wronskian relation and a recurrence relation for the Bessel function, we see that (6.10) reduces to Schläfli's formula (3.2). Similarly, in the case $\alpha = \pi/2$ we are led to Schafheitlin's formula [19]

$$\frac{dy}{d\nu} = \frac{\pi^2 y J_\nu^2(y)}{4}\left[1 - 2\nu\int_y^\infty s^{-1}Y_\nu^2(s)\,ds\right] \tag{6.14}$$

for the zeros of $Y_\nu(t)$.

Remark 1. A formula that is somewhat more useful than (3.2) for the derivative with respect to order of a zero of a Bessel function is (5.2) (due to Watson [23, p. 508]), which is valid for all zeros of cylinder functions throughout the interval in which they are continuous functions of ν. Because of the simple nature (positive, decreasing, etc.) of $K_0(t)$, the formula (5.2) has been used to remarkable effect in several discussions of monotonicity, convexity, etc., of the zeros; see [2]–[5] and the references therein. In [15] it was shown how to derive (5.2) by a differential-equations method essentially by showing that the corresponding Q (in the notation of the current paper) satisfies (4.12) in the special case involved. It would be interesting to be able to do this for a class of differential equations, that is, to find a method for solving (4.12) that would lead to formulas like (5.2) in much the same way as variation of parameters leads to formulas of the Schläfli type; see [16] for further remarks on this topic.

In contrast to (5.2), there are many approaches to the Schläfli formula (3.2). In addition to the approach in this section, there is an approach based on the Hellmann–Feynman theorem in [14]. A formula for $dj/d\nu$ ([11]; see also [13]) involving infinite sums has been shown [12] to be derivable from the Schläfli formula by using classical results for Bessel functions.

Remark 2. Since we know independently from (5.2) that $dc/d\nu > 0$ and since every positive number can be realized as a zero of some $C_\nu(t)$, we find from (6.10) the inequality

$$J_\nu^2(t) + Y_\nu^2(t) > 2\nu \int_t^\infty s^{-1} \left| \begin{array}{cc} J_\nu(t) & Y_\nu(t) \\ J_\nu(s) & Y_\nu(s) \end{array} \right|^2 ds, \qquad t > 0. \tag{6.15}$$

Acknowledgments. We thank the referees for their constructive comments. We are also grateful to Angelo Mingarelli for drawing our attention to the work of Richardson [18].

REFERENCES

[1] P. APPELL, *Sur les transformations des équations différentielles linéaires*, C. R. Acad. Sci. Paris, 91 (1880), pp. 211–214.

[2] Á. ELBERT, *Concavity of the zeros of Bessel functions*, Studia Sci. Math. Hungar., 12 (1977), pp. 81–88.

[3] Á. ELBERT AND A. LAFORGIA, *On the square of the zeros of Bessel functions*, SIAM J. Math. Anal., 15 (1984), pp. 206–212.

[4] ———, *On the convexity of the zeros of Bessel functions*, SIAM J. Math. Anal., 16 (1985), pp. 614–619.

[5] ———, *Further results on the zeros of Bessel functions*, Analysis, 5 (1986), pp. 71–86.

[6] Á. ELBERT AND M. E. MULDOON, *Inequalities for the zeros of Hermite functions*, in preparation.

[7] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[8] ———, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1954.

[9] A. LAFORGIA AND M. E. MULDOON, *Monotonicity of zeros of generalized Airy functions*, Z. Angew. Math. Phys., 39 (1988), pp. 267–271.

[10] P. HARTMAN, *Ordinary Differential Equations*, third ed., Birkhäuser Verlag, Basel, Switzerland, 1982.

[11] E. K. IFANTIS AND P. D. SIAFARIKAS, *A differential equation for the zeros of Bessel functions*, Appl. Anal., 20 (1985), pp. 269–281.

[12] M. E. H. ISMAIL, *Zeros of Bessel functions*, Appl. Anal., 22 (1986), pp. 167–168.

[13] M. E. H. ISMAIL AND M. E. MULDOON, *On the variation with respect to a parameter of zeros of Bessel and q-Bessel functions*, J. Math. Anal. Appl., 135 (1988), pp. 187–207.

[14] J. T. LEWIS AND M. E. MULDOON, *Monotonicity and convexity properties of zeros of Bessel functions*, SIAM J. Math. Anal., 8 (1977), pp. 171–178.

[15] M. E. MULDOON, *A differential equations proof of a Nicholson-type formula*, Z. Angew. Math. Mech., 61 (1981), pp. 598–599.

[16] ———, *On the zeros of some special functions: Differential equations and Nicholson-type formulas*, Lecture Notes in Mathematics, 1192, Springer-Verlag, New York, 1986, pp. 155–160.

[17] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[18] R. G. D. RICHARDSON, *Contribution to the study of oscillation properties of the solutions of linear differential equations of the second order*, Amer. J. Math., 40 (1918), pp. 283–316.

[19] P. SCHAFHEITLIN, *Die Lage der Nullstellen der Besselschen Funktionen zweiter Art*, Sitzungsber. Berl. Math. Ges., 5 (1906), pp. 82–93.

[20] L. SCHLÄFLI, *Über die Convergenz der Entwicklung einer arbiträren Function $f(x)$ nach den Bessel'schen Functionen $J^{\alpha}(\beta_1 x)$, $J^{\alpha}(\beta_2 x)$, $J^{\alpha}(\beta_3 x), \ldots$, wo β_1, β_2, β_3, ... die positiven Wurzeln der Gleichung $J^{\alpha}(\beta)$ vorstellen*, Math. Ann., 10 (1876), pp. 137–142.

[21] C. STURM, *Mémoire sur les équations différentielles linéaires du second ordre*, J. Math. Pures Appl., 1 (1836), pp. 106–186.

[22] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, Vol. 23, 4th ed., American Mathematicsl Society, Providence, RI, 1975.

[23] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, London, 1944.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 365–383, March 1994

008

REMEZ- AND NIKOLSKII-TYPE INEQUALITIES FOR LOGARITHMIC POTENTIALS*

TAMÁS ERDÉLYI†, XIN LI‡, AND E. B. SAFF§

Abstract. Remez- and Nikolskii-type inequalities on line segments, on circles, and on certain bounded domains of the complex plane are established for exponentials of logarithmic potentials with respect to probability measures on $\mathbb{C}$ having compact support.

Key words. potentials, Remez inequalities, Nikolskii inequalities, logarithmic capacity, Green function

AMS subject classifications. 41A17, 31A15

1. Introduction and notation.

Generalized nonnegative polynomials of the form

$$f(z) = |\omega| \prod_{j=1}^{k} |z - z_j|^{r_j} \qquad (\omega \in \mathbb{C},\ z_j \in \mathbb{C},\ 0 < r_j \in \mathbb{R},\ j = 1, 2, \dots k) \tag{1.1}$$

were studied in a sequence of recent papers [1], [2], [3], [4], [6], [7]. Several important polynomial inequalities were extended to this class by utilizing the generalized degree

$$N := \sum_{j=1}^{k} r_j \tag{1.2}$$

of f in place of the ordinary one. Since

$$\log f(z) = \sum_{j=1}^{k} r_j \log |z - z_j| + \log |\omega|, \tag{1.3}$$

a generalized nonnegative polynomial can be considered as a constant times the exponential of a logarithmic potential with respect to a finite Borel measure on $\mathbb{C}$ that is supported in finitely many points (the measure has mass $r_j > 0$ at each $z_j, j = 1, 2, \dots, k$). This suggests that some of the inequalities holding for generalized nonnegative polynomials may be true for exponentials of logarithmic potentials of the form

$$Q_{\mu,c}(z) = \exp\left(\int_{\mathbb{C}} \log |z - t| d\mu(t) + c \right), \tag{1.4}$$

where μ is a finite, nonnegative Borel measure on $\mathbb{C}$ having compact support and $c \in \mathbb{R}$. The quantity $\mu(\mathbb{C})$ plays the role of the generalized degree N, defined by (1.2). In this paper we extend a number of classical polynomial inequalities for exponentials

* Received by the editors July 27, 1992; accepted for publication (in revised form) February 8, 1993.

† Department of Mathematics, Ohio State University, 231 West Eighteenth Avenue, Columbus, Ohio 43210. This author's research was supported in part by National Science Foundation grant DMS-902-4901.

‡ Department of Mathematics, University of Central Florida, Orlando, Florida 32816.

§ Institute for Constructive Mathematics, Department of Mathematics, University of South Florida, Tampa, Florida 33620. This author's research was supported in part by National Science Foundation grants DMS-881-4026 and DMS-891-2423.

of logarithmic potentials. Typically such extensions are not straightforward; indeed our proofs are far from simple density arguments.

Denote by $\mathcal{P}_n^r$ the set of all algebraic polynomials of degree at most n with real coefficients and let $\mathcal{P}_n^c$ be the set of all algebraic polynomials of degree at most n with complex coefficients. Let $\mathcal{M}$ denote the set of all probability measures on $\mathbb{C}$ with compact support. For $\mu \in \mathcal{M}$ and $c \in \mathbb{R}$ we define

$$P_{\mu,c}(z) := \int_{\mathbb{C}} \log|z-t| d\mu(t) + c \qquad (z \in \mathbb{C}) \tag{1.5}$$

and

$$Q_{\mu,c}(z) := \exp(P_{\mu,c}(z)) \qquad (z \in \mathbb{C}). \tag{1.6}$$

Associated with $\mu \in \mathcal{M}$ and $c \in \mathbb{R}$ we introduce the sets

$$\begin{aligned} E_{\mu,c} :&= \{x \in [-1,1] : \ P_{\mu,c}(x) \le 0\} \\ &= \{x \in [-1,1] : \ Q_{\mu,c}(x) \le 1\}. \end{aligned} \tag{1.7}$$

We will denote by $m_1(A)$ and $m_2(B)$ the one-dimensional Lebesgue measure of a set $A \subset \mathbb{R}$, and the two-dimensional Lebesgue measure of a set $B \subset \mathbb{C}$, respectively.

The Remez inequality [12] asserts that

$$\max_{-1\le x\le 1} |p(x)| \le T_n\left(\frac{2+s}{2-s}\right) \tag{1.8}$$

for every $p \in P_n^r$ such that

$$m_1(\{x \in [-1,1] : \ |p(x)| \le 1\}) \ge 2-s \qquad (0 < s < 2), \tag{1.9}$$

where T_n is the Chebyshev polynomial of degree n, defined by $T_n(x) := \cos n\theta, x = \cos\theta$. Proofs of this inequality appear in [9, pp. 119–121] and [5]. In Theorem 2.1 we establish a sharp upper bound for $\max_{-1\le x\le 1} Q_{\mu,c}(x)$ when $m_1(E_{\mu,c}) \ge 2-s$, and (assuming $Q_{\mu,c}(x)$ is continuous) we find all $\mu \in \mathcal{M}$ and $c \in \mathbb{R}$ with $m_1(E_{\mu,c}) \ge 2-s$ for which this sharp upper bound is achieved.

In Theorem 2.2 we establish pointwise upper bounds for $Q_{\mu,c}(x)$ for fixed $x \in [-1,1]$, if $\mu \in \mathcal{M}$, $c \in \mathbb{R}$, and $m_1(E_{\mu,c}) \ge 2-s$. An obvious bound for $Q_{\mu,c}(x)$ follows immediately from Theorem 2.1, but it turns out that for any fixed $-1 < x < 1$ this can be substantially improved. Indeed, Theorem 2.2 establishes essentially sharp upper bounds, which extend the validity of a pointwise Remez-type inequality [4, Thm. 4] proved for generalized nonnegative polynomials.

In Corollary 2.3 we offer another, slightly weaker version of the Remez-type inequality of Theorem 2.1, and in Theorem 2.4 we establish an analogue of Corollary 2.3, where the interval $[-1,1]$ is replaced by the closure of a bounded domain $\Omega \subset \mathbb{C}$ with C^2 boundary, and the one-dimensional Lebesgue measure m_1 is replaced by m_2. Such two-dimensional Remez-type inequalities seem to be new even for ordinary polynomials and for special domains, such as the open unit disk. Therefore we formulate a two-dimensional Remez-type inequality in this special case first, which turns out to be essentially sharp by Theorem 2.6.

Concerning L_p-versions of Remez's inequality, we study the following question: How large can the ratio

$$\frac{\int_{-1}^{1}(Q_{\mu,c}(x))^p dx}{\int_{A}(Q_{\mu,c}(x))^p dx}$$

be if $\mu \in \mathcal{M}$, $c \in \mathbb{R}$, $A \subset [-1,1]$, $m_1(A) \geq 2-s$, $0 < s < 2$, and $p > 0$? We give an essentially sharp answer in Theorem 2.7 for the case when $0 < s \leq \frac{1}{2}$. In Theorem 2.8 we establish an essentially sharp upper bound for the ratio

$$\frac{\int_{\Omega}(Q_{\mu,c}(z))^p dm_2(z)}{\int_{A}(Q_{\mu,c}(z))^p dm_2(z)},$$

when $\Omega \subset \mathbb{C}$ is a bounded domain with C^2 boundary, $\mu \in \mathcal{M}$, $c \in \mathbb{R}$, $A \subset \overline{\Omega}$, $m_2(A) \geq m_2(\Omega) - s, s > 0$ sufficiently small, and $p > 0$. In Theorems 2.9 and 2.10 we give essentially the best possible Remez-type inequalities for exponentials of logarithmic potentials on the unit circle. The Remez-type inequalities of Corollary 2.3 and Theorems 2.4 and 2.9 will play a central role in establishing the Nikolskii-type inequalities for exponentials of potentials on $[-1,1]$, on the unit circle and on bounded domains of $\mathbb{C}$ with C^2 boundary. These Nikolskii-type inequalities are formulated in Theorems 3.1, 3.2, and 3.3.

2. Remez-type inequalities: statement of results. In this section we state our main results concerning Remez-type inequalities for logarithmic potentials on $[-1,1]$, on the unit circle and on bounded domains having smooth boundaries. The proofs of these results will be given in §§6, 7, 8, and 9.

THEOREM 2.1. *Let $\mu \in \mathcal{M}, c \in \mathbb{R}$, and $E_{\mu,c}$ be defined as in* (1.7). *Then*

$$m_1(E_{\mu,c}) \geq 2-s \qquad (0 < s < 2) \tag{2.1}$$

implies

$$\max_{-1\leq x\leq 1} Q_{\mu,c}(x) \leq \frac{\sqrt{2}+\sqrt{s}}{\sqrt{2}-\sqrt{s}}. \tag{2.2}$$

Furthermore, if $Q_{\mu,c}$ restricted to $[-1,1]$ is continuous on $[-1,1]$, then the equality holds in (2.2) *if and only if*

$$\mu = \mu^*_{[-1,1-s]} \quad \text{or} \quad \mu = \mu^*_{[-1+s,1]}$$

and

$$c = -\log\frac{2-s}{4},$$

where μ^*_K denotes the equilibrium measure (cf. [14, §III.2]*) of a compact set $K \subset \mathbb{C}$.*

We remark that $Q_{\mu,c}$ is upper semicontinuous on $\mathbb{C}$, so the maximum on $[-1,1]$ is attained.

Concerning pointwise upper bounds for $Q_{\mu,c}(x)$ we shall prove the following result that extends the validity of Theorem 4 of [4].

THEOREM 2.2. *There is an absolute constant k_1 such that*

$$Q_{\mu,c}(x) \le \exp\left(k_1 \min\left\{\frac{s}{\sqrt{1-x^2}}, \sqrt{s}\right\}\right) \tag{2.3}$$

for every $-1 \le x \le 1, \mu \in \mathcal{M}$ and $c \in \mathbb{R}$ satisfying

$$m_1(E_{\mu,c}) \ge 2 - s \qquad (0 < s \le 1). \tag{2.4}$$

Here we do not examine what happens when $1 < s < 2$; the case $0 < s \le 1$ is more important in applications. The sharpness of Theorem 2.2 (in the corresponding polynomial case) is shown in [4, §12].

Observe that the first assertion of Theorem 2.1 is equivalent to the following.

THEOREM 2.1*. *For every $\mu \in \mathcal{M}, c \in \mathbb{R}$ and $0 < t < 1$,*

$$m_1\left(\left\{x \in [-1,1] : Q_{\mu,c}(x) > \frac{1-\sqrt{t}}{1+\sqrt{t}} \max_{-1\le y\le 1} Q_{\mu,c}(y)\right\}\right) \ge 2t. \tag{2.5}$$

Consequently, we obtain the following.

COROLLARY 2.3. *There is an absolute constant $k_2 > 0$ such that*

$$m_1\left(\left\{x \in [-1,1] : Q_{\mu,c}(x) > \exp(-\sqrt{s}) \max_{-1\le y\le 1} Q_{\mu,c}(y)\right\}\right) \ge k_2 s \tag{2.6}$$

for every $\mu \in \mathcal{M}$, $c \in \mathbb{R}$, and $0 < s < 2$.

In our next theorem we establish the analogue of Corollary 2.3 for the case when $[-1,1]$ is replaced by the closure of a bounded domain $\Omega \subset \mathbb{C}$ with C^2 boundary.

THEOREM 2.4. *Let $\Omega \subset \mathbb{C}$ be a bounded domain with C^2 boundary. Then, there is a constant $k_3 = k_3(\Omega) > 0$ depending only on Ω such that*

$$m_2\left(\left\{z \in \overline{\Omega} : Q_{\mu,c}(z) > \exp(-\sqrt{s}) \max_{w\in\overline{\Omega}} Q_{\mu,c}(w)\right\}\right) \ge k_3 s \tag{2.7}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}$, and $0 < s < m_2(\Omega)$.

Actually, in the above theorem it suffices to have a somewhat weaker geometric assumption for the boundary of Ω, namely, the following: there is an $r > 0$ depending only on Ω such that for each $z \in \partial\Omega$ there is an open disk D_z with radius r such that $D_z \subset \Omega$ and $\overline{D}_z \cap \partial\Omega = \{z\}$. It is well known that if $\partial\Omega$ is a C^2 curve, then this property holds.

To prove Theorem 2.4 we will need the following result for polynomials. To formulate this, we introduce the notation

$$D := \{z \in \mathbb{C} : |z| < 1\} \tag{2.8}$$

and

$$\mathcal{P}_n^c(\overline{D}, s) := \{p \in \mathcal{P}_n^c : m_2(\{z \in \overline{D} : |p(z)| \le 1\}) \ge \pi - s\} \qquad (0 < s < \pi), \tag{2.9}$$

along with the analogous definition for $\mathcal{P}_n^r(\overline{D}, s)$.

THEOREM 2.5. *There is an absolute constant $k_4 > 0$ such that*

$$\max_{u\in\overline{D}} |p(u)| \le \exp(k_4 n\sqrt{s}) \tag{2.10}$$

for every $p \in \mathcal{P}_n^c(\overline{D}, s)$ and $0 < s \le \frac{1}{4}$.

Our next theorem shows that the result of Theorem 2.5 is essentially sharp.

THEOREM 2.6. *There is an absolute constant $k_5 > 0$ such that*

$$\sup(\{|p(1)| : p \in \mathcal{P}_{3n}^r(\overline{D}, s)\}) \geq \exp(k_5 n\sqrt{s}) \tag{2.11}$$

for every $0 < s \leq \frac{1}{2}$.

Using Theorems 2.1 and 2.4, we establish Remez-type inequalities in $L_p(0 < p < \infty)$ for exponentials of potentials on both $[-1, 1]$ and bounded domains $\Omega \subset \mathbb{C}$ and C^2 boundary.

THEOREM 2.7. *There is an absolute constant $k_6 > 0$ such that*

$$\begin{aligned} \int_{-1}^{1} (Q_{\mu,c}(x))^p dx &\leq \left(1 + \left(\frac{1+\sqrt{s}}{1-\sqrt{s}}\right)^p\right) \int_A (Q_{\mu,c}(x))^p dx \\ &\leq (1 + \exp(k_6 p\sqrt{s})) \int_A (Q_{\mu,c}(x))^p dx \end{aligned} \tag{2.12}$$

for every $\mu \in \mathcal{M}$, $c \in \mathbb{R}$, $p > 0$, $0 < s \leq \frac{1}{2}$, and $A \subset [-1, 1]$ with $m_1(A) \geq 2 - s$. If $0 < s \leq \frac{1}{4}$, then $k_6 = 4$ is a suitable choice.

THEOREM 2.8. *Let $\Omega \subset \mathbb{C}$ be a bounded domain with C^2 boundary. Then there are constants $0 < k_7 = k_7(\Omega)$ and $0 < k_8 = k_8(\Omega)$ depending only on Ω such that*

$$\int_{\Omega} (Q_{\mu,c}(z))^p dm_2(z) \leq (1 + \exp(k_7 p\sqrt{s})) \int_A (Q_{\mu,c}(z))^p dm_2(z) \tag{2.13}$$

for every $\mu \in \mathcal{M}$, $c \in \mathbb{R}$, $p > 0$, $0 < s \leq k_8$, and $A \subset \overline{\Omega}$ with $m_2(A) \geq m_2(\Omega) - s$.

The following theorem establishes a Remez-type inequality for exponentials of logarithmic potentials on the unit circle, extending a Remez-type inequality for trigonometric polynomials [4, Thm. 3].

THEOREM 2.9. *There is an absolute constant $k_9 > 0$ such that*

$$\max_{-\pi \leq t \leq \pi} Q_{\mu,c}(e^{it}) \leq \exp(k_9 s) \tag{2.14}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}$ and $0 < s \leq \pi/2$ whenever

$$m_1(\{t \in [-\pi, \pi) : Q_{\mu,c}(e^{it}) \leq 1\}) \geq 2\pi - s. \tag{2.15}$$

From this we will easily obtain the following.

THEOREM 2.10. *We have*

$$\int_{-\pi}^{\pi} (Q_{\mu,c}(e^{it}))^p dt \leq (1 + \exp(2k_9 ps)) \int_A (Q_{\mu,c}(e^{it}))^p dt$$

for every $\mu \in \mathcal{M}$, $c \in \mathbb{R}$, $p > 0$, $0 < s \leq \pi/4$, and $A \subset [-\pi, \pi)$ with $m_1(A) \geq 2\pi - s$. Here k_9 is the same as in Theorem 2.9.

We have formulated each of our results for *probability* measures on $\mathbb{C}$ with compact support. This was done only for the sake of brevity. As an example, we rewrite the result of Theorem 2.1 for all finite Borel measures on $\mathbb{C}$ with compact support.

COROLLARY 2.11. *Let μ be a finite Borel measure on $\mathbb{C}$ with compact support, and let $c \in \mathbb{R}$ and $E_{\mu,c}$ be defined as in (1.7). Then*

$$m_1(E_{\mu,c}) \geq 2 - s \qquad (0 < s < 2) \tag{2.16}$$

implies

$$\max_{-1\le x\le 1} Q_{\mu,c}(x) \le \left(\frac{\sqrt{2}+\sqrt{s}}{\sqrt{2}-\sqrt{s}}\right)^{\mu(\mathbb{C})}. \tag{2.17}$$

Furthermore, if $Q_{\mu,c}$ restricted to $[-1,1]$ is continuous, then equality holds in (2.17) *if and only if*

$$\mu = \mu(\mathbb{C})\mu^*_{[-1,1-s]} \quad \text{or} \quad \mu = \mu(\mathbb{C})\mu^*_{[-1+s,1]}$$

and

$$c = -\mu(\mathbb{C}) \log \frac{2-s}{4},$$

where μ^*_K denotes the equilibrium measure of a compact set $K \subset \mathbb{C}$.

3. Nikolskii-type inequalities: statement of results. Using Corollary 2.3 and Theorem 2.4 we will prove the following Nikolskii-type inequalities. The proofs will be given in §10.

THEOREM 3.1. *There is an absolute constant $k_{10} > 0$ such that*

$$\|Q_{\mu,c}\|_{L_p(-1,1)} \le (k_{10}(1+q^2))^{1/q-1/p} \|Q_{\mu,c}\|_{L_q(-1,1)} \tag{3.1}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}$, and $0 < q < p \le \infty$.

THEOREM 3.2. *Let $\Omega \subset \mathbb{C}$ be a bounded domain with C^2 boundary. There exists a constant $k_{11} = k_{11}(\Omega) > 0$ depending only on Ω such that*

$$\|Q_{\mu,c}\|_{L_p(\Omega)} \le (k_{11}(1+q^2))^{1/q-1/p} \|Q_{\mu,c}\|_{L_q(\Omega)} \tag{3.2}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}$, and $0 < q < p \le \infty$.

We remark that Theorem 3.1 is an extension of [7, Thm. 6], where the same inequality was proved when the support of μ is a finite set.

THEOREM 3.3. *There is an absolute constant $k_{12} > 0$ such that*

$$\|Q_{\mu,c}(e^{it})\|_{L_p(-\pi,\pi)} \le (k_{12}(1+q))^{1/q-1/p} \|Q_{\mu,c}(e^{it})\|_{L_q(-\pi,\pi)} \tag{3.3}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}$, and $0 < q < p \le \infty$.

For general finite measures μ, Theorem 3.1 yields the following.

COROLLARY 3.4. *There is an absolute constant $k_{10} > 0$ such that*

$$\|Q_{\mu,c}\|_{L_p(-1,1)} \le (k_{10}(1+(q\mu(\mathbb{C}))^2))^{1/q-1/p} \|Q_{\mu,c}\|_{L_q(-1,1)} \tag{3.4}$$

for every finite Borel measure μ on C with compact support, $c \in \mathbb{R}$ and $0 < q < p \le \infty$.

Theorems 3.2 and 3.3 have similar straightforward extensions.

4. Lemmas for Theorem 2.1. To prove Theorem 2.1 we need a series of lemmas, which we state in this section and prove in §5.

For a compact set $K \subset \mathbb{C}$ containing infinitely many points, let $T_{n,K} \in \mathcal{P}^c_n$ be the nth degree monic Chebyshev polynomial with respect to K, i.e.,

$$\|T_{n,K}\|_K = \inf_{p \in \mathcal{P}^c_{n-1}} \|z^n - p(z)\|_K, \tag{4.1}$$

where $\|\cdot\|_K$ denotes the uniform norm on K. We also define the normalized Chebyshev polynomials

$$\widehat{T}_{n,K} := \frac{T_{n,K}}{\|T_{n,K}\|_K}. \tag{4.2}$$

LEMMA 4.1. *Let $0 \le \delta < 2, \delta < s < 2$, and $z \in \mathbb{C}$ with $\operatorname{Re} z \ge 1 - \delta$ fixed. Then*

$$\sup |p(z)| = |\widehat{T}_{n,[-1,1-s]}(z)|, \tag{4.3}$$

where the supremum in (4.3) is taken over all $p \in \mathcal{P}_n^r$ satisfying

$$m_1(\{x \in [-1, 1-\delta] : |p(x)| \le 1\}) \ge 2 - s. \tag{4.4}$$

If $K \subset \mathbb{C}$ is a compact set we denote by $D_\infty(K)$ the unbounded component of the complement $\mathbb{C}\backslash K$. This domain is referred to as the *outer domain* of K and its boundary $\partial D_\infty(K)$ is called the *outer boundary* of K. If K has positive logarithmic capacity [14, p. 55], we denote by $g_{D_\infty(K)}(z, \infty)$ the Green function with pole at ∞ for $D_\infty(K)$. We remark that $g_{D_\infty(K)}(z, \infty)$ is the smallest positive harmonic function in $D_\infty(K)\backslash\{\infty\}$ that behaves like $\log |z| +$ const. near ∞ (cf. [11, p. 333]).

LEMMA 4.2. *Let $K \subset [-1, 1]$ be compact with $m_1(K) \ge 2 - s$ $(0 < s < 2)$. Then the inequality*

$$g_{D_\infty(K)}(z, \infty) \le g_{D_\infty([-1,1-s])}(z, \infty) \tag{4.5}$$

holds for all z such that $\operatorname{Re} z \ge \sup(K)$.

To prove Lemma 4.2 we need the following result of Myrberg and Lega [11, Thm. 11.1, p. 333].

LEMMA 4.3. *Let $K \subset \mathbb{C}$ be compact with* $\operatorname{cap}(K) > 0$, *where* "cap" *denotes the logarithmic capacity. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{|F_{n,K}(z)|}{\|F_{n,K}\|_K} = g_{D_\infty(K)}(z, \infty) \tag{4.6}$$

for every $z \in D_\infty(K)$, where $F_{n,K}$ denotes an nth *degree monic Fekete polynomial for K. The convergence in (4.6) is locally uniform in $D_\infty(K)$.*

LEMMA 4.4. *Let $\mu \in \mathcal{M}$ and $c \in \mathbb{R}$. If*

$$m_1(E_{\mu,c}) \ge 2 - s \qquad (0 < s < 2), \tag{4.7}$$

then the inequality

$$P_{\mu,c}(z) \le g_{D_\infty([-1,1-s])}(z, \infty) \tag{4.8}$$

holds for all z such that $\operatorname{Re} z \ge \sup(E_{\mu,c})$.

To formulate our last lemma, for $0 < s < 2$ we introduce the notation

$$\mathcal{M}(s) := \{(\mu, c) \in \mathcal{M} \times \mathbb{R} : m_1(E_{\mu,c}) \ge 2 - s\}. \tag{4.9}$$

LEMMA 4.5. *Let $0 < s < 2$ be fixed. Then*

$$\sup\left(\max_{-1\le x\le 1} P_{\mu,c}(x)\right) = \sup(P_{\mu,c}(1)), \tag{4.10}$$

where the supremum on each side of (4.10) is taken over all $(\mu, c) \in \mathcal{M}(s)$.

5. Proofs of the lemmas for Theorem 2.1.

Proof of Lemma 4.1. We prove the lemma only when $\delta = 0$, since the case when $0 < \delta < 2$ can be handled similarly. For the sake of brevity we introduce the classes

$$\mathcal{P}_n^r([-1,1],s) := \{p \in \mathcal{P}_n^r : m_1(\{x \in [-1,1] : |p(x)| \le 1\}) \ge 2 - s\} \quad (n = 0,1,2,\dots;\quad 0 < s < 2). \tag{5.1}$$

It is easy to see that $\mathcal{P}_n^r([-1,1],s)$ is a closed and bounded subset of $\mathcal{P}_n^r$ in the uniform norm on $[-1,1]$; hence it is compact. If $z \in \mathbb{C}$ is fixed, then the map $p \to |p(z)|$ is continuous; therefore, there exists a $p^* \in \mathcal{P}_n^r([-1,1],s)$ such that

$$|p^*(z)| = \sup_{p \in \mathcal{P}_n^r([-1,1],s)} |p(z)|. \tag{5.2}$$

Now we show that $\operatorname{Re} z \ge 1$ implies

$$p^* = \pm \widehat{T}_{n,[-1,1-s]}. \tag{5.3}$$

To see this, we analyze the properties of p^*.

PROPOSITION 5.1. *p^* has only real zeros.*

Proof. Assume to the contrary that p^* has a nonreal zero w. Then

$$p(x) = (1+\eta)p^*(x)\left(1 - \frac{\varepsilon(x-z)(x-\overline{z})}{(x-w)(x-\overline{w})}\right) \in \mathcal{P}_n^r([-1,1],s) \tag{5.4}$$

with sufficiently small $\eta > 0$ and $\varepsilon > 0$ contradicts the maximality of p^*. This proves the proposition. □

PROPOSITION 5.2. *All zeros of p^* are in $[-1,1]$.*

Proof. Assume to the contrary that p^* has a nonreal zero w outside $[-1,1]$, which is real by Proposition 5.1. We now distinguish three cases.

Case 1. $w > \operatorname{Re} z$. Let $w^* \in \mathbb{R}$ be the symmetric image of w with respect to $\operatorname{Re} z$, i.e., $w^* := 2\operatorname{Re} z - w$. Then

$$p(x) := (1+\eta)p^*(x)\frac{x-w^*}{x-w} \in \mathcal{P}_n^r([-1,1],s) \tag{5.5}$$

with a sufficiently small $\eta > 0$ contradicts the maximality of p^*.

Case 2. $1 < w \le \operatorname{Re} z$. Now

$$p(x) := (1+\eta)p^*(x)\frac{x-1}{x-w} \in \mathcal{P}_n^r([-1,1],s) \tag{5.6}$$

with a sufficiently small $\eta > 0$ contradicts the maximality of p^*.

Case 3. $w < -1$. Observe that $(x+1)(x-w)^{-1}$ is strictly increasing in $[-1,\infty)$, and so

$$\left|\frac{\operatorname{Re} z + 1}{\operatorname{Re} z - w}\right| \le \left|\frac{z+1}{z-w}\right|. \tag{5.7}$$

Then

$$p(x) := (1+\eta)p^*(x)\frac{(x+1)(\operatorname{Re} z - w)}{(x-w)(\operatorname{Re} z + 1)} \in \mathcal{P}_n^r([-1,1],s) \tag{5.8}$$

with a sufficiently small $\eta > 0$ contradicts the maximality of p^*.

By considering Cases 1, 2, and 3, Proposition 5.2 is completely proved. □

Now we introduce the notation

$$I := \{x \in [-1,1] : |p^*(x)| \le 1\}. \tag{5.9}$$

Obviously I is the union of pairwise disjoint subintervals of $[-1,1]$ that will be called the components of I. Every component of I contains at least one zero of p^*; otherwise a routine application of Rolle's Theorem, together with Propositions 5.1 and 5.2 would imply that p'^* has at least as many zeros as p^*, a contradiction. Using this observation we prove the following.

PROPOSITION 5.3. *The set I is a single interval; in fact, $I = [-1, 1-s]$.*

Proof. To see that I is an interval, assume to the contrary that I has at least two components, and let I_1 be the component closest to 1. Let η and η' be the left-hand endpoint of I_1 and the right-hand endpoint of the component closest to I_1, respectively. If $w_j (j = 1, 2, \dots, m)$ are the zeros of p^* lying in I_1, then it is easy to check that

$$p(x) := p^*(x) \frac{\prod_{j=1}^{m} (x - w_j + h)}{\prod_{j=1}^{m} (x - w_j)} \in \mathcal{P}_n^r([-1,1], s) \tag{5.10}$$

with $0 < h \le \eta - \eta'$ contradicts the maximality of p^*. Therefore I is an interval with $m_1(I) \ge 2 - s$. Now if $I \ne [-1, 1-s]$, then

$$p(x) := p^*(x + \varepsilon) \in \mathcal{P}_n^r([-1,1], s) \tag{5.11}$$

with sufficiently small $\varepsilon > 0$ contradicts the maximality of p^*, which proves the proposition. □

Now Proposition 5.3 together with a result of Erdős [13, p. 64] yield that $p^* \equiv \widehat{T}_{n,[-1,1-s]}$ and Lemma 4.1 is proved. □

Proof of Lemma 4.2. Let $F_{n,K}$ denote an nth degree monic Fekete polynomial for K and set

$$\widehat{F}_{n,K}(x) := \frac{F_{n,K}(x)}{\|F_{n,K}\|_K} \quad \text{and} \quad \delta := 1 - \sup(K).$$

Then

$$K \subset \{x \in [-1, 1-\delta] : |\widehat{F}_{n,K}(x)| \le 1\}.$$

Therefore,

$$m_1(\{x \in [-1, 1-\delta] : |\widehat{F}_{n,K}(x)| \le 1\}) \ge 2 - s,$$

and Lemma 4.1 implies that

$$|\widehat{F}_{n,K}(z)| \le |\widehat{T}_{n,[-1,1-s]}(z)| \tag{5.12}$$

holds for every $z \in \mathbb{C}$ such that $\operatorname{Re} z \ge \sup(K)$. Since $\operatorname{cap}(K) \ge m_1(K)/4 > 0$ [14, Cor. 4, p. 84], by Lemma 4.3 we have

$$\lim_{n\to\infty} |\widehat{F}_{n,K}(z)|^{1/n} = \exp(g_{D_\infty(K)}(z, \infty)), \qquad z \in D_\infty(K), \tag{5.13}$$

and, as is well known,

$$\lim_{n\to\infty} |\widehat{T}_{n,[-1,1-s]}(z)|^{1/n} = \exp\{g_{D_\infty([-1,1-s])}(z,\infty)\} \tag{5.14}$$

for every $z \in \mathbb{C}$ such that $z \notin [-1, 1-s]$. Now (5.12)–(5.14) yield the lemma except for the point $z_0 = \sup(K) \geq 1-s$. That (4.5) also holds at z_0 can be seen from the limiting argument given in the next proof. □

Proof of Lemma 4.4. For a fixed $0 < \varepsilon < 2-s$ we choose a compact set $K \subset E_{\mu,c}$ such that $\sup(E_{\mu,c}) > \sup(K)$ and

$$m_1(E_{\mu,c}\backslash K) \leq \varepsilon.$$

The last inequality, together with (4.7), yields $m_1(K) \geq 2-s-\varepsilon$. Note that the function

$$g_{D_\infty(K)}(z,\infty) - P_{\mu,c}(z)$$

is superharmonic on $\overline{\mathbb{C}}\backslash K$ and, since $K \subset E_{\mu,c}$,

$$\liminf_{\substack{z\to K \\ z\in D_\infty(K)}} (g_{D_\infty(K)}(z,\infty) - P_{\mu,c}(z)) \geq 0.$$

Therefore the minimum principle for superharmonic functions gives

$$g_{D_\infty(K)}(z,\infty) - P_{\mu,c}(z) \geq 0 \tag{5.15}$$

for all $z \in D_\infty(K)$, and in particular, for all $z \in \mathbb{C}$ with $\operatorname{Re} z \geq \sup(E_{\mu,c}) > \sup(K)$. On the other hand, by the preceding proof,

$$g_{D_\infty(K)}(z,\infty) \leq g_{D_\infty([-1,1-s-\varepsilon])}(z,\infty)$$

for all $z \in \mathbb{C}$ with $\operatorname{Re} z > \sup(K)$. This, together with (5.15) yields

$$P_{\mu,c}(z) \leq g_{D_\infty([-1,1-s-\varepsilon])}(z,\infty) \tag{5.16}$$

for all $z \in \mathbb{C}$ with $\operatorname{Re} z \geq \sup(E_{\mu,c})$. Taking the limit in (5.16) as $\varepsilon \to 0+$, we obtain the desired result. □

Proof of Lemma 4.5. Note that $P_{\mu,c}$ is upper semicontinuous; hence there exists $y \in [-1,1]$ such that

$$P_{\mu,c}(y) = \max_{-1\leq x\leq 1} P_{\mu,c}(x).$$

If $(\mu,c) \in \mathcal{M}(s)$, then either

$$m_1(\{x \in [-1,y] : P_{\mu,c}(x) \leq 0\}) \geq \tfrac{1}{2}(1+y)(2-s) \tag{5.17}$$

or

$$m_1(\{x \in [y,1] : P_{\mu,c}(x) \leq 0\}) \geq \tfrac{1}{2}(1-y)(2-s). \tag{5.18}$$

We may assume that $y > -1$ and that (5.17) holds; otherwise, we study $(\mu(-t),c) \in \mathcal{M}(s)$. Now (5.17) impiles that $(\nu,\hat{c}) \in \mathcal{M}(s)$, where

$$\nu(t) := \mu\left(\frac{y+1}{2}t + \frac{y-1}{2}\right), \qquad \hat{c} := c - \log\left(\frac{2}{y+1}\right),$$

and

$$\max_{-1\leq x\leq 1} P_{\mu,c}(x) = P_{\mu,c}(y) = P_{\nu,\hat{c}}(1),$$

which proves the lemma. □

6. Proofs of Theorems 2.1 and 2.2.

Proof of Theorem 2.1. From Lemmas 4.4 and 4.5 we deduce that

$$\max_{-1\le x\le 1} Q_{\mu,c}(x) \le \exp(g_{D_\infty([-1,1-s])}(1,\infty)), \tag{6.1}$$

whenever $m_1(E_{\mu,c}) \ge 2 - s$. Note that

$$\exp\{g_{D_\infty([-1,1-s])}(1,\infty)\} = \frac{\sqrt{2}+\sqrt{s}}{\sqrt{2}-\sqrt{s}}, \tag{6.2}$$

which, together with (6.1), yields the first part of the theorem.

Now we prove the unicity part of the theorem. Assume that $(\mu, c) \in \mathcal{M}(s)$, $Q_{\mu,c}$ restricted to $[-1, 1]$ is continuous and

$$Q_{\mu,c}(1) = \frac{\sqrt{2}+\sqrt{s}}{\sqrt{2}-\sqrt{s}}. \tag{6.3}$$

Then, by continuity, $\sup(E_{\mu,c}) < 1$. First we show that

$$P_{\mu,c}(z) = g_{D_\infty([-1,1-s])}(z,\infty) \tag{6.4}$$

for all z in the half plane

$$\mathcal{H} := \{z \in \mathbb{C} : \operatorname{Re} z > \sup(E_{\mu,c})\}.$$

Indeed, (6.3) can be written as $h(1) = 0$, where

$$h(z) := g_{D_\infty([-1,1-s])}(z,\infty) - P_{\mu,c}(z).$$

Since h is superharmonic in the domain and $1 \in \mathcal{H}$, Lemma 4.4 and the minimum principle for superharmonic functions imply that $h(z) \equiv 0$ in $\mathcal{H}$. Thus (6.4) holds in $\mathcal{H}$.

Next we show that

$$\operatorname{supp}(\mu) \subset \mathbb{R}.$$

Assume that $\operatorname{supp}(\mu)\backslash\mathbb{R} \ne \emptyset$. If $w \in \operatorname{supp}(\mu)\backslash\mathbb{R}$, then the disk

$$D(w,\varepsilon) := \{z \in \mathbb{C} : |z - w| < \varepsilon\}$$

has positive μ-measure for every $\varepsilon > 0$. Now let

$$A := \operatorname{supp}(\mu) \cap D(w,\varepsilon) \quad \text{with } \varepsilon = \tfrac{1}{3}|\operatorname{Im} w|. \tag{6.5}$$

We define the linear transformation $\varphi : \mathbb{C} \to \mathbb{C}$ by

$$\varphi(z) := 1 + (z-1)\exp(i(\pi - \arg(w-1))). \tag{6.6}$$

Obviously,

$$|1 - \varphi(t)| = |1 - t| \quad \text{for all } t \in \mathbb{C}, \tag{6.7}$$

and there is a $0 < \delta < 1$ depending only on w and $\sup(E_{\mu,c})$ such that

$$|x - \varphi(t)| < \delta|x - t| \quad \text{for all } t \in A \quad \text{and} \quad -1 \le x \le \sup(E_{\mu,c}). \tag{6.8}$$

We denote the restriction of a measure ν on a measurable set B by $\nu|_B$, and define the measure $\sigma(t) := \mu(\varphi^{-1}(t))$. Then (6.7), (6.8), and $\mu \in \mathcal{M}$ imply

$$\begin{aligned}\int_{\mathbb{C}} \log|1-t|d\mu|_A(t) &= \int_A \log|1-\varphi(t)|d\mu(t)\\ &= \int_{\varphi(A)} \log|1-t|d\sigma(t)\end{aligned} \tag{6.9}$$

and

(6.10)

$$\begin{aligned}\int_{\mathbb{C}} \log|x-t|d\sigma|_{\varphi(A)} &= \int_A \log|x-\varphi(t)|d\mu(t)\\ &< \int_A \log|x-t|d\mu(t) + \mu(A)\log\delta \quad \text{for all } -1 \le x \le \sup(E_{\mu,c}).\end{aligned}$$

Now let

$$\hat{\mu}(t) := \mu|_{\mathbb{C}\setminus A}(t) + \sigma|_{\varphi(A)}(t). \tag{6.11}$$

We have $\hat{\mu} \in \mathcal{M}$, since $\mu \in \mathcal{M}$ and

$$\begin{aligned}\int_{\mathbb{C}} d\hat{\mu} &= \int_{\mathbb{C}\setminus A} d\mu + \int_{\varphi(A)} d\sigma\\ &= \int_{\mathbb{C}\setminus A} d\mu + \int_A d\mu = \int_{\mathbb{C}} d\mu = 1.\end{aligned}$$

From (6.9)–(6.11) we obtain

$$\int_{\mathbb{C}} \log|1-t|d\hat{\mu}(t) = \int_{\mathbb{C}} \log|1-t|d\mu(t) \tag{6.12}$$

and, for $-1 \le x \le \sup(E_{\mu,c})$,

$$\int_{\mathbb{C}} \log|x-t|d\hat{\mu}(t) < \int_{\mathbb{C}} \log|x-t|d\mu(t) + \mu(A)\log\delta. \tag{6.13}$$

Now (6.13) and $(\mu, c) \in \mathcal{M}(s)$ imply

$$(\hat{\mu}, c - \mu(A)\log\delta) \in \mathcal{M}(s),$$

while (6.12) and $0 < \delta < 1$ yield

$$P_{\hat{\mu},c-\mu(A)\log\delta}(1) = P_{\mu,c}(1) - \mu(A)\log\delta > P_{\mu,c}(1),$$

which contradicts the extremal property of $P_{\mu,c}$. Therefore $\operatorname{supp}(\mu) \subset \mathbb{R}$.

Let $[\alpha, \beta]$ be the smallest interval containing $\operatorname{supp}(\mu) \cup [-1, 1-s]$. Since the function $g_{D_\infty([-1,1-s])} - P_{\mu,c}$ is harmonic on $\overline{\mathbb{C}}\setminus[\alpha,\beta]$ and vanishes in the half plane $\mathcal{H}$, we have

$$g_{D_\infty([-1,1-s])}(z,\infty) \equiv P_{\mu,c}(z) \tag{6.14}$$

for all $z \in \mathbb{C}\backslash[\alpha,\beta]$. In particular, letting $z \to \infty$ in (6.14),

$$c = -\log\frac{2-s}{4}.$$

Since (6.14) can be written as

$$\int_{\mathbb{C}} \log|z-t|d\mu(t) = \int_{\mathbb{C}} \log|z-t|d\mu^*_{[-1,1-s]}(t)$$

for all $z \in \mathbb{C}\backslash[\alpha,\beta]$, the result of [10, Thm. 1.12′, p. 76] yields $\mu = \mu^*_{[-1,1-s]}$.

Finally, if $(\mu,c) \in \mathcal{M}(s)$ and

$$Q_{\mu,c}(y) = \frac{\sqrt{2}+\sqrt{s}}{\sqrt{2}-\sqrt{s}}, \qquad y \in [-1,1],$$

then, as in the proof of Lemma 4.5, we have either

$$(\hat{\nu},\hat{c}) \in \mathcal{M}(s) \quad \text{or} \quad (\tilde{\nu},\tilde{c}) \in \mathcal{M}(s),$$

where

$$\hat{\nu}(t) := \mu((y+1)t/2 + (y-1)/2), \qquad \hat{c} := c - \log\left(\frac{2}{y+1}\right)$$

and

$$\tilde{\nu}(t) := \mu((1-y)t/2 + (1+y)/2), \qquad \tilde{c} := c - \log\left(\frac{2}{1-y}\right).$$

If $(\hat{\nu},\hat{c}) \in \mathcal{M}(s)$, then

$$Q_{\hat{\nu},\hat{c}}(1) = Q_{\mu,c}(y) = \frac{\sqrt{2}+\sqrt{s}}{\sqrt{2}-\sqrt{s}},$$

and so, by the first part of the proof,

$$\hat{\nu} = \mu^*_{[-1,1-s]} \quad \text{and} \quad \hat{c} = -\log\frac{2-s}{4}.$$

Since $Q_{\hat{\nu},\hat{c}}(1) \geq Q_{\mu,c}(1) = Q_{\hat{\nu},\hat{c}}((3-y)/(1+y))$, it follows that $y = 1$. Hence

$$\mu = \hat{\nu} = \mu^*_{[-1,1-s]} \quad \text{and} \quad c = \hat{c} = -\log\frac{2-s}{4}.$$

If $(\tilde{\nu},\tilde{c}) \in \mathcal{M}(s)$, then in exactly the same way we obtain

$$\mu = \tilde{\nu} = \mu^*_{[-1+s,1]} \quad \text{and} \quad c = \tilde{c} = -\log\frac{2-s}{4},$$

which completes the proof. □

Proof of Theorem 2.2. We assume that $0 < s < 1$, since the case $s = 1$ can be obtained by Theorem 2.1. Let $\mu \in \mathcal{M}$ and $c \in \mathbb{R}$ be such that (2.4) holds. For a fixed $0 < \varepsilon < 1 - s$ we choose a compact set $K \subset E_{\mu,c}$ such that

$$m_1(E_{\mu,c}\backslash K) \leq \varepsilon.$$

Then, as in the proof of Lemma 4.4, we deduce that

$$g_{D_\infty(K)}(z,\infty) - P_{\mu,c}(z) \geq 0 \tag{6.15}$$

for all $z \in D_\infty(K)$. Note that assumption (2.4) and the choice of K imply $m_1(K) \geq 2 - s + \varepsilon$. Applying [4, Thm. 4] to the Fekete polynomials $F_{n,K} \in \mathcal{P}_n^r$, we have

$$\frac{1}{n} \log \frac{|F_{n,K}(x)|}{\|F_{n,K}\|_K} \leq k_1 \min\left\{ \frac{s+\varepsilon}{\sqrt{1-x^2}}, \sqrt{s+\varepsilon} \right\} \tag{6.16}$$

for every $-1 \leq x \leq 1$, where $k_1 > 0$ is an absolute constant. By Lemma 4.3 and $m_1(K) \geq 2 - s - \varepsilon > 0$, the limit of the left-hand side of (6.16), as $n \to \infty$, exists for every $x \in [-1,1]\backslash K$, and equals $g_{D_\infty(K)}(x,\infty)$. Therefore (6.16) and Lemma 4.3 imply

$$g_{D_\infty(K)}(x,\infty) \leq k_1 \min\left\{ \frac{s+\varepsilon}{\sqrt{1-x^2}}, \sqrt{s+\varepsilon} \right\}$$

for every $x \in [-1,1]\backslash K$, and together with (6.15) this yields

$$P_{\mu,c}(x) \leq k_1 \min\left\{ \frac{s+\varepsilon}{\sqrt{1-x^2}}, \sqrt{s+\varepsilon} \right\} \tag{6.17}$$

for every $x \in [-1,1]\backslash K$. Since $P_{\mu,c}(x) \leq 0$ for every $x \in K \subset E_{\mu,c}$, (6.17) holds for every $x \in [-1,1]$. Taking the limit in (6.17) as $\varepsilon \to 0+$, we get the desired result. □

7. Proofs of Theorems 2.4, 2.5, and 2.6. Denote by $\mathcal{T}_n$ the set of all real trigonometric polynomials of degree at most n. Note that $p \in \mathcal{P}_n^c$ implies that $q_r(t) := |p(re^{it})|^2 \in \mathcal{T}_n$ for every $r > 0$. This follows immediately from the identity

$$\begin{aligned} |z - z_j|^2 &= |re^{it} - r_j e^{it_j}|^2 = (re^{it} - r_j e^{it_j})(re^{-it} - r_j e^{-it_j}) \\ &= r^2 + r_j^2 - 2rr_j \cos(t - t_j) \end{aligned} \tag{7.1}$$

$$(z = re^{it}, z_j = r_j e^{it_j}, t, t_j \in \mathbb{R}, r > 0, r_j > 0).$$

In the proof of Theorem 2.5 a Remez-type inequality on the size of trigonometric polynomials will play a central role. To formulate this we introduce the notation

$$\mathcal{T}_n(s) := \{q \in \mathcal{T}_n : m_1(\{t \in [-\pi,\pi) : |q(t)| \leq 1\}) \geq 2\pi - s\} \qquad (0 < s < 2\pi). \tag{7.2}$$

LEMMA 7.1. *There is an absolute constant $k_{13} > 0$ such that*

$$\max_{-\pi \leq t \leq \pi} |q(t)| \leq \exp(k_{13} n s) \qquad (0 < s \leq \pi/2)$$

for every $q \in \mathcal{T}_n(s)$.

Lemma 7.1 is proved in [4, Thm. 3]. Our next lemma is a well-known, simple consequence of the maximum principle for analytic functions.

LEMMA 7.2. *Let $\overline{D} := \{z \in \mathbb{C} : |z| \leq 1\}$. We have*

$$\max_{u \in \overline{D}} |p(u)| \leq (1-r)^{-n} \max_{|u| \leq 1-r} |p(u)|$$

for every $p \in \mathcal{P}_n^c$ and $0 < r < 1$.

Theorem 2.5 will be used in the proof of Theorem 2.4, so we prove Theorem 2.5 first. The proof of Theorem 2.4 will be given at the end of this section.

Proof of Theorem 2.5. Let $p \in \mathcal{P}_n^c(\overline{D}, s)(0 < s \leq \frac{1}{4})$. Observe that if $q_r(t) := |p(re^{it})|^2 \notin \mathcal{T}_n(2\sqrt{s})$ for every $1 - \sqrt{s} \leq r \leq 1$, then

$$m_2(\{z \in \overline{D} : |p(z)|^2 > 1\}) > \int_{1-\sqrt{s}}^{1} 2\sqrt{sr}dr \geq \sqrt{s}2\sqrt{s}(1-\sqrt{s}) \geq s \tag{7.3}$$

($0 < s \leq \frac{1}{4}$ was used in the last inequality), which contradicts the fact that $p \in \mathcal{P}_n^c(\overline{D}, s)$. Thus there exists an $r_0, (1 - \sqrt{s} \leq r_0 \leq 1)$ such that

$$q_{r_0}(t) = |p(r_0 e^{it})|^2 \in \mathcal{T}_n(2\sqrt{s}). \tag{7.4}$$

Then, by Lemma 7.1, we obtain

$$\max_{-\pi \leq t \leq \pi} |p(r_0 e^{it})|^2 = \max_{-\pi \leq t \leq \pi} |q_{r_0}(t)| \leq \exp(2k_{13} n\sqrt{s}). \tag{7.5}$$

Furthermore, Lemma 7.2, together with $1 - \sqrt{s} \leq r_0 \leq 1$ and $0 < s \leq \frac{1}{4}$, yields

$$\begin{aligned} \max_{u \in \overline{D}} |p(u)|^2 &\leq (1-\sqrt{s})^{-2n} \max_{|u| \leq r_0} |p(u)|^2 \\ &\leq \exp(4n\sqrt{s}) \max_{|u| \leq r_0} |p(u)|^2 = \exp(4n\sqrt{s}) \max_{-\pi \leq t \leq \pi} |p(r_0 e^{it})|^2. \end{aligned} \tag{7.6}$$

Now (7.5) and (7.6) give the theorem with $k_4 := k_{13} + 2$. □

Proof of Theorem 2.6. Let $T_n(x) = \cos(n \arccos x)(-1 \leq x \leq 1)$ be the Chebyshev polynomial of degree n. For $0 < s \leq 1$, define the polynomials

$$T_{n,s}(z) := T_n\left(\frac{z}{\cos\sqrt{s}}\right) \tag{7.7}$$

and

$$Q_{3n,s}(z) := z^{2n} T_{n,s}\left(\frac{z + z^{-1}}{2}\right). \tag{7.8}$$

Obviously,

$$\max_{|u| \leq 1} |Q_{3n,s}(u)| = |Q_{3n,s}(1)| = T_n\left(\frac{1}{\cos\sqrt{s}}\right) \geq T_n\left(1 + \frac{s}{2}\right). \tag{7.9}$$

Let

$$\begin{aligned} D_{s,c} := \{z \in \mathbb{C} : |z| \leq 1, \arg z \in [\sqrt{s}, \pi - \sqrt{s}] \cup [\pi + \sqrt{s}, 2\pi - \sqrt{s}]\} \\ \cup \{z \in \mathbb{C} : |z| \leq 1 - c\sqrt{s}\}, \end{aligned} \tag{7.10}$$

where $0 < c \leq 1$ will be chosen later. We examine the maximum of $|Q_{3n,s}|$ on $D_{s,c}$. By the maximum principle, it is sufficient to examine the maximum of $|Q_{3n,s}|$ on the boundary of $D_{s,c}$. For z satisfying $|z| = 1, \arg z \in [\sqrt{s}, \pi - \sqrt{s}] \cup [\pi + \sqrt{s}, 2\pi - \sqrt{s}]$, we have

$$|Q_{3n,s}(z)| \leq \max_{|x| \leq \cos\sqrt{s}} \left| T_n\left(\frac{x}{\cos\sqrt{s}}\right) \right| \leq 1. \tag{7.11}$$

Furthermore, by the maximum principle, we have for $|z| = 1 - c\sqrt{s}$,

(7.12)
$$\begin{aligned} |Q_{3n,s}(z)| &\le (1-c\sqrt{s})^n \max_{|u|\le 1-c\sqrt{s}} \left| u^n T_{n,s}\left(\frac{u+u^{-1}}{2}\right)\right| \\ &\le \exp(-cn\sqrt{s}) \max_{|u|=1} \left| u^n T_{n,s}\left(\frac{u+u^{-1}}{2}\right)\right| = \exp(-cn\sqrt{s}) T_n\left(\frac{1}{\cos\sqrt{s}}\right). \end{aligned}$$

Now let

(7.13) $$z = r(\cos\sqrt{s} + i\sin\sqrt{s}) \quad \text{with } 1 - c\sqrt{s} \le r \le 1.$$

If $c = \frac{1}{8}$ and $0 < s \le 1$ in (7.13), then

(7.14)
$$\begin{aligned} \left|\frac{z+z^{-1}}{2} - \cos\sqrt{s}\right| &= \left|\left(\frac{r+r^{-1}}{2} - 1\right)\cos\sqrt{s} + i\frac{r-r^{-1}}{2}\sin\sqrt{s}\right| \\ &\le \frac{1}{2}\frac{c^2 s}{1-c\sqrt{s}}\cos\sqrt{s} + \frac{c\sqrt{s}}{1-c\sqrt{s}}\sin\sqrt{s} \le \frac{s}{4}\cos\sqrt{s}. \end{aligned}$$

Therefore, using the fact that the zeros of $T_{n,s}$ are in $(-\cos\sqrt{s}, \cos\sqrt{s})$, we easily conclude

(7.15) $$|Q_{3n,s}(z)| \le \left|T_{n,s}\left(\frac{z+z^{-1}}{2}\right)\right| \le T_{n,s}\left(\left(1+\frac{s}{4}\right)\cos\sqrt{s}\right) = T_n\left(1+\frac{s}{4}\right).$$

By the reason of symmetry (7.15) holds when $z = r(\pm\cos\sqrt{s} \pm i\sin\sqrt{s}), 1 - \sqrt{s}/8 \le r \le 1$, and $0 < s \le 1$. We define

(7.16) $$K(n,s) := \max\left\{T_n\left(\frac{1}{\cos\sqrt{s}}\right)\exp\left(-\frac{1}{8}n\sqrt{s}\right), T_n\left(1+\frac{s}{4}\right)\right\}$$

and

(7.17) $$P_{3n,s}(z) := \frac{Q_{3n,s}(z)}{K(n,s)}.$$

By (7.11), (7.12), (7.15)–(7.17) and the maximum principle we can easily deduce that

(7.18) $$|P_{3n,s}(z)| \le 1 \quad \text{for } z \in D_{s,1/8}.$$

Hence

(7.19) $$P_{3n,s} \in \mathcal{P}^r_{3n}(\overline{D}, s) \qquad (0 < s \le 1).$$

Finally, by (7.9), (7.16), and (7.17) we obtain

(7.20)
$$\begin{aligned} P_{3n,s}(1) &= \frac{T_n(1/\cos\sqrt{s})}{K(n,s)} \\ &\ge \min\left\{\exp\left(\frac{1}{8}n\sqrt{s}\right), \frac{T_n(1+s/2)}{T_n(1+s/4)}\right\} = \exp\left(\frac{1}{8}n\sqrt{s}\right), \end{aligned}$$

which completes the proof. □

Proof of Theorem 2.4. Denote the boundary of Ω by Γ. Since Γ is a C^2 curve, there is an $r > 0$ depending only on Ω such that for each $z \in \Gamma$ there is an open disk D_z with radius r such that $D_z \subset \Omega$ and $\overline{D}_z \cap \Gamma = \{z\}$. By the maximum principle for analytic functions, for every $p \in \mathcal{P}_n^c$ there is a $z_0 \in \Gamma$ such that $|p(z_0)| = \max_{u\in\overline{\Omega}} |p(u)|$. It follows from Theorem 2.5, by a linear transformation, that there are constants $k_{14} := k_4\sqrt{1/r^2} = k_4/r > 0$ and $k_{15} := r^2/4 > 0$ depending only on Ω (k_4 is the same as in Theorem 2.5) such that

$$\max_{u\in\overline{\Omega}} |p(u)| = \max_{u\in\overline{D}_{z_0}} |p(u)| \le \exp(k_4 n\sqrt{s/r^2}) = \exp(k_{14} n\sqrt{s}) \qquad (0 < s \le k_{15}) \tag{7.21}$$

for every $p \in \mathcal{P}_n^c$ satisfying

$$m_2(\{z \in \overline{\Omega} : |p(z)| \le 1\}) \ge m_2(\overline{\Omega}) - s. \tag{7.22}$$

Now let $\mu \in \mathcal{M}, c \in \mathbb{R}$,

$$E_{\Omega,\mu,c} := \{z \in \overline{\Omega} : Q_{\mu,c}(z) \le 1\}, \tag{7.23}$$

and assume that

$$m_2(E_{\Omega,\mu,c}) > m_2(\overline{\Omega}) - s. \tag{7.24}$$

For a fixed $0 < \varepsilon < m_2(\overline{\Omega}) - s$ we choose a compact set $K \subset E_{\Omega,\mu,c}$ such that

$$m_2(E_{\Omega,\mu,c}\backslash K) \le \varepsilon. \tag{7.25}$$

This, together with (7.24), gives

$$m_2(K) \ge m_2(\overline{\Omega}) - (s + \varepsilon). \tag{7.26}$$

As in the proof of Lemma 4.4, we obtain that

$$g_{D_\infty(K)}(z, \infty) - P_{\mu,c}(z) \ge 0 \tag{7.27}$$

for all $z \in D_\infty(K)$. Applying (7.21) to the normalized Fekete polynomials

$$\widehat{F}_{n,K} := \frac{F_{n,K}}{\|F_{n,K}\|_K}, \tag{7.28}$$

we obtain

$$\frac{1}{n} \log |\widehat{F}_{n,K}(z)| \le K_{14}\sqrt{s + \varepsilon} \tag{7.29}$$

for every $z \in \overline{\Omega}$ and $0 < s + \varepsilon \le k_{15}$. By Lemma 4.3, the limit of the left-hand side of (7.29), as $n \to \infty$, exists for every $z \in D_\infty(K)$ and equals $g_{D_\infty(K)}(z, \infty)$. Therefore, (7.29) and Lemma 4.3 imply

$$g_{D_\infty(K)}(z, \infty) \le k_{14}\sqrt{s + \varepsilon} \tag{7.30}$$

for every $z \in D_\infty(K) \cap \overline{\Omega}$ and $0 < s + \varepsilon \le k_{15}$, and together with (7.27) this yields

$$P_{\mu,c}(z) \le k_{14}\sqrt{s + \varepsilon} \tag{7.31}$$

for every $z \in D_\infty(K) \cap \overline{\Omega}$ and $0 < s + \varepsilon \le k_{15}$. Since $K \subset E_{\Omega,\mu,c}$ and $P_{\mu,c}(z)$ is subharmonic in $\mathbb{C}$, it follows from the maximum principle that (7.31) holds for all $z \in \overline{\Omega}$. Taking the limit as $\varepsilon \to 0+$, we get

$$Q_{\mu,c}(z) \le \exp(k_{14}\sqrt{s}) \tag{7.32}$$

for every $z \in \overline{\Omega}$ and $0 < s \le k_{15}$, which completes the proof. □

8. Proofs of Theorems 2.7 and 2.8.

Proof of Theorem 2.7. From Theorem 2.1* we can easily deduce that

$$m_1(E_{\mu,c,s,A}) \geq 2s - s = s \tag{8.1}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}, 0 < s < 1$, and $A \subset [-1,1]$ with $m_1(A) \geq 2 - s$, where

$$E_{\mu,c,s,A} = \left\{ x \in A : Q_{\mu,c}(x) > \frac{1-\sqrt{s}}{1+\sqrt{s}} \max_{-1 \leq y \leq 1} Q_{\mu,c}(y) \right\}. \tag{8.2}$$

Hence

$$\begin{aligned} \int_{[-1,1]\setminus A} (Q_{\mu,c}(x))^p dx &\leq s \max_{-1 \leq y \leq 1} (Q_{\mu,c}(y))^p \\ &< \left(\frac{1+\sqrt{s}}{1-\sqrt{s}} \right)^p \int_{E_{\mu,c,s,A}} (Q_{\mu,c}(x))^p dx \\ &< \left(\frac{1+\sqrt{s}}{1-\sqrt{s}} \right)^p \int_A (Q_{\mu,c}(x))^p dx \\ &\leq \exp(k_6 p \sqrt{s}) \int_A (Q_{\mu,c}(x))^p dx \end{aligned} \tag{8.3}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}, 0 < s \leq \frac{1}{2}$ and $A \subset [-1,1]$ with $m_1(A) \geq 2 - s$, where $k_6 = 4$ is a suitable choice. From (8.3) we immediately get (2.12). □

Theorem 2.8 follows from Theorem 2.4 by straightforward modifications of the proof of Theorem 2.7.

9. Proofs of Theorems 2.9 and 2.10.

Proof of Theorem 2.9. Assume that $p \in \mathcal{P}_n^c$ and

$$m_1(\{t \in [-\pi,\pi) : |p(e^{it})| \leq 1\}) \geq 2\pi - s \qquad (0 < s \leq \pi/2). \tag{9.1}$$

Applying Lemma 7.1 to $q(t) := |p(e^{it})|^2 \in \mathcal{T}_n(s)$, we obtain

$$|p(e^{it})| \leq \exp(k_{13} n s) \qquad (t \in \mathbb{R}). \tag{9.2}$$

The above polynomial inequality can be extended to exponentials of logarithmic potentials with compact support by the technique used in the proof of Theorems 2.2 and 2.4; so we omit the details. □

Theorem 2.10 follows immediately from Theorem 2.9 in exactly the same way as Theorem 2.7 was obtained from Theorem 2.1* ; so we omit the details.

10. Proofs of Theorems 3.1, 3.2, and 3.3.

Proof of Theorem 3.1. For the sake of brevity we denote the norm $\|\cdot\|_{L_p(-1,1)}$ by $\|\cdot\|_p$. It is sufficient to prove the theorem when $p = \infty$, and then a simple argument gives the desired result for arbitrary $0 < q < p < \infty$. To see this, assume that

$$\|f\|_\infty \leq M^{1/q} \|f\|_q$$

for an $f \in L_\infty$ and $0 < q < \infty$, with some factor M. Then

$$\|f\|_p^p = \| \, |f|^{p-q+q}\|_1 < \|f\|_\infty^{p-q}\|f\|_q^q$$

$$\leq M^{p/q-1}\|f\|_q^{p-q}\|f\|_q^q,$$

and therefore

$$\|f\|_p \leq M^{1/q-1/p}\|f\|_q$$

for every $0 < q < p < \infty$. Thus, in the sequel let $0 < q < p = \infty$. Applying Corollary 2.3 with

$$s = \min\{1, q^{-2}\}, \tag{10.1}$$

we obtain

$$m_1\left(\left\{x \in [-1,1] : (Q_{\mu,c}(x))^q \geq e^{-1} \max_{-1\leq y\leq 1}(Q_{\mu,c}(y))^q\right\}\right) \geq k_2 s \tag{10.2}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}$, and $q > 0$. Now, integrating only on the subset E of $[-1,1]$, where

$$(Q_{\mu,c}(x))^q \geq e^{-1} \max_{-1\leq y\leq 1}(Q_{\mu,c}(y))^q, \tag{10.3}$$

and using (10.1) and (10.2), we conclude that

$$\|Q_{\mu,c}\|_\infty^q \leq \frac{e}{m_1(E)} \int_E (Q_{\mu,c}(x))^q dx \leq \frac{e}{k_2}(1+q^2)\|Q_{\mu,c}\|_q^q \tag{10.4}$$

for every $\mu \in \mathcal{M}, c \in \mathbb{R}$, and $q > 0$, and the theorem follows by taking the qth root. □

Theorems 3.2 and 3.3 follow from Theorems 2.4 and 2.9, respectively, by straightforward modifications of the proof of Theorem 3.1.

REFERENCES

[1] P. Borwein and T. Erdélyi, *Remez-, Nikolskii- and Markov-type inequalities for generalized non-negative polynomials with restricted zeros*, Constr. Approx., 8 (1992), pp. 343–362.

[2] T. Erdélyi, *Bernstein and Markov type inequalities for generalized non-negative polynomials*, Canad. J. Math., 43 (1991), pp. 1–11.

[3] ———, *Nikolskii-type inequalities for generalized polynomials and zeros of orthogonal polynomials*, J. Approx. Theory, 67 (1991), pp. 80–92.

[4] ———, *Remez-type inequalities on the size of generalized polynomials*, J. London Math. Soc., 45 (1992), pp. 255–264.

[5] ———, *The Remez inequality on the size of polynomials*, in Approximation Theory VI, Vol. 1, C. K. Chui, L. L. Schumaker, and J. D. Ward, eds., Academic Press, Boston, MA, 1989, pp. 243–246.

[6] ———, *Weighted Markov and Bernstein type inequalities for generalized non-negative polynomials*, J. Approx. Theory, 68 (1992), pp. 283–305.

[7] T. Erdélyi, A. Máté , and P. Nevai, *Inequalities for generalized non-negative polynomials*, Constr. Approx., 8 (1992), pp. 241–255.

[8] T. Erdélyi and P. Nevai, *Generalized Jacobi weights, Christoffel functions and zeros of orthogonal polynomials*, J. Approx. Theory, 68 (1992), pp. 111–132.

[9] G. Freud, *Orthogonal Polynomials*, Pergamon Press, Oxford, 1971.

[10] N. S. Landkoff, *Foundations of Modern Potential Theory*, Springer-Verlag, New York, 1972.

[11] C. Pommerenke, *Univalent Functions*, Vardenhoeck und Rupreicht, Gottingen, 1975.

[12] E. J. Remez, *Sur une propriété des polynômes de Tchebycheff*, Comm. Inst. Sci. Kharkow, 13 (1936), pp. 93–95.

[13] T. J. Rivlin, *The Chebyshev Polynomials*, 2nd ed., John Wiley, New York, 1990.

[14] M. Tsuji, *Potential Theory in Modern Function Theory*, Dover, New York, 1959.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 384–391, March 1994

009

ON NECESSARY MULTIPLIER CONDITIONS FOR LAGUERRE EXPANSIONS II*

GEORGE GASPER† AND WALTER TREBELS‡

Abstract. The necessary multiplier conditions for Laguerre expansions derived by Gasper and Trebels [*Canad. J. Math*, 43 (1991), pp. 1228–1242] are supplemented and modified. This allows the authors to place Markett's Cohen-type inequality [*SIAM J. Math. Anal.*, 14 (1983), pp. 819–833] (up to the log case) in the general framework of necessary conditions.

Key words. Laguerre polynomials, necessary multiplier conditions, Cohen-type inequalities, fractional differences, weighted Lebesgue spaces

AMS subject classifications. 33C65, 42A45, 42C10

1. Introduction. The purpose of this sequel to [3] is to obtain a better insight into the structure of Laguerre multipliers on L^p spaces from the point of view of necessary conditions. We recall that in [3] there occurs the annoying phenomenon that, e.g., the optimal necessary conditions in the case $p = 1$ do not give the "right" unboundedness behavior of the Cesàro means. By slightly modifying these conditions we not only remedy this defect but also derive Markett's Cohen-type inequality [6] (up to the log case) as an immediate consequence.

For the convenience of the reader we briefly repeat the notation. We consider the Lebesgue spaces

$$L^p_{w(\gamma)} = \left\{ f : \|f\|_{L^p_{w(\gamma)}} = \left(\int_0^\infty |f(x)e^{-x/2}|^p x^\gamma \, dx \right)^{1/p} < \infty \right\}, \qquad 1 \le p < \infty,$$

denote the classical Laguerre polynomials by $L_n^\alpha(x)$, $\alpha > -1$, $n \in \mathbf{N}_0$ (see Szegö [8, p. 100]), and set

$$R_n^\alpha(x) = L_n^\alpha(x)/L_n^\alpha(0), \qquad L_n^\alpha(0) = A_n^\alpha = \binom{n+\alpha}{n} = \frac{\Gamma(n+\alpha+1)}{\Gamma(n+1)\Gamma(\alpha+1)}.$$

Associate to f its formal Laguerre series

$$f(x) \sim (\Gamma(\alpha+1))^{-1} \sum_{k=0}^\infty \hat{f}_\alpha(k) L_k^\alpha(x),$$

where the Fourier Laguerre coefficients of f are defined by

$$\hat{f}_\alpha(n) = \int_0^\infty f(x) R_n^\alpha(x) x^\alpha e^{-x} \, dx \tag{1}$$

(if the integrals exist). A sequence $m = \{m_k\}$ is called a (bounded) multiplier on $L^p_{w(\gamma)}$, notation $m \in M^p_{w(\gamma)}$, if

$$\left\| \sum_{k=0}^\infty m_k \hat{f}_\alpha(k) L_k^\alpha \right\|_{L^p_{w(\gamma)}} \le C \|f\|_{L^p_{w(\gamma)}}$$

* Received by the editors February 24, 1992; accepted for publication July 14, 1992.
† Department of Mathematics, Northwestern University, Evanston, Illinois 60208. The work of this author was supported in part by National Science Foundation grant DMS-9103177.
‡ Fachbereich Mathematik, Technische Hochschule Darmstadt, D-64289 Darmstadt, Germany.

for all polynomials f; the smallest constant C for which this holds is called the multiplier norm $\|m\|_{M^p_{\alpha,\gamma}}$. The necessary conditions will be given in certain smoothness properties of the multiplier sequence in question. To this end we introduce a fractional difference operator of order δ by

$$\Delta^\delta m_k = \sum_{j=0}^{\infty} A_j^{-\delta-1} m_{k+j}$$

(whenever the sum converges), the first-order difference operator Δ_2 with increment 2 by

$$\Delta_2 m_k = m_k - m_{k+2},$$

and the notation

$$\Delta_2 \Delta^\delta m_k = \Delta^{\delta+1} m_k + \Delta^{\delta+1} m_{k+1}.$$

Generic positive constants that are independent of the functions (and sequences) will be denoted by C. Within the setting of the $L^p_{w(\gamma)}$-spaces, our main results now read (with $1/p + 1/q = 1$) as the following theorem.

THEOREM 1.1. *Let α, $a > -1$, and $\alpha + a > -1$. If $f \in L^p_{w(\gamma)}$, $1 \le p < 2$, $\gamma > -1$, then*

$$\left(\sum_{k=0}^{\infty} |(k+1)^{(\gamma+1)/p - 1/2} \Delta_2 \Delta^a \hat{f}_\alpha(k)|^q \right)^{1/q} \le C \|f\|_{L^p_{w(\gamma)}}, \tag{2}$$

provided

$$\frac{\gamma+1}{p} \le \frac{\alpha+a}{p} + 1 \quad \text{if } \alpha + a \le \frac{1}{2},$$

$$\frac{\gamma+1}{p} \le \frac{\alpha+a}{2} + 1 + \frac{1}{2}\left(\frac{1}{p} - \frac{1}{2}\right) \quad \text{if } \alpha + a > \frac{1}{2}.$$

If we note that

$$|m_k| \, \|L^\alpha_k\|_{L^p_{w(\gamma)}} = \|m_k L^\alpha_k\|_{L^p_{w(\gamma)}} \le \|m\|_{M^p_{\alpha,\gamma}} \|L^\alpha_k\|_{L^p_{w(\gamma)}}, \qquad \gamma > -1,$$

implies $M^p_{w(\gamma)} \subset l^\infty$, we immediately obtain, as in [3] (see there the proof of Lemma 2.3), the following theorem.

THEOREM 1.2. *Let $m = \{m_k\} \in M^p_{w(\gamma)}$, $1 \le p < 2$, and let α, γ, and a be as in Theorem 1.1. Then*

$$\sup_n \left(\sum_{k=n}^{2n} |(k+1)^{(2\gamma+1)/p - (2\alpha+1)/2} \Delta_2 \Delta^a m_k|^q \frac{1}{k+1} \right)^{1/q} \le C \|m\|_{M^p_{\alpha,\gamma}} \tag{3}$$

An extension of Theorem 1.2 to $2 < p < \infty$ easily follows by duality

$$M^p_{w(\gamma)} = M^q_{w(\alpha q - \gamma q/p)}, \quad -1 < \gamma < p(\alpha+1) - 1, \quad 1 < p < \infty.$$

In view of the results in [3], [6] and for an easy comparison we want to emphasize the cases $\gamma = \alpha$ and $\gamma = \alpha p/2$. Therefore, we state the following corollary.

COROLLARY 1.3. (a) *Let* $m \in M^p_{w(\alpha)}$, $1 \le p < 2$, $\alpha > -1$, *and let* $\lambda := (2\alpha+1)(1/p - 1/2)$. *Then, with* $\lambda > 0$,

$$\sup_n \left(\sum_{k=n}^{2n} |(k+1)^\lambda \Delta_2 \Delta^{\lambda-1} m_k|^q \frac{1}{k+1} \right)^{1/q} \le C \|m\|_{M^p_{w(\alpha)}}$$

if $\alpha + \lambda \ge 3/2$; *if* $\alpha + \lambda < 3/2$, *it has additionally to be assumed that* $\lambda \ge 2 - p$. *In the case* $\alpha + \lambda < 3/2$ *and* $\lambda < 2 - p$ *an analogous result holds when the difference operator* $\Delta_2 \Delta^{\lambda-1}$ *is replaced by* Δ_2.

(b) *Let* $m \in M^p_{w(\alpha p/2)}$, $1 \le p < \frac{4}{3}$, *and* $(\alpha - 1)(1/p - 1/2) \ge -\frac{1}{2}$. *Then*

$$\sup_n \left(\sum_{k=n}^{2n} |(k+1)^{1/p-1/2} \Delta_2 m_k|^q \frac{1}{k+1} \right)^{1/q} \le C \|m\|_{M^p_{w(\alpha p/2)}}.$$

Remark 1. For polynomial $f(x) = \sum_{k=0}^n c_k L_k^\alpha(x)$ Theorem 1.1 yields, by taking only the $k = n$ term on the left-hand side of (2),

$$|c_n|(n+1)^{(\gamma+1)/p-1/2} \le C \|f\|_{L^p_{w(\gamma)}}, \qquad 1 \le p < 2$$

(under the restrictions on γ of Theorem 1.1). In particular, if we choose $\gamma = \alpha$, this comprises [6, form. (1.13)] for his basic case $\beta = \alpha$. For $\gamma = \alpha p/2$ it even extends [6, form. (1.14)] to negative α's, as described in part (b) of Corollary 1.3. The case $2 < p < \infty$ can be solved by an application of a Nikolskii inequality; see [6].

Remark 2. Analogously, Cohen-type inequalities follow from Theorem 1.2; in particular, Corollary 1.3 yields the following corollary.

COROLLARY 1.4. *Let* $m = \{m_k\}_{k=0}^n$ *be a finite sequence,* $1 \le p < 2$, *and* $\alpha > -1$.

(a) *If* $m \in M^p_{w(\alpha)}$, *then*

$$(n+1)^{(2\alpha+2)(1/p-1/2)-1/2}|m_n| \le C \|m\|_{M^p_{w(\alpha)}}, \qquad 1 \le p < \frac{4\alpha+4}{2\alpha+3}.$$

(b) *If* $m \in M^p_{w(\alpha p/2)}$ *and* $(\alpha - 1)(1/p - 1/2) \ge -\frac{1}{2}$, *then*

$$(n+1)^{2/p-3/2}|m_n| \le C \|m\|_{M^p_{w(\alpha p/2)}}, \qquad 1 \le p < 4/3.$$

With the exception of the crucial log case, i.e., $p_0 = (4\alpha+4)/(2\alpha+3)$ or $p_0 = \frac{4}{3}$, Corollary 1.4 contains [6, Thm. 1] and extends it to negative α's. In particular we obtain for the Cesàro means of order $\delta \ge 0$, represented by its multiplier sequence $m^\delta_{k,n} = A^\delta_{n-k}/A^\delta_n$, the "right" unboundedness behavior (see [4])

$$\|\{m^\delta_{k,n}\}\|_{M^p_{w(\alpha)}} \ge C(n+1)^{(2\alpha+2)(1/p-1/2)-1/2-\delta}, \qquad 1 \le p < \frac{4\alpha+4}{2\alpha+3+2\delta}.$$

Remark 3. There arises the question of how far are necessary conditions of the type given in [3] comparable with the present ones. Let $\lambda > 1$. Since $\Delta_2 m_k = \Delta m_k + \Delta m_{k+1}$, we obviously have

$$\sup_n \left(\sum_{k=n}^{2n} |(k+1)^{(2\gamma+1)/p-(2\alpha+1)/2} \Delta_2 \Delta^{\lambda-1} m_k|^q \frac{1}{k+1} \right)^{1/q} \tag{4}$$

$$\leq C \sup_n \left(\sum_{k=n}^{2n} |(k+1)^{(2\gamma+1)/p-(2\alpha+1)/2} \Delta^\lambda m_k|^q \frac{1}{k+1} \right)^{1/q}.$$

In general, a converse cannot hold, as can be seen by the following example: choose $\gamma = \alpha$, $\lambda = (2\alpha+1)(1/p - 1/2)$, and $m_k = (-1)^k (k+1)^{-\varepsilon}$, $0 < \varepsilon < 1$. Then

$$\sup_n \left(\sum_{k=n}^{2n} |(k+1)\Delta m_k|^q \frac{1}{k+1} \right)^{1/q} = \infty,$$

and hence by the embedding properties of the *wbv*-spaces (see [2]) the right-hand side of (4) cannot be finite for all $\lambda > 1$. But since $\Delta_2 \Delta^{\lambda-1} m_k = \Delta^{\lambda-1} \Delta_2 m_k \sim (k+1)^{-\varepsilon-\lambda}$, the left-hand side of (4) is finite for all $\lambda > 1$.

Theorem 1.1 will be proved in §2 by interpolating between (L^1, l^∞)- and (L^2, l^2)-estimates. The $a \neq 0$ case is an easy consequence of the case $a = 0$ when one uses the basic formula (see [3, form. 3 and Remark 3 preceding §3])

$$(5)\quad \Delta^a R_k^\alpha(x) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+a+1)} x^a R_k^{\alpha+a}(x), \qquad x > 0,\ a > -1 - \min\{\alpha, \alpha/2 - 1/4\},$$

where in the case $a > -(2\alpha+1)/4$ the series for the fractional difference converges absolutely. In §3 a necessary (L^1, l^1)-estimate is derived and is compared with a corresponding sufficient (l^1, L^1)-estimate.

2. Proof of Theorem 1.1. Let us first handle the (L^2, l^2)-estimate. Since

$$\Delta_2 \Delta^a \hat{f}_\alpha(k) = \Delta^{1+a} \hat{f}_\alpha(k) + \Delta^{1+a} \hat{f}_\alpha(k+1),$$

it follows from the Parseval formula preceding Corollary 2.5 in [3] that

$$(6)\quad \left(\sum_{k=0}^{\infty} |\sqrt{A_k^{\alpha+1+a}} \Delta_2 \Delta^a \hat{f}_\alpha(k)|^2 \right)^{1/2} \leq C \left(\int_0^\infty |f(t) e^{-t/2} t^{(\alpha+1+a)/2}|^2 \, dt \right)^{1/2}.$$

Concerning the (L^1, l^∞)-estimate, we first restrict ourselves to the case $a = 0$. Define $\mu \in \mathbf{R}$ by

$$2\left(\frac{1}{p} - \frac{1}{2}\right)\mu = \frac{\gamma}{p} - \frac{\alpha+1}{2};$$

with the notation $\mathcal{L}_k^\alpha(t) = (A_k^\alpha/\Gamma(\alpha+1))^{1/2} R_k^\alpha(t) e^{-t/2} t^{\alpha/2}$ it follows that

$$\begin{aligned}
|\Delta_2 \hat{f}_\alpha(k)| &= C \left| \int_0^\infty f(t) \left\{ \mathcal{L}_k^\alpha(t)/\sqrt{A_k^\alpha} - \mathcal{L}_{k+2}^\alpha(t)/\sqrt{A_{k+2}^\alpha} \right\} e^{-t/2} t^{\alpha/2} \, dt \right| \\
&\leq C(k+1)^{-1-\alpha/2} \int_0^\infty |f(t)| |t^{-\mu-1/2} \mathcal{L}_k^\alpha(t)| e^{-t/2} t^{(\alpha+1)/2+\mu} \, dt \\
&\quad + C(k+1)^{-\alpha/2} \int_0^\infty |f(t)| |t^{-\mu-1/2} \{\mathcal{L}_k^\alpha(t) - \mathcal{L}_{k+2}^\alpha(t)\}| \\
&\qquad\qquad e^{-t/2} t^{(\alpha+1)/2+\mu} \, dt \\
&= \mathrm{I} + \mathrm{II}.
\end{aligned}$$

We distinguish the two cases $\alpha \le \frac{1}{2}$ and $\alpha > \frac{1}{2}$ as follows:

First consider the case $\alpha \le \frac{1}{2}$. By the asymptotic estimates for $\mathcal{L}_k^\alpha(t) - \mathcal{L}_{k+2}^\alpha(t)$ in Askey and Wainger [1, p. 699] (see also [6, form. (2.12)]) it follows for $\gamma \le \alpha + p - 1$ that

$$\left\| t^{-\mu-1/2} \{ \mathcal{L}_k^\alpha(t) - \mathcal{L}_{k+2}^\alpha(t) \} \right\|_\infty \le C(k+1)^{-1-\mu},$$

so that

$$\text{(7)} \qquad \text{II} \le C(k+1)^{-1-\mu-\alpha/2} \int_0^\infty |f(t)| e^{-t/2} t^{(\alpha+1)/2+\mu}\, dt, \qquad \gamma \le \alpha + p - 1.$$

By [5, Lemma 1, case 4]

$$\left\| t^{-\mu-1/2} \mathcal{L}_k^\alpha(t) \right\|_\infty \le C(k+1)^{-\mu-5/6},$$

so that trivially

$$\text{I} \le C(k+1)^{-1-\mu-\alpha/2} \int_0^\infty |f(t)| e^{-t/2} t^{(\alpha+1)/2+\mu}\, dt, \quad \frac{\gamma+1}{p} \le \frac{\alpha+1}{2} - \frac{1}{3p} + \frac{2}{3}.$$

By [5, Lemma 1, case 5]

$$\left\| t^{-\mu-1/2} \mathcal{L}_k^\alpha(t) \right\|_\infty \le C(k+1)^{\mu+1/2},$$

so that

$$\begin{aligned} \text{I} &\le C(k+1)^{\mu-1/2-\alpha/2} \int_0^\infty |f(t)| e^{-t/2} t^{(\alpha+1)/2+\mu}\, dt \\ &\le C(k+1)^{-1-\mu-\alpha/2} \int_0^\infty |f(t)| e^{-t/2} t^{(\alpha+1)/2+\mu}\, dt, \quad \frac{\gamma+1}{p} > \frac{\alpha+1}{2} - \frac{1}{3p} + \frac{2}{3}, \end{aligned}$$

provided that $\mu - (\alpha+1)/2 \le -1 - \mu - \alpha/2$, which is equivalent to $\mu \le -\frac{1}{4}$ or $\gamma \le 3p/4 - \frac{1}{2} + \alpha p/2$. But this is no further restriction since for $\alpha \le \frac{1}{2}$ there holds $\alpha + p - 1 \le 3p/4 - \frac{1}{2} + \alpha p/2$. Summarizing, for $-1 < \alpha \le \frac{1}{2}$, $\gamma \le \alpha + p - 1$, and $\mu = (\gamma/p - (\alpha+1)/2)/2(1/p - \frac{1}{2})$ we have that

$$\text{(8)} \qquad \sup_k |(k+1)^{1+\mu+\alpha/2} \Delta_2 \hat{f}_\alpha(k)| \le C \int_0^\infty |f(t)| e^{-t/2} t^{(\alpha+1)/2+\mu}\, dt.$$

Now consider the case $\alpha > 1/2$. Then, by [6, form. (2.12)], (7) is obviously true when $(\gamma+1)/p \le \alpha/2 + 1 + (1/p - 1/2)/2$. Again, the application of [5, Lemma 1] requires $\gamma \le \alpha + p - 1$, which for $\alpha > \frac{1}{2}$ is less restrictive than $(\gamma+1)/p \le \alpha/2 + 1 + (1/p - \frac{1}{2})/2$. Now [5, Lemma 1, case 4] leads to

$$\text{I} \le C(k+1)^{-11/6-\mu-\alpha/2} \int_0^\infty |f(t)| e^{-t/2} t^{(\alpha+1)/2+\mu}\, dt, \qquad \frac{\gamma+1}{p} \le \frac{\alpha+1}{2} - \frac{1}{3p} + \frac{2}{3},$$

and [5, Lemma 1, case 5] leads to

$$\text{I} \le C(k+1)^{\mu-1/2-\alpha/2} \int_0^\infty |f(t)| e^{-t/2} t^{(\alpha+1)/2+\mu}\, dt, \qquad (\gamma+1)/p > \frac{\alpha+1}{2} - \frac{1}{3p} + \frac{2}{3}.$$

But $\mu - \frac{1}{2} - \alpha/2 \le -\mu - 1 - \alpha/2$ if $(\gamma+1)/p \le \alpha/2 + 1 + (1/p - \frac{1}{2})/2$, so that, summarizing, (8) also holds under this restriction for $\alpha > \frac{1}{2}$.

Now an application of the Stein and Weiss interpolation theorem (see [7]) with $Tf = \{Tf(k)\}$ and $Tf(k) = \sqrt{A_k^{\alpha+1}}\Delta_2 \hat{f}_\alpha(k)$ gives the assertion of Theorem 1.1 in the case $a = 0$.

If $a \neq 0$, then by (1), the definition of $\Delta_2\Delta^a$, and (5)

$$\Delta_2\Delta^a \hat{f}_\alpha(k) = C\{\Delta \hat{f}_{\alpha+a}(k) + \Delta \hat{f}_{\alpha+a}(k+1)\} = C\Delta_2 \hat{f}_{\alpha+a}(k)$$

since already the condition $\gamma < \alpha + a + 1$ (which implies no new restriction) gives absolute convergence of the infinite sum and integral involved (see the formula following (9) in [3]) and Fubini's theorem can be applied. Hence all the previous estimates remain valid when α is replaced by $\alpha + a$.

3. A variant for integrable functions. Theorem 1.1 gives a necessary condition for a sequence $\{f_k\}$ to generate with respect to L_k^α an $L^1_{w(\gamma)}$-function. But this condition is hardly comparable with the following sufficient one, which is a slight modification of [3, Lemma 2.2].

THEOREM 3.1. *Let $\alpha > -1$ and $\delta > 2\gamma - \alpha + \frac{1}{2} \ge 0$. If $\{f_k\}$ is a bounded sequence with $\lim_{k\to\infty} f_k = 0$ and*

$$\sum_{k=0}^{\infty} (k+1)^{\delta+\alpha-\gamma} |\Delta^{\delta+1} f_k| \le K_{\{f_k\}},$$

then there exists a function $f \in L^1_{w(\gamma)}$ with $\hat{f}_\alpha(k) = f_k$ for all $k \in \mathbf{N}_0$ and

$$\|f\|_{L^1_{w(\gamma)}} \le C\, K_{\{f_k\}}$$

for some constant C independent of the sequence $\{f_k\}$.

The proof follows along the lines of [3, Lemma 2.2] since the norm of the Cesàro kernel

$$\chi_n^{\alpha,\delta}(x) = (A_n^\delta \Gamma(\alpha+1))^{-1} \sum_{k=0}^{n} A_{n-k}^\delta L_k^\alpha(x) = (A_n^\delta \Gamma(\alpha+1))^{-1} L_n^{\alpha+\delta+1}(x)$$

can be estimated with the aid of [5, Lemma 1] by

$$\left\| \chi_k^{\alpha,\delta} \right\|_{L^1_{w(\gamma)}} \le C(k+1)^{\alpha-\gamma}, \qquad \delta > 2\gamma - \alpha + \tfrac{1}{2}.$$

The variant of Theorem 1.1 in the case $p = 1$ is the following theorem.

THEOREM 3.2. *If $\alpha > -1$ and $\gamma > \max\{-\frac{1}{3}, \alpha/2 - \frac{1}{6}\}$, then*

$$\sum_{k=0}^{\infty} (k+1)^{\gamma-2/3} |\Delta^{2\gamma-\alpha+1/3} \hat{f}_\alpha(k)| \le C \|f\|_{L^1_{w(\gamma)}}.$$

A comparison of the sufficient condition and the necessary one nicely shows where the $L^1_{w(\gamma)}$-functions live; in particular, we see that the smoothness gap (the difference of the orders of the difference operators) is just greater than $\frac{7}{6}$. It is clear that Theorem 3.2 can be modified by using the Δ_2 operator. Theorem 3.2 does not follow

from the $p = 1$ case of [3, Lemma 2.1] since that estimate would lead to the divergent sum $\sum_{k=0}^{\infty}(k+1)^{-1}\|f\|_{L^1_{w(\gamma)}}$.

Proof. By formula (5) we have

$$\Delta^{2\gamma-\alpha+1/3}\hat{f}_\alpha(k) = C\int_0^\infty f(t)R_k^{2\gamma+1/3}(t)t^{2\gamma+1/3}e^{-t}\,dt$$
$$= C(k+1)^{-\gamma-1/6}\int_0^\infty f(t)\mathcal{L}_k^{2\gamma+1/3}(t)t^{\gamma+1/6}e^{-t/2}\,dt,$$

and hence

$$\sum_{k=0}^{\infty}(k+1)^{\gamma-2/3}|\Delta^{2\gamma-\alpha+1/3}\hat{f}_\alpha(k)|$$
$$\le C\int_0^\infty |f(t)|\sum_{k=0}^{\infty}(k+1)^{-5/6}|t^{1/6}\mathcal{L}_k^{2\gamma+1/3}(t)|t^\gamma e^{-t/2}\,dt$$

if the right-hand side converges. To show this we discuss for $j \in \mathbf{Z}$

$$\sup_{2^j\le t\le 2^{j+1}}\sum_{k=0}^{\infty}(k+1)^{-5/6}|t^{1/6}\mathcal{L}_k^{2\gamma+1/3}(t)|$$

and prove that this quantity is uniformly bounded in j, whence the assertion.

First consider those $j \ge 0$ for which there exists a nonnegative integer n such that $0 \le k \le 2^n$ implies $3\nu/2 := 3(2k + 2\gamma + 4/3) \le 2^j$ but such that this inequality fails to hold for $k \ge 2^{n+1}$; the latter assumption, in particular, implies that essentially $\nu/2 \ge 2^{j+1}$ for $k \ge 2^{n+4}$. Since $\|t^{1/6}\mathcal{L}_k^{2\gamma+1/3}(t)\|_\infty \le C(k+1)^{-1/6}$ by [5, Lemma 1], we obviously have

$$\sum_{k=0}^{\infty}(k+1)^{-5/6}|t^{1/6}\mathcal{L}_k^{2\gamma+1/3}(t)| \le \left(\sum_{k=0}^{2^n}+\sum_{k=2^{n+4}}^{\infty}\right)+O(1). \tag{9}$$

For $k = 0,\dots,2^n$ we can apply [5, form. (2.5), case 4] to obtain $|t^{1/6}\mathcal{L}_k^{2\gamma+1/3}(t)| \le Ce^{-\mu 2^j}$ for some positive constant μ, and the first sum on the right-hand side of (9) is bounded uniformly in j. In consequence of the choice of n, [5, form. (2.5), case 2] can be used for $k \ge 2^{n+4}$, giving

$$\sum_{k=2^{n+4}}^{\infty}(k+1)^{-5/6}|t^{1/6}\mathcal{L}_k^{2\gamma+1/3}(t)| \le Ct^{-1/12}\sum_{k=2^{n+4}}^{\infty}(k+1)^{-13/12} = O(1)$$

since $2^j \le t \le 2^{j+1}$ and j and n are comparable.

Now consider the remaining j's: We have to split up the sum $\sum_{k=0}^{\infty}\cdots$ into two parts, one where k is such that $2^j\nu \ge 1$ (this contribution has just been seen to be uniformly bounded in j) and the other where k is such that $2^j\nu \le 1$. To deal with the last case again choose n to be the greatest integer such that $2^{n+2} + 4\gamma + 8/3 \le 2^{-j}$; this time, n and $-j$ are comparable and we obtain by [5, form. (2.5), case 1]

$$\sum_{k=0}^{2^n}(k+1)^{-5/6}|t^{1/6}\mathcal{L}_k^{2\gamma+1/3}(t)| \le Ct^{\gamma+1/3}\sum_{k=0}^{2^n}(k+1)^{\gamma-2/3} = O(1)$$

if $2^j \le t \le 2^{j+1}$, $\gamma > -\frac{1}{3}$, which completes the proof. □

REFERENCES

[1] R. ASKEY AND S. WAINGER, *Mean convergence of expansions in Laguerre and Hermite series*, Amer. J. Math., 87 (1965), pp. 695–708.

[2] G. GASPER AND W. TREBELS, *A characterization of localized Bessel potential spaces and applications to Jacobi and Hankel multipliers*, Studia Math., 65 (1979), pp. 243–278.

[3] ———, *Necessary multiplier conditions for Laguerre expansions*, Canad. J. Math., 43 (1991), pp. 1228–1242.

[4] E. GÖRLICH AND C. MARKETT, *A convolution structure for Laguerre series*, Indag. Math. N.S., 44 (1982), pp. 161–171.

[5] C. MARKETT, *Mean Cesàro summability of Laguerre expansions and norm estimates with shifted parameter*, Anal. Math., 8 (1982), pp. 19–37.

[6] ———, *Cohen type inequalities for Jacobi, Laguerre and Hermite expansions*, SIAM J. Math. Anal., 14 (1983), pp. 819–833.

[7] E. M. STEIN AND G. WEISS, *Interpolation of operators with change of measures*, Trans. Amer. Math. Soc., 87 (1958), pp. 159–172.

[8] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., American Mathematical Society Colloq. Publication 23, American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 392–419, March 1994

010

SCATTERING THEORY, ORTHOGONAL POLYNOMIALS, AND q-SERIES*

JEFFREY S. GERONIMO†

Abstract. The techniques of scattering theory and Banach algebras are used to study orthogonal polynomials. The coefficients in the three-term recurrence formula are assumed to converge geometrically to their asymptotic limits. The results are used to investigate certain properties of the Askey–Wilson polynomials.

Key words. orthogonal polynomials, scattering theory, Askey–Wilson polynomials, Banach algebras, q-series

AMS subject classification. 42C05

1. Introduction.

Beginning with the three-term recurrence formula

$$\begin{aligned} &a(n+1)p(\lambda,n+1)+b(n)p(\lambda,n)+a(n)p(\lambda,n-1)=\lambda p(\lambda,n),\\ &p(\lambda,0)=1, \qquad p(\lambda,-1)=0, \qquad n=0,1,2,\dots, \end{aligned} \tag{1.1}$$

with $a(n)>0$ and $b(n-1)$ real for $n=1,2,\dots$, one can construct a sequence of polynomials which are, according to a famous result of Favard [7], orthogonal with respect to some (not necessarily unique) probability measure. In [8], [10], and [11] equation (1.1) was studied in the case when $a(n)\to a(\infty)>0$, $b(n)\to b(\infty)$, and

$$\sum_{n=1}^{\infty} n\nu(2n)\left\{\left|1-\frac{a(n)^2}{a(\infty)^2}\right|+|B(n-1)|\right\}<\infty, \tag{1.2}$$

where

$$B(n)=\frac{b(n)-b(\infty)}{a(\infty)} \tag{1.3}$$

and $\nu(n)$ has the following properties:

$$\begin{aligned} &\nu(n)=\nu(-n),\\ &\nu(n)\le\nu(n+1), \qquad n\ge 0,\\ &\nu(n)\le\nu(m)\nu(n-m), \qquad n,m\ge 0,\\ &\nu(0)=1, \qquad \nu(n)\ge 1, \end{aligned} \tag{1.4}$$

and

$$\limsup_{n\to\infty}(\nu(n))^{1/n}=R,$$

with $R=1$. In this case of course the measure $\rho(\lambda)$ is unique and $\{a(n)\}$ and $\{b(n)\}$ are related to $\rho(\lambda)$ by the standard formulas

$$\begin{aligned} a(n)&=\int \lambda p(\lambda,n)p(\lambda,n-1)d\rho(\lambda), && n=1,2,\dots,\\ b(n)&=\int \lambda p(\lambda,n)^2 d\rho(\lambda), && n=0,1,2,\dots, \end{aligned}$$

*Received by the editors August 31, 1992; accepted for publication (in revised form) March 5, 1993. This work was supported in part by National Science Foundation grant DMS-8620079.

†School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

where the integral is taken over the support of $\rho(\lambda)$.

In Geronimo and Nevai [11] (see also Guseinov [13]) necessary and sufficient conditions were found relating (1.2) to the measure for $R = 1$. Here the case $R \geq 1$ will be investigated. It will be shown that the decay of the coefficients manifests itself in the decay of the Fourier coefficients of the absolutely continuous part (after suitable modifications) of the measure. Another consequence of (1.2) with $R > 1$ is that under certain circumstances one can specify the measure just in terms of the absolutely continuous part and the location of the mass points (see Theorem 4). Note that (1.2) takes into account the exponential decay of the coefficients to their asymptotic values. Thus many of the results obtained here are directly applicable to the q-analogs of some classical orthogonal polynomials, (see [1], [2], and [14]).

We proceed as follows: In §2 the notation is established and some of the analytic consequences of (1.2) are discussed; the techniques of scattering theory and Banach algebras are introduced also. In §3 the addition or removal of mass points is discussed. Also investigated is the addition or removal of polynomial factors (see also [17]) from the measure. Returning in §4 to the absolutely continuous part, the connection between the decay of the coefficients in the recurrence formula and decay of the Fourier coefficients of the absolutely continuous part of the measure is investigated. Finally, in §5 we consider certain examples (the Askey–Wilson polynomials) and obtain some general results on polynomials of this type.

2. Analytic properties. Without loss of generality let

$$a(n) \to \tfrac{1}{2} \quad \text{and} \quad b(n) \to 0. \tag{2.1}$$

Then (1.1) can be written as

$$\Phi(z,n) = C(n)\Phi(z,n-1), \tag{2.2}$$

where

$$\Phi(z,n) = \begin{pmatrix} p(\lambda,n) \\ \psi(z,n) \end{pmatrix},$$

$$C(n) = \frac{1}{2a(n)} \begin{bmatrix} z - 2b(n-1) & 1/z \\ (1-4a(n)^2)z - 2b(n-1) & 1/z \end{bmatrix}, \tag{2.3}$$

$$\lambda = \frac{(z+1/z)}{2},$$

and

$$\Phi(z,0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{2.4}$$

Here we choose the branch

$$z = \lambda - \sqrt{\lambda^2 - 1}. \tag{2.5}$$

To proceed further, it is convenient to introduce the techniques of Banach algebras. Let $\nu(n)$ be given as in (1.4) with $R > 1$. Let A_ν denote the class of functions

$$g(z) = \sum_{n=-\infty}^{\infty} g(n)z^n, \qquad \frac{1}{R} \leq |z| \leq R,$$

with

$$\|g\|_\nu = \sum_{n=-\infty}^{\infty} \nu(n)|g(n)|,$$

and A_ν^+ and A_ν^- denote those functions in A_ν of the form

$$g(z) = \sum_{n=0}^{\infty} g(n)z^n$$

and

$$h(z) = \sum_{n=-\infty}^{0} h(n)z^n,$$

respectively.

If $\|\cdot\|_\nu$ is the norm on A_ν, A_ν^+, and A_ν^-, then A_ν, A_ν^+, and A_ν^- are Banach algebras. We also set $A_\nu = A$, $A_\nu^+ = A^+$, and $A_\nu^- = A^-$ in the case when $v(n) = 1$ for all n.

Note that if (2.1) holds, then (1.2) becomes

$$\sum_{n=1}^{\infty} n\nu(2n)\{|1 - 4a(n)^2| + |2b(n-1)|\} < \infty. \tag{2.6}$$

We begin with the following.

LEMMA 1 (Krein [16]). *Suppose $g(z) \in A_\nu^+$ and $g(z_0) = 0$ for $|z_0| < R$. Then $(g(z)/z - z_0) \in A_\nu^+$.*

Proof. Let $g(z) = \sum_{n=0}^{\infty} g_n z^n$ and define $g_n' = g_n R^n$, $z = Rz'$, and $\nu(n) = R^n\nu'(n)$. Then $g(z) = \sum_{n=0}^{\infty} g_n'(z')^n = \sum_{n=0}^{\infty} g_n'(z_0')^n\left(\frac{z'}{z_0'}\right)^n$. Summation by parts gives

$$g(z) = \sum_{n=1}^{\infty}\sum_{m=n}^{\infty} g_m'(z_0')^m\left(\frac{z'}{z_0'}\right)^{n-1}\left(\frac{z'}{z_0'} - 1\right),$$

where the fact that $g(z_0) = \sum_{n=0}^{\infty} g_n'(z_0')^n = 0$ has been used. The above equation shows that

$$\frac{g(z)}{\left(\frac{z}{z_0} - 1\right)} = \sum_{n=1}^{\infty}\left(\frac{z'}{z_0'}\right)^{n-1}\sum_{m=n}^{\infty} g_m'(z_0')^m.$$

Consequently,

$$\begin{aligned}\left\|\frac{g(z)}{\frac{z}{z_0} - 1}\right\|_\nu &\le \sum_{n=1}^{\infty}\frac{\nu'(n-1)}{|z_0'^{n-1}|}\sum_{m=n}^{\infty}|g_m'||z_0'|^m,\\ &\le \sum_{m=1}^{\infty}\nu'(m)|g_m'||z_0'|^m\sum_{n=0}^{m-1}|1/z_0|^n,\\ &\le \frac{|z_0'|}{1 - |z_0'|}\sum_{m=1}^{\infty}\nu(m)|g_m|,\end{aligned}$$

since $|z_0'| < 1$. □

THEOREM 1. *If (2.6) holds then there exists a function $f_+(z)$ such that*

(i) $\lim_{n\to\infty}\|z^n\psi(z,n) - 2zf_+(z)\|_\nu = 0$;

(ii) *Furthermore, the zeros of $f_+(z)$ for $|z| < 1$ are real, simple, finite in number, and the points where the orthogonal polynomials are square summable;*

(iii) *If $f_+(z)$ has zeros on the unit circle they must be at $z = \pm 1$, and they must be simple.*

(iv) $(zf_+(z)/d(z)) \in A_\nu^+$, *where*

$$d(z) = \begin{cases} 1 - z & \text{if } f(1) = 0, \\ 1 + z & \text{if } f(-1) = 0, \\ 1 - z^2 & \text{if } f(1) = f(-1) = 0; \end{cases}$$

(v) $(1 - z^2)(dzf_+(z)/dz) \in A_\nu^+$; *finally, if*

$$\sum_{n=1}^{\infty} n^2\{|1 - 4a(n)^2| + 2|b(n-1)|\} < \infty, \tag{2.7}$$

then

(vi) $(dzf(z)/dz) \in A^+$.

Proof. Parts (ii) and (iii) are proved in Geronimo and Case [10] (see also Chihara and Nevai [6], and Geronimo [9]). For $R > 1$ (iv) follows from Lemma 1 while for $R = 1$ see [10]. Part (v) follows as in Geronimo [8, Thm. 2.3]. To prove (i) first define $\psi^*(z,n) = z^n\psi(z,n)$ and $\alpha(n) = \prod_{i=1}^n (1/2a(i))$, $\alpha(0) = 1$. Multiplying the lower component of (2.2) by z^n, then iterating down yields

$$\frac{\psi^*(z,n)}{\alpha(n)} = 1 + \sum_{i=0}^{n-1} \{(1 - 4a(i+1)^2)z^2 - 2b(i)z\} \frac{z^i p(\lambda,i)}{\alpha(i)}. \tag{2.8}$$

After suitable manipulation of (2.2) (see [10, eq. (B.4)] or [17, Thm. 3]) the following formula can be obtained for $p(\lambda,n)$:

$$\begin{aligned} \frac{z^n p(\lambda,n)}{\alpha(n)} = \frac{1 - z^{2n+2}}{1 - z^2} + \sum_{i=0}^{n-1} \Bigg\{ & (1 - 4a(i+1)^2)z^2 \left(\frac{1 - z^{2n-2i-2}}{1 - z^2}\right) \\ & - 2b(i)z \left(\frac{1 - z^{2n-2i}}{1 - z^2}\right) \Bigg\} z^i \frac{p(\lambda,i)}{\alpha(i)}. \end{aligned} \tag{2.9}$$

To get a bound on $z^n(p(\lambda,n)/\alpha(n))$ note that the above equation yields

$$\begin{aligned} \left\| \frac{z^n p(\lambda,n)}{\alpha(n)} \right\|_\nu \le{}& (n+1)\nu(2n+2) \\ & + \sum_{i=0}^{n-1} \{|1 - 4a(i+1)^2| + 2|b(i)|\} (n-i)\nu(2n-2i) \left\| \frac{z^1 p(\lambda,i)}{\alpha(i)} \right\|. \end{aligned}$$

Since $(n-i)\nu(2n-2i) < (n+1)\nu(2n+2)$ for $i = 0, 1, \ldots, n-1$ we find

$$\|p^*(\lambda,n)\|_\nu \le 1 + \sum_{i=0}^{n-1} \{|1 - 4a(i+1)^2| + 2|b(i)|\}(i+1)\nu(2i+2)\|p^*(\lambda,i)\|_\nu,$$

where $p^*(\lambda,n) = z^n p(\lambda,n)/(n+1)\nu(2n+2)\alpha(n)$. The discrete Gronwall inequality (see Van Assche [21]) now yields

$$\left\|\frac{z^n p(\lambda,n)}{\alpha(n)}\right\|_\nu \le (n+1)\nu(2n+2) \times \exp\sum_{i=1}^{n} i\nu(2i)\left\{|1-4a(i)^2| + |2b(i-1)|\right\}. \tag{2.10}$$

Consequently,

$$\left\|\frac{\psi^*(z,n)}{\alpha(n)}\right\|_\nu \le 1 + \sum_{i=1}^{n} i\nu(2i)\left\{|1-4a(i)^2| + |2b(i-1)|\right\} \times \exp\sum_{j=1}^{n-1} j\nu(2j)\left\{|1-4a(j)^2| + |2b(j-1)|\right\}. \tag{2.11}$$

Furthermore, from (2.8) we find

$$\left\|\frac{\psi^*(z,n)}{\alpha(n)} - \frac{\psi^*(z,m)}{\alpha(m)}\right\|_\nu \le \sum_{i=m}^{n-1} i\nu(2i)\left\{|1-4a(i)^2| + |2b(i-1)|\right\} \times \exp\sum_{j=1}^{i-1} j\nu(2j)\left\{|1-4a(j)^2 + |2b(j-1)|\right\},$$

proving (i). Since $\psi(z,n) \in A_\nu^+$ for all n, $zf(z) \in A_\nu^+$. To prove (vi), differentiate (2.8) and (2.9), then use (2.10) with $v(n) = 1$, for all n, the equation $\|(d/dz)((1-z^{2(n+1)})/(1-z^2))\| = n(n+1)$, and Gronwall's inequality to obtain

$$\left\|\frac{z^n p(\lambda,n)'}{\alpha(n)}\right\| \le (n+1)^2 C \exp C\sum_{i=1}^{n} i^2\left\{|1-4a(i)^2| + 2|b(i-1)|\right\},$$

where C is independent of z and n. This implies

$$\left\|\frac{\psi^*(z,n)'}{\alpha(n)} - \frac{\psi^*(z,m)'}{\alpha(m)}\right\| \le C_1\sum_{i=m}^{n-1} i^2\{|1-4a(i)^2| + 2|b(i-1)|\}, \qquad |z| \le 1;$$

hence $(zf_+(z))' \in A_v^+$ which yields the result. □

This leads to the following.

COROLLARY 1. *If* (2.6) *holds, then* $zf_+(z)$ *is analytic for* $|z| < R \ge 1$ (*see* (1.4)) *and continuous for* $|z| \le R$. *It also follows from* (2.8) *and* (2.9) *that for* $|z| < 1$,

$$z^n p(\lambda,n) - \frac{\psi^*(z,n)}{1-z^2} = o(1). \tag{2.12}$$

So far a consequence of the results above is the extension of the region of analyticity of $zf_+(z)$ to the open disk of radius R. As a function of λ this means f_+ can be continued onto the second Riemann sheet, $(z = \lambda + \sqrt{\lambda^2-1})$. In analogy with scattering theory we shall call the principal sheet $(z = \lambda - \sqrt{\lambda^2-1})$ the physical sheet.

It is possible to include polynomials of the second kind in this scheme by defining

$$\Phi^1(z,n) = \begin{pmatrix} p^1(\lambda,n) \\ \psi^1(z,n) \end{pmatrix}, \qquad n \geq 1, \tag{2.13}$$

satisfying (2.2) with boundary conditions

$$p^1(\lambda,1) = \psi^1(z,1) = 1/a(1).$$

From (2.2) and (2.13) the following equations for $\psi^1(z,n)$ and $p^1(\lambda,n)$ can be derived:

$$\frac{\psi^{*1}(z,n)}{\alpha(n)} = 2z + \sum_{i=1}^{n-1} \{(1-4a(i+1)^2)z^2 - 2b(i)z\} \frac{z^i p^1(\lambda,i)}{\alpha(i)},$$

and

$$\frac{z^n p^1(\lambda,n)}{\alpha(n)} = 2z\left(\frac{1-z^{2n}}{1-z^2}\right) + \sum_{i=1}^{n-1}\left\{(1-4a(i+1)^2)z^2\left(\frac{1-z^{2n-2i-2}}{1-z^2}\right) - 2b(i)z\left(\frac{1-z^{2n-2i}}{1-z^2}\right)\right\}\frac{z^i p^1(\lambda,i)}{\alpha(i)}.$$

Let

$$zf_+^1(z) = \tfrac{1}{2}\lim_{n\to\infty} z^{n-1}\psi^1(z,n). \tag{2.14}$$

Then the above methods show that

$$zf_+^1(z) \in A_\nu. \tag{2.15}$$

Two other useful solutions (see [10]) of (2.2) are

$$\Phi_+(z,n) = \begin{pmatrix} p_+(z,n) \\ \psi_+(z,n) \end{pmatrix} \tag{2.16}$$

and

$$\Phi_-(z,n) = \begin{pmatrix} p_-(z,n) \\ \psi_-(z,n) \end{pmatrix}, \tag{2.17}$$

satisfying the following boundary conditions

$$\lim_{n\to\infty} |z^{-n}p_+(z,n) - 1| = 0, \qquad |z| \leq 1, \tag{2.18a}$$

$$\lim_{n\to\infty} |z^{-n}\psi_+(z,n)| = 0, \qquad |z| \leq 1, \tag{2.18b}$$

and

$$\lim_{n\to\infty} |z^{n}p_-(z,n) - 1| = 0, \qquad |z| \geq 1, \tag{2.19a}$$

$$\lim_{n\to\infty} |z^n\psi_-(z,n) - (1-z^2)| = 0, \qquad |z| \geq 1. \tag{2.19b}$$

THEOREM 2. *If* (2.6) *holds, then* $z^{-n}p_+(z,n)$ *and* $z^{-n}\psi_+(z,n)$ *are elements of* A_ν^+, *which implies that* $\Phi_+(z,n)$ *is analytic for* $|z| < R$. *Likewise* $z^n p_-(z,n)$ *and* $z^n\psi_-(z,n)$ *are elements of* A_ν^-. *Furthermore,* $\Phi_+(z,n)$ *and* $\Phi_-(z,n)$ *are linearly independent for* $1/R \leq |z| \leq R$ *except at* $z = \pm 1$ *and* $p_+(1/z,n) = p_-(z,n)$. *If* (2.7) *holds, then* $p'_+(z,n)$ *is continuous for* $|z| \leq 1$.

Proof. Inverting $C(n)$ we find from (2.2) that

$$p_+(z,n) = \frac{1}{2a(n+1)z}(p_+(z,n+1) - \psi_+(z,n+1)). \tag{2.20}$$

Iterating this equation plus the lower component of (2.2) upwards and using the boundary conditions (2.18) yield a discrete integral equation for $p_+(z,n)$:

$$\begin{aligned}\frac{p_+(z,n)}{\gamma(n+1)} = z^n + \sum_{i=n+1}^{\infty}\sum_{m=i}^{\infty}\left\{\prod_{j=i+1}^{m}(2a(j))^2\right\} \\ \times\{(1-4a(m+1)^2)z - 2b(m)\}\, z^{m-2i+n+1}\frac{p_+(z,m)}{\gamma(m+1)},\end{aligned} \tag{2.21}$$

with $\gamma(n) = \prod_{i=n}^{\infty}(1/2a(i))$. The following bound on $p_+(z,n)$ can be obtained using the discrete Gronwall inequality,

$$\left\|\frac{z^{-n}p_+(z,n)}{\gamma(n+1)}\right\|_\nu \leq \exp\left[C\sum_{m=n+1}^{\infty} m\nu(2m)\left\{|1-4a(m+1)^2| + |2b(m)|\right\}\right], \tag{2.22}$$

with $C = \max_{i,j}\left\{\prod_{k=i+1}^{j}(2a(k))^2\right\}$. Furthermore, using the above equation in (2.21) yields

$$\left\|\frac{z^{-n}p_+(z,n)}{\gamma(n+1)} - 1\right\|_\nu \leq D\sum_{m=n+1}^{\infty} m\nu(2m)\left\{|1-4a(m+1)^2| + |2b(m)|\right\},$$

where D is a constant independent of z and n. Since $p_+(z,n)$ and $p_-(z,n)$ satisfy (1.1) with the boundary conditions (2.18a) and (2.19a), respectively, we see that $p_-(z,n) = p_+(1/z,n)$ for $1/R \leq |z| \leq |R|$, which implies that $z^n p_-(z,n) \in A_\nu^-$. Since the Wronskian of any two solutions Φ_1 and Φ_2, $W[\Phi_1(n),\Phi_2(n)] = \det|\Phi_1(n),\Phi_2(n)|$ is independent of n [10], we find that $W[\Phi_+(n),\Phi_-(n)] = 1 - z^2$, which implies that $\Phi_+(z,n)$ and $\Phi_-(z,n)$ are linearly independent except at $z = \pm 1$.

The properties of ψ_+ follow from (2.20) while those of ψ_- follow an analogous equation. If (2.7) holds, multiply (2.21) by z^{-n}, differentiate, then use (2.22) and Gronwall's inequality to obtain

$$\|(z^{-n}p_+(z,n))'\| \leq \hat{C}\exp C\sum_{m=n+1}^{\infty} m^2|1-4a(m+1)^2| + 2|b(m)|, \tag{2.23}$$

which says that $(z^{-n}p_+(z,n))' \in A$. □

Since Φ_+ and Φ_- are linearly independent except at $z = \pm 1$ we find, using the fact that the Wronskian is independent of n, and the boundary conditions (2.18) and (2.19), the useful equation

$$\Phi(z,n) = \frac{2}{z - 1/z}(f_+(1/z)\Phi_+(z,n) - f_+(z)\Phi_-(z,n)), \quad \frac{1}{R} \le |z| \le R. \tag{2.24}$$

Here we have used (2.18) and (i) of Theorem 1 to make the identification

$$f_+(z) = \frac{1}{2z} W[\Phi_+, \Phi]. \tag{2.25}$$

Now using (2.2) to eliminate ψ_+ and ψ in the above equation yields

$$\begin{aligned} f_+(z) &= a(n+1)[p(\lambda, n+1)p_+(z,n) - p_+(z,n+1)p(\lambda,n)] \\ &= a(0)p_+(z,-1), \qquad |z| \le R. \end{aligned} \tag{2.26}$$

Since Φ and Φ^1 are linearly independent except at $z = 0$, i.e., $W[\Phi, \Phi^1] = 2z$, we also find that

$$\Phi_+(z,n) = f_+^1(z)\Phi(z,n) - f_+(z)\Phi^1(z,n), \qquad |z| \le R, \quad n \ge 1. \tag{2.27}$$

Let

$$\frac{p_+(z,n)}{\gamma(n+1)} = z^n \left(1 + \sum_{i=1}^{\infty} \alpha(n,i) z^i\right). \tag{2.28}$$

Then (2.24) allows us to develop useful asymptotic formulas for $p(\lambda, n)$.

THEOREM 3. *If* (2.6) *holds, then*

$$\begin{aligned} \sin\theta p(\cos\theta, n) = 2|f_+(e^{i\theta})|\gamma(n+1)\{&\sin((n+1)\theta - \arg e^{i\theta} f_+(e^{i\theta})) \\ &+ \sum_{i=1}^{\ell} \alpha(n,i)\sin((n+i+1)\theta - \arg e^{i\theta} f_+(e^{i\theta}))\} \\ + O\left(\sum_{m=[\ell/2]}^{\infty} (2m - \ell + 1)\{|1 - 4a(m+n+1)^2| + 2|b(m+n)|\}\right), \\ &0 \le \theta \le \pi. \end{aligned} \tag{2.29}$$

This implies that the zeros of $p(\cos\theta, n)$ *in* $[-1,1]$ *are located at*

$$\theta = \frac{k\pi}{n+1} + \frac{\arg e^{i\theta} f_+(e^{i\theta})}{n+1} + o(1/n), \qquad k = 1, 2, \ldots, n.$$

If (2.7) *holds, then*

$$\begin{aligned} p(\cos\theta, n) = 2|f_+(e^{i\theta})|\gamma(n+1)\Big\{&\sin((n+1)\theta - \arg e^{i\theta} f_+(e^{i\theta})) \\ &+ \sum_{i=1}^{\ell} \alpha(n,i)\sin((n+i+1)\theta - \arg e^{i\theta} f_+(e^{i\theta}))\Big\}/\sin\theta \\ + O\left(\sum_{k=[\ell/2]}^{\infty} k^2\{|1 - 4a(k+n+1)^2| + 2|b(k+n)|\}\right), &\quad 0 \le \theta \le \pi. \end{aligned} \tag{2.30}$$

Furthermore, from (2.29) *the error term for the zeros of* $p(\cos\theta, n)$ *can be improved to* $O(1/n^2)$.

Proof. Everything in (2.29) but the error term follows from Theorems 1 and 2, (2.28), the upper component of (2.24) and the fact that $f_+(e^{-i\theta}) = \overline{f_+(e^{i\theta})}$. Substituting (2.28) into the left hand side of (2.21) yields

$$\sum_{k=1}^{\infty} \alpha(n,k) z^k = \sum_{i=0}^{\infty} z^{2i+1} \sum_{m=i}^{\infty} C(m+n+1, m-i+n+2) K(m+n+1) \left[1 + \sum_{j=1}^{\infty} \alpha(m+n+1, j) z^j \right],$$

where $C(m,i) = \prod_{j=i+1}^{m} (2a(j))^2$ and $K(m) = (1 - 4a(m+1)^2)z - 2b(m)$. Equating coefficients of z^k we find

$$|\alpha(n,k)| \le C \sum_{i=\left[\frac{k-1}{2}\right]}^{\infty} |K(n+i+1)| + C \sum_{i=n+1}^{\infty} \sum_{j=0}^{\left[\frac{k-2}{2}\right]} |K(i+j)||\alpha(i+j, k-1-2j)|,$$

with $C = \max_{m,i} C(m,i)$. Since $C \sum_{i=n}^{\infty} i|K(i)| < 1$ for all $n \ge n_0$ large enough we find by iterating the above equation that

$$|\alpha(n,k)| \le \frac{C \sum_{i=\left[\frac{k-1}{2}\right]}^{\infty} |K(i+n+1)|}{1 - C \sum_{i=1}^{\infty} i|K(i+n)|}$$

for $n \ge n_0$ large enough. It now follows by induction on n that

$$|\alpha(n,k)| \le O\left(\sum_{m=[(k-1)/2]}^{\infty} |1 - 4a(m+n+2)|^2 + 2|b(m+n+1)| \right). \tag{2.31}$$

Summing the above equation from $k = \ell + 1$ to infinity and using the fact that $p_-(e^{i\theta}, n) = \overline{p_+(e^{i\theta}, n)}$ yields the error term in (2.29). The representation for the zeros of $p(\cos\theta, n)$ now follows from (2.29).

To show (2.30) note that since $p_-(e^{i\theta}, n) = p_+(e^{-i\theta}, n)$, $\frac{p_-(e^{i\theta},n) - p_+(e^{i\theta},n)}{\sin\theta}$ exists for $\theta = 0$ and π by Theorem 1 (vi) and Theorem 2. The same is true for $\frac{f_+(e^{-i\theta}) - f_+(e^{i\theta})}{\sin\theta}$. Thus the error term in (2.30) follows from Theorem 1 (vi), the fact that

$$\left\| \frac{z^{-n} p_+(z,n) - z^n p_-(z,n)}{z - 1/z} \right\| \le \sum_{i=1}^{\infty} |\alpha(n,i)| \left\| \frac{z^i - z^{-i}}{z - 1/z} \right\| \le \sum_{i=1}^{\infty} i|\alpha(n,i)|,$$

and (2.31). □

Estimates for the zeros of $p(\lambda, n)$ for λ near one may be obtained by letting $\theta \to \pi - \theta$.

Equation (2.27) can be used to obtain a useful integral representation for $p_+(z,n)$.

LEMMA 2. *Suppose $p_+(z,n)$ exists for $n = 0,1,2\ldots$, then for $|z| < 1$,*

$$p_+(z,n) = f_+(z)\int \frac{p(\lambda',n)}{\lambda-\lambda'}d\rho(\lambda'), \qquad \lambda = \frac{z+1/z}{2}, \quad n = 0,1,2,\ldots. \tag{2.32}$$

Proof. (This is a modified version of an unpublished proof of Nevai.) Since $p^1(\lambda,n)$ is a polynomial of the second kind,

$$p^1(\lambda,n) = \int \frac{(p(\lambda,n)-p(\lambda',n))}{\lambda-\lambda'}d\rho(\lambda'), \qquad n \geq 1,$$

and from the upper component of (2.27) we find for λ not in the convex hull of the support of $\rho(\lambda)$,

$$\frac{p_+(z,n)}{p(\lambda,n)} = f_+^1(z) - f_+(z)\int \frac{1}{\lambda-\lambda'}d\rho(\lambda') + \frac{f_+(z)}{p(\lambda,n)}\int \frac{p(\lambda',n)}{\lambda-\lambda'}d\rho(\lambda'). \tag{2.33}$$

Since $|z| < 1$ and z is not in the convex hull of the support of $\rho(\lambda)$, $\lim_{n\to\infty}|p(\lambda,n)| = \infty$ and $\lim_{n\to\infty} p_+(z,n) = 0$. These equations, plus the fact that the integral in the last term on the right-hand side of the above equation is uniformly bounded on compact subsets of the unit disk that do not include the support of $\rho(\lambda)$, show that

$$f_+^1(z) = f_+(z)\int \frac{d\rho(\lambda')}{\lambda-\lambda'}, \qquad \lambda = \frac{(z+1/z)}{2}, \qquad |z| < 1, z \notin \text{convex hull supp } \rho.$$

Extending this equation by analytic continuation to $|z| < 1$, using Theorem 4 below, then substituting this result into (2.27) gives the lemma. □

3. The distribution function.

THEOREM 4. *If* (2.6) *holds, then*

$$\int_{-\infty}^{\infty} p(\lambda,n)p(\lambda,m)d\rho(\lambda) = \delta_{n,m}, \tag{3.1}$$

where

$$d\rho(\lambda) = \begin{cases} \sigma(\theta)d\lambda, & \lambda = \cos\theta, \quad 0 \leq \theta \leq \pi, \\ \sum_{i=1}^{N} \rho_i \delta(\lambda-\lambda_i)d\lambda, & \lambda \text{ not as above}, \quad N < \infty, \end{cases}$$

with

$$\sigma(\theta)d\lambda = \frac{\sin\theta}{2\pi|f_+(z)|^2}d\lambda, \qquad \lambda = \cos\theta, \quad z = e^{i\theta}, \tag{3.2}$$

and

$$\rho_i = \frac{p_+(z_i,0)}{f_+'(\lambda_i)}, \qquad \lambda_i = \frac{(z_i+1/z_i)}{2}, \qquad |z_i| < \frac{1}{R}, \tag{3.3}$$

$$\rho_i = \frac{(z_i-1/z_i)}{2}\frac{1}{f_+(1/z_i)f_+'(\lambda_i)}, \qquad \frac{1}{R} \leq |z_i| < 1. \tag{3.4}$$

Here $\{\lambda_i\}$ denote the roots of $f_+(z)$ for $|z| < 1$, and the above differentiation is with respect to λ.

Proof. All but (3.4) has been proved in [10]. Furthermore, it was shown there that (3.3) holds for $|z_i| < 1$. To prove (3.4), note that if (2.6) holds, then (2.24) is valid for all z such that $1/R \le |z| \le R$. Therefore, let $f_+(z_i) = 0$, $1/R \le |z_i| < 1$. Then from (2.24),

$$p(\lambda_i, n) = \frac{2}{z_i - 1/z_i} f_+\left(\frac{1}{z_i}\right) p_+(z_i, n), \qquad n = 0, 1, \ldots,$$

and the result follows by eliminating $p_+(z_i, 0)$ in (3.3). □

If

$$f_+(z) \neq 0, \qquad |z| = R, \tag{3.5}$$

then inside the disk of radius R, $f_+(z)$ has only a finite number of zeros, each having a finite multiplicity. Let $z_0, z_1, z_2, \ldots, z_n$ be the zeros of $f_+(z)$ including multiplicities; then by Lemma 1,

$$\frac{z f_+(z)}{\kappa(z)} \in A_\nu^+, \tag{3.6}$$

where

$$\kappa(z) = \prod_{i=0}^{n} (z - z_i).$$

This leads to the following.

COROLLARY 2. *If* (2.6) *and* (3.5) *hold, then*

$$\ln \frac{\sigma(z)\kappa(z)\kappa(1/z)}{z - 1/z} \in A_\nu.$$

Proof. From (3.2),

$$\sigma(z) = \frac{1}{4\pi i} \frac{z - 1/z}{f_+(z) f_+(1/z)}, \qquad z = e^{i\theta}. \tag{3.7}$$

It follows from (3.6) that $f_+(1/z)/z\kappa(1/z) \in A_\nu^-$. Furthermore, $f_+(1/z)/z\kappa(1/z)$ and $z f_+(z)/\kappa(z)$ are nonzero for $1/R \le |z| \le R$. Thus the result follows from the Wiener–Levy theorem. □

The above result shows that $\sigma(z)$ is a meromorphic function for $1/R < |z| < R$. If $\sigma(z)$ has a pole at $|z_0| < 1$, (3.7) forces $\sigma(z)$ to have a pole at $1/z_0$ and these are the only places $\sigma(z)$ may have poles. Consequently, we need only study the poles of $\sigma(z)$ for $|z| \le 1$, or in terms λ, we can stay on the physical sheet. As seen from (3.7) the poles of $\sigma(z)/(z - 1/z)$ come from the zeros of $f_+(z)$ and $f_-(z)$, and (2.4) prevents $f_+(z)$ and $f_-(z)$ from vanishing at the same value of z for $1/R \le |z| \le R$, $z \neq \pm 1$, since the zeros of $p(\lambda, n)$ alternate with those of $p(\lambda, n+1)$.

The following gives a useful representation of $f_+(z)$ in terms of $\sigma(\theta)$.

COROLLARY 3 (Geronimo and Case [10]). *If* (2.6) *holds with* $\nu(n) = 1$ *for all* n, *then*

$$f_+(z) = \frac{1}{z} \prod_{i=1}^{N} \frac{|z_i|}{z_i} \frac{(z_i - z)}{1 - z_i z} \exp \frac{-1}{4\pi} \int_{-\pi}^{\pi} \left(\frac{e^{i\theta'} + z}{e^{i\theta'} - z}\right) \ln\left(\frac{\sigma(\theta') 2\pi}{\sin\theta'}\right) d\theta', \qquad |z| < 1.$$

What we would like to consider now is the effect of adding or removing the poles (zeros) of $\sigma(z)$ ($f_+(z)$). To this end Geronimo and Nevai [11] have proven the following.

THEOREM 5. *Let $d\rho(\lambda)$ be given as*

$$d\rho(\lambda) = \begin{cases} \sigma(\theta)d\lambda, & \lambda = \cos\theta, \\ \sum_{m=1}^{N} \rho_m \delta(\lambda - \lambda_m)d\lambda, & \lambda \text{ not as above}, \end{cases} \quad N \geq 1,\ |\lambda_i| \geq |\lambda_{i-1}|,$$

and let

$$d\rho^*(\lambda) = \begin{cases} \sigma(\theta)d\lambda, \\ \sum_{m=1}^{N-1} \rho_m \delta(\lambda - \lambda_m)d\lambda. \end{cases}$$

Furthermore let $\{a(n)\}$, $\{b(n)\}$, $\{a^*(n)\}$, and $\{b^*(n)\}$ be the coefficients associated with $d\rho(\lambda)$ and $d\rho^*(\lambda)$, respectively. If $\sum_{n=1}^{\infty} n\nu(2n)\{|1 - 4a(n)^2| + |b(n-1)|\} < \infty$, then $\sum_{n=1}^{\infty} n\nu(2n)\{|1 - 4a^*(n)^2| + |b^*(n-1)|\} < \infty$.

THEOREM 6. *Let $d\rho(\lambda)$, $\{a(n)\}$, and $\{b(n)\}$ be given as in Theorem 5. Let*

$$d\hat{\rho}(\lambda) = \begin{cases} \sigma(\theta)d\lambda, \\ \sum_{m=1}^{N+1} \rho_m \delta(\lambda - \lambda_m)d\lambda, \end{cases}$$

where $\lambda_{N+1} = (z_{N+1} + 1/z_{N+1})/2$, $|z_{N+1}| < 1/R, |\lambda_{N+1}| \geq |\lambda_N|$. Furthermore, let $\{\hat{a}(n)\}$ and $\{\hat{b}(n)\}$ be the coefficients associated with $d\hat{\rho}(\lambda)$. If

$$\sum_{n=1}^{\infty} n\nu(2n)\{|1 - 4a(n)^2| + |b(n-1)|\} < \infty,$$

then $\sum_{n=1}^{\infty} n\nu(2n)\{|1 - 4\hat{a}(n)^2| + |\hat{b}(n-1)|\} < \infty$.

Thus from Theorem 5 it is clear that one can remove all the mass points and the rate of convergence of the coefficients associated with the new measure will be at least as fast as the rate of convergence of the coefficients associated with the original measure. Theorem 6 shows that if $|z_{N+1}| < 1/R$, that is, if z_{N+1} is not in the maximal ideal space of A_ν, then one can add a finite number of mass points, the only restrictions being that each one must be positive and of finite magnitude, and the rate of convergence of the new coefficients will be at least as fast as the rate of convergence of the original coefficients. It should be noted that the mass points are added or removed without altering the absolutely continuous part of the distribution function.

LEMMA 3. *If (2.6) holds, then*

$$\sum_{i=1}^{\infty} |z^{-i}\psi_+(z,i)| < \infty, \qquad |z| \leq R, \tag{3.8}$$

$$\sum_{i=1}^{\infty} i\nu(2i)|z^{-2i}(p_+(z,i)^2 - p_+(z,i+1)p_+(z,i-1))| < \infty, \qquad |z| \leq R, \tag{3.9}$$

and

$$\sum_{i=1}^{\infty} i\nu(2i)|z^{-2i}(\lambda p_+(z,i)p_+(z,i+1) - a(i+1)p_+(z,i)^2 - a(i+1)p_+(z,i+1)^2)| < \infty, \qquad |z| \leq R. \tag{3.10}$$

Proof. From (2.2) and the boundary conditions satisfied by $\psi_+(z,n)$ we find

$$\frac{z^{-n}\psi_+(z,n)}{\gamma(n+1)}$$
$$= -\sum_{j=n}^{\infty}\left\{\prod_{k=n+1}^{j}(2a(k))^2\right\}\{(1-4a(j+1)^2)z-2b(j)\}\,z^{2j-2n+1}\frac{z^{-j}p_+(z,j)}{\gamma(j+1)}.$$

To obtain (3.8), multiply the above equation by z^{-n}, take magnitudes, then sum on n and use the fact that $\{\prod_{k=n+1}^{j}(2a(k))^2\}|\frac{z^{-j}p_+(z,j)|}{\gamma(j)}| < c$ (see (2.22)).

To show (3.9) and (3.10) set $\hat{w}(n) = p_+(z,n+1) - zp_+(z,n)$ and $w(n) = p_+(z, n+1) - (1/z)p_+(z,n)$. Then

$$p_+(z,n)^2 - p_+(z,n-1)p_+(z,n+1)$$
$$= \frac{-1}{2\lambda}\left[\left\{\hat{w}(n) - \frac{1}{z}\hat{w}(n-1)\right.\right.$$
$$\left.\left. + w(n) - zw(n-1)\right\}p_+(z,n) - w(n)\hat{w}(n-1) - w(n-1)\hat{w}(n)\right].$$

From (2.6) and (2.22) it is apparent that (3.10) will follow if it can be shown that

$$z^{-2n}\left[\lambda p_+(z,n)p_+(z,n+1) - \frac{p_+(z,n)^2}{2} - \frac{p_+(z,n+1)^2}{2}\right] = -z^{-2n}\frac{w(n)\hat{w}(n)}{2}$$

converges fast enough. Since

$$w(n)\hat{w}(n-1) + w(n-1)\hat{w}(n)$$
$$= w(n)\left[\hat{w}(n-1) - \frac{1}{z}\hat{w}(n)\right]$$
$$+ \hat{w}(n)\left[w(n-1) - zw(n)\right] + \left(z + \frac{1}{z}\right)w(n)\hat{w}(n),$$

the result will follow if it can be shown that

$$\sum_{i=1}^{\infty} i\nu(2i)\left|z^{-2i}\left(\hat{w}(i) - \frac{1}{z}\hat{w}(i-1)\right)p_+(z,i)\right| < \infty, \tag{3.11}$$

$$\sum_{i=1}^{\infty} i\nu(2i)|z^{-2i}(w(i) - zw(i-1))p_+(z,i)| < \infty, \tag{3.12}$$

and

$$\sum_{i=1}^{\infty} i\nu(2i)|z^{-2i}w(i)\hat{w}(i)| < \infty. \tag{3.13}$$

From the definition of w and $\hat{w}$ we find

$$\hat{w}(i) - \frac{1}{z}\hat{w}(i-1) = w(i) - zw(i-1)$$
$$= (1-2a(i+1))p_+(z,i+1) - 2b(i)p_+(z,i)$$
$$+ (1-2a(i))p_+(z,i-1).$$

Combining this with (2.22) gives (3.9) and (3.10). It is a consequence of (3.8) and (2.2) that

$$\sum_{i=1}^{\infty}|z^{-i}\hat{w}(i)| \le C\Big\{\sum_{i=1}^{\infty}|1-2a(i+1)||z^{-i+1}p_+(z,i)| + \sum_{i=1}^{\infty}|2b(i)||z^{-i}p_+(z,i)| + \sum_{i=1}^{\infty}|z^{-i-1}\psi_+(z,i)|\Big\} < \infty.$$

Therefore, with $w^*(n) = z^{-n}w(n)$ we find

$$\sum_{i=1}^{\infty} i\nu(2i)|z^{-i}\hat{w}(i)w^*(i)| \le \sum_{i=1}^{\infty}|z^{-i}\hat{w}(i)| \sum_{j=i+1}^{\infty} j\nu(2j)|w^*(j)-w^*(j-1)| \le \sum_{i=1}^{\infty}|z^{-i}\hat{w}(i)| \sum_{j=2}^{\infty} j\nu(2j)|w^*(j)-w^*(j-1)| < \infty;$$

the result follows from (3.12) and the fact that $\lim_{i\to\infty}(z^{-i}p_+(z,i))/\gamma(i+1) = 1$. □

Theorem 6 deals with mass points whose locations are in the region $0 < |z| < 1/R$. The next theorem demonstrates that if one wants to add mass points whose locations lie in $1/R \le |z| < 1$ and still have the coefficients converge at least as fast as the original coefficients, then the normalizations are strictly determined.

THEOREM 7. *Let*

$$d\rho(\lambda) = \begin{cases} \sigma(\theta)d\lambda, & \lambda = \cos\theta, \quad 0 \le \theta \le \pi, \\ \sum_{i=1}^{N-1}\rho_i\delta(\lambda-\lambda_i)d\lambda, & \lambda_i = \frac{(z_i+1/z_i)}{2}, \quad 1 > |z_i| > 1/R, \end{cases}$$

and

$$d\hat{\rho}(\lambda) = \frac{d\rho(\lambda)}{\lambda_N - \lambda} + \rho_N\delta(\lambda-\lambda_N)d\lambda, \qquad |\lambda_N| > |\lambda_i|, \lambda_N > 0, \quad \lambda_N = \frac{(z_N+1/z_N)}{2}, \quad 1 > z_N \ge 1/R.$$

Let $\{p(\lambda,n)\}$, $\{a(n)\}$, $\{b(n)\}$, $\{\hat{p}(\lambda,n)\}$, $\{\hat{a}(n)\}$, *and* $\{\hat{b}(n)\}$ *be the orthonormal polynomials and coefficients associated with* $d\rho(\lambda)$ *and* $d\hat{\rho}(\lambda)$, *respectively. If*

$$\sum_{n=1}^{\infty} n\nu(2n)\{|1-4a(n)^2| + |b(n-1)|\} < \infty$$

and

$$\rho_N = \frac{-(z_N - 1/z_N)}{2f_+(1/z_N)f_+(z_N)}, \tag{3.14}$$

then $\sum_{n=1}^{\infty} n\nu(2n)\{|1-4\hat{a}(n)^2| + |\hat{b}(n-1)|\} < \infty$.

Proof. Expanding $\hat{p}(\lambda,n)$ in terms of $p(\lambda,n)$ yields

$$\hat{p}(\lambda,n) = \frac{\hat{k}(n)}{k(n)}p(\lambda,n) - \frac{k(n-1)}{\hat{k}(n)}p(\lambda,n-1), \tag{3.15}$$

where $\hat{k}(n)$ and $k(n)$ are the leading coefficients of $\hat{p}(\lambda,n)$ and $p(\lambda,n)$, respectively. Multiplying by $d\hat{\rho}(\lambda)$ and integrating gives

$$\hat{k}(n)^2 = k(n)k(n-1)\frac{\int p(\lambda,n-1)d\hat{\rho}(\lambda)}{\int p(\lambda,n)d\hat{\rho}(\lambda)}, \qquad n>0.$$

Since $\lambda_N > |\lambda_i|$, $i=1,\dots,N-1$ the functions of the second kind associated with $\rho(\lambda)$ evaluated at λ_N are positive. This, plus the fact that $p(\lambda_N,n)>0$ for $n\geq 0$, implies that the above integrals do not vanish. Therefore

$$\hat{a}(n+1)^2 = a(n+1)a(n)\frac{\int p(\lambda,n-1)d\hat{\rho}(\lambda)\int p(\lambda,n+1)d\hat{\rho}(\lambda)}{\left(\int p(\lambda,n)d\hat{\rho}(\lambda)\right)^2}, \qquad n>0. \tag{3.16}$$

Squaring (3.15) then multiplying by $d\rho(\lambda)$ and integrating yields

$$\int \hat{p}(\lambda,n)^2(\lambda_N-\lambda)d\hat{\rho}(\lambda) = \frac{\hat{k}(n)^2}{k(n)^2}+\frac{k(n-1)^2}{\hat{k}(n)^2}.$$

Consequently,

$$\lambda_N - \hat{b}(n) = a(n)\left[\frac{\int p(\lambda,n-1)d\hat{\rho}(\lambda)}{\int p(\lambda,n)d\hat{\rho}(\lambda)}+\frac{\int p(\lambda,n)d\hat{\rho}(\lambda)}{\int p(\lambda,n-1)d\hat{\rho}(\lambda)}\right], \qquad n>0,$$

which gives

$$\begin{aligned}\hat{b}(n) = \Bigg[&\lambda_N\int p(\lambda,n)d\hat{\rho}(\lambda)\int p(\lambda,n-1)d\hat{\rho}(\lambda) - a(n)\left(\int p(\lambda,n-1)d\hat{\rho}(\lambda)\right)^2\\ &- a(n)\left(\int p(\lambda,n)d\hat{\rho}(\lambda)\right)^2\Bigg]\Big/\int p(\lambda,n-1)d\hat{\rho}(\lambda)\int p(\lambda,n)d\hat{\rho}(\lambda), \quad n>0.\end{aligned} \tag{3.17}$$

A consequence of (2.10), (2.22), and (2.32) is that $\int p(\lambda,i)d\hat{\rho}(\lambda) = O(z_N^{-i})$ for large enough i. Thus, from (3.16) and (3.17) the result follows if it can be shown that

$$\sum_{i=1}^{\infty} i\nu(2i)|z_N^{2i}|\left|\int p(\lambda,i)d\hat{\rho}(\lambda)^2 - \int p(\lambda,i-1)d\hat{\rho}(\lambda)\int p(\lambda,i+1)d\hat{\rho}(\lambda)\right| < \infty \tag{3.18}$$

and

$$\begin{aligned}\sum_{i=1}^{\infty} i\nu(2i)|z_N^{2i}|\Bigg|&\lambda_N\int p(\lambda,i)d\hat{\rho}(\lambda)\int p(\lambda,i-1)d\hat{\rho}(\lambda)\\ &- a(i)\left(\int p(\lambda,i-1)d\hat{\rho}(\lambda)\right)^2 - a(i)\left(\int p(\lambda,i)d\hat{\rho}(\lambda)\right)^2\Bigg| < \infty.\end{aligned} \tag{3.19}$$

Since $\int p(\lambda,n)d\hat{\rho}(\lambda) = (p_+(z_N,n)/f_+(z_N)) + \rho_N p(\lambda_N,i)$, the terms between the magnitude signs in (3.18) can be recast as

$$\begin{aligned}&= \left[\frac{p_+(z_N,i)}{f_+(z_N)}+\rho_N p(\lambda_N,i)\right]^2 z_N^{2i}\\ &\quad - \left[\frac{p_+(z_N,i+1)}{f_+(z_N)}+\rho_N p(\lambda_N,i+1)\right]\left[\frac{p_+(z_N,i-1)}{f_+(z_N)}+\rho_N p(\lambda_N,i-1)\right]z_N^{2i},\end{aligned}$$

which can be rewritten as

$$
\begin{aligned}
&= \frac{z_N^{2i}}{f_+(z_N)^2}\left[p_+(z_N,i)^2 - p_+(z_N,i+1)p_+(z_N,i-1)\right] \\
&\quad + \frac{\rho_N}{f_+(z_N)} z_N^{2i}\Big[2p_+(z_N,i)p(\lambda_N,i) - p_+(z_N,i-1)p(\lambda_N,i+1) \\
&\qquad\qquad - p_+(z_N,i+1)p(\lambda_N,i-1)\Big] \\
&\quad + z_N^{2i}\rho_N^2\left[p(\lambda_N,i)^2 - p(\lambda_N,i+1)p(\lambda_N,i-1)\right].
\end{aligned}
$$

From Lemma 3 it is apparent that one need only consider the last two terms in the above equation. Since $1 > |z_N| > 1/R$, equation (2.24) can be used, yielding

(3.20)

$$
\begin{aligned}
&= z_N^{2i}\left[\frac{4\rho_N f_+(1/z_N)}{f_+(z_N)(z_N - 1/z_N)} + \frac{4\rho_N^2 f_+(1/z_N)^2}{(z_N - 1/z_N)^2}\right] \\
&\quad \times \Big[p_+(z_N,i)^2 - p_+(z_N,i-1)p_+(z_N,i+1)\Big] \\
&\quad + \frac{4\rho_N^2 f_+(z_N)^2}{(z_N - 1/z_N)^2}\left[p_-(z_N,i)^2 - p_-(z_N,i-1)p_-(z_N,i+1)\right] z_N^{2i} \\
&\quad - \left[\frac{2\rho_N}{(z_N - 1/z_N)} + \frac{4\rho_N^2 f_+(z_N)f_+(1/z_N)}{(z_N - 1/z_N)^2}\right] \\
&\quad \times \left[2p_+(z_N,i)p_-(z_N,i) - p_+(z_N,i+1)p_-(z_N,i-1) - p_+(z_N,i-1)p_-(z_N,i+1)\right] z_N^{2i}.
\end{aligned}
$$

Thus the convergence of (3.16) follows from Lemma 3, (3.14), the fact that $p_-(z,i) = p_+(1/z,i)$, and the fact that $\nu(n) = \nu(-n)$.

Recast (3.17) as

$$
\begin{aligned}
&= \frac{O(z_N^{2i})}{f_+(z_N)^2}\left[\lambda_N p_+(z_N,i)p_+(z_N,i-1) - a(i)p_+(z_N,i-1)^2 - a(i)p_+(z_N,i)^2\right] \\
&\quad + \frac{\rho_N O(z_N^{2i})}{f_+(z_N)}\left[\lambda_N\left(p_+(z_N,i)p(\lambda_N,i-1) + p_+(z_N,i-1)p(\lambda_N,i)\right)\right. \\
&\qquad\qquad \left. - 2a(i)p_+(z_N,i-1)p(\lambda_N,i-1) - 2a(i)p_+(z_N,i)p(\lambda_N,i)\right] \\
&\quad + \rho_N^2 O(z_N^{2i})\left[\lambda_N p(\lambda_N,i)p(\lambda_N,i-1) - a(i)p(\lambda_N,i-1)^2 - a(i)p(\lambda_N,i)^2\right].
\end{aligned}
$$

The convergence of the first term follows from Lemma 3. The convergence of the last two terms can be demonstrated by using the same manipulations that led to (3.20), then by using Lemma 3, (3.14), the fact that $p_-(z,i) = p_+(1/z,i)$, and the symmetry of $v(n)$. □

An analogous result holds for $\lambda_N < 0$, by letting $d\rho(\lambda)/(\lambda_N - \lambda) \to d\rho(\lambda)/(\lambda - \lambda_N)$.

A consequence of (3.2) is the following.

COROLLARY 4.

$$\hat{f}_+(z) = \frac{(z_N - z)}{\sqrt{2z_N}} f_+(z).$$

The following theorem allows one to eliminate mass points in a fashion that inverts the procedure given in Theorem 7.

THEOREM 8. *Let*

$$
d\rho(\lambda) = \begin{cases} \sigma(\theta)d\lambda, & \lambda = \cos\theta, \quad 0 \le \theta \le \pi, \\ \sum_{i=1}^{N} \rho_i \delta(\lambda - \lambda_i)d\lambda, & \lambda \text{ not as above}, \quad \lambda_N > 0, \lambda_N \ge |\lambda_i|, \end{cases}
$$

and

$$d\rho^*(\lambda) = (\lambda_N - \lambda)d\rho(\lambda).$$

Let $\{p(\lambda,n)\}$, $\{a(n)\}$, $\{b(n)\}$, $\{p^*(n)\}$, *and* $\{b^*(n)\}$ *be the orthonormal polynomials and coefficients associated with* $d\rho(\lambda)$ *and* $d\rho^*(\lambda)$, *respectively. If* $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4a(i)^2| + |b(i-1)|\} < \infty$, *then* $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4a^*(i)^2| + |b^*(i-1)|\} < \infty$ *and*

$$f_+^*(z) = \frac{\sqrt{2z_N}}{z_N - z} f_+(z). \tag{3.21}$$

An analogous result holds for $\lambda_N < 0$.

Proof. Expanding $(\lambda_N - \lambda)p^*(\lambda,n)$ in a Fourier series gives

$$(\lambda_N - \lambda)p^*(\lambda,n) = \frac{k(n)}{k^*(n)}p(\lambda,n) - \frac{k^*(n)}{k(n+1)}p(\lambda,n+1), \tag{3.22}$$

where $k^*(n)$ and $k(n)$ are the leading coefficients of $p^*(\lambda,n)$ and $p(\lambda,n)$, respectively. At $\lambda = \lambda_N$ we find

$$\frac{1}{k^*(n)} = \frac{1}{k(n+1)k(n)}\frac{p(\lambda_N,n+1)}{p(\lambda_N,n)}, \tag{3.23}$$

or

$$a^*(n)^2 = a(n)a(n+1)\frac{p(\lambda_N,n+1)p(\lambda_N,n-1)}{p(\lambda_N,n)^2}.$$

Squaring (3.22), then multiplying by $d\rho(\lambda)$ and integrating yields

$$\lambda_N - b^*(n) = \frac{k(n)^2}{k^*(n)^2} + \frac{k^*(n)^2}{k(n+1)^2}.$$

Now, using (3.23) gives

$$\begin{aligned} b^*(n) = (&\lambda_N p(\lambda_N,n)p(\lambda_N,n+1) - a(n+1)p(\lambda_N,n)^2 \\ &- a(n+1)p(\lambda_N,n+1)^2)/p(\lambda_N,n)p(\lambda_N,n+1). \end{aligned} \tag{3.24}$$

Since the mass points occur at the zeros of $f_+(z)$ for $|z| < 1$, we find

$$a^*(n)^2 = a(n)a(n+1)\frac{p_+(z_N,n+1)p_+(z_N,n-1)}{p_+(z_N,n)^2},$$

and

$$\begin{aligned} b^*(n) = (&\lambda_N p_+(z_N,n)p_+(z_N,n+1) - a(n+1)p(z_N,n)^2 \\ &- a(n+1)p_+(z_N,n+1)^2)/p_+(z_N,n)p_+(z_N,n+1). \end{aligned}$$

The convergence of the series now follows from Lemma 3 and the fact that $p_+(z_N,n) \neq 0$ for all finite n. Equation (3.21) follows from (3.2). □

Theorems 7 and 8 demonstrate how to add zeros to $f_+(z)$ for $1 > |z| \geq 1/R$ or remove the zeros of $f_+(z)$ for $1 > |z| > 0$ without decreasing the rate of convergence of the new coefficients. What about the zeros of $f_+(1/z)$ in the region $1 > |z| \geq 1/R$? This breaks down into two cases: the roots of $f_+(1/z)$ that are real and those that are complex. Since the coefficients of $f_+(1/z)$ are real the complex roots of $f_+(1/z)$ come in conjugate pairs. Thus the investigation reduces to considering the effect, on the

coefficients, of multiplying or dividing the distribution function by linear or quadratic polynomials. (For other results in this area see Nevai [17, Chap. 6].)

THEOREM 9. *Given* $d\rho(\lambda)$ *and* $d\hat{\rho}(\lambda) = d\rho(\lambda)/(\lambda_0 - \lambda)$ *with* $\infty > \lambda_0 > 1$, λ_0 *not in the convex hull of the support of* $\rho(\lambda)$, $\lambda_0 = (z_0 + 1/z_0)/2$, $0 < z_0 < 1$. *If* $\sum_{i=1}^{\infty} i\nu(2i)\{|1-4a(i)^2| + |b(i)|\} < \infty$, *then* $\sum_{i=1}^{\infty} i\nu(2i)\{|1-4\hat{a}(i)^2| + |\hat{b}(i)|\} < \infty$ *and*

$$\hat{f}_+(z) = \frac{(1 - z_0 z)}{\sqrt{2z_0}} f_+(z). \tag{3.25}$$

Proof. Use the same procedures that led to (3.16) and (3.17). Then the theorem is a consequence of (2.32), Lemma 3, and (3.2). □

It should be noted that the only restriction on z_0 is that $0 < z_0 < 1$, which may put $1/z_0$ outside the region of analyticity of $zf_+(z)$. However, if $1/z_0$ is within this region or if the region of analyticity can be extended by analytic continuation (see examples) to include the point $1/z_0$, then $\hat{f}_+(z)$ will have a zero there. Analogous results holds for $\lambda_0 < -1$.

THEOREM 10. *Let*

$$d\rho^*(\lambda) = (\lambda_0 - \lambda)d\rho(\lambda),$$

$\lambda_0 > 1$, λ_0 *not in the convex hull of the support of* $\rho(\lambda)$, $\lambda_0 = (z_0 + 1/z_0)/2$ *with* $1 > z_0 \geq 1/R$. *If* $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4a(i)^2| + |b(i)|\} < \infty$, *and* $f_+(1/z_0) = 0$; *then* $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4a^*(i)^2| + |b^*(i)|\} < \infty$, *with* $f_+^*(z) = (\sqrt{2z_0}/(1 - z_0 z))f_+(z)$. *Analogous results hold for* $\lambda_0 < -1$.

Proof. Using procedures similar to those of Theorem 8 we find

$$a^*(n)^2 = a(n)a(n+1)\frac{p(\lambda_0, n+1)p(\lambda_0, n-1)}{p(\lambda_0, n)^2},$$

and

$$\begin{aligned} b^*(n) = (\lambda_0 p(\lambda_0, n)p(\lambda_0, n+1) - a(n+1)p(\lambda_0, n+1)^2 \\ - a(n+1)p(\lambda_0, n)^2)/p(\lambda_0, n)p(\lambda_0, n+1). \end{aligned}$$

Substituting (2.24) into the above equations yields

$$a^*(n)^2 = a(n)a(n+1)\frac{p_-(z_0, n+1)p_-(z_0, n-1)}{p_-(z_0, n)^2},$$

and

$$\begin{aligned} b^*(n) = (\lambda_0 p_-(z_0, n)p_-(z_0, n+1) - a(n+1)p_-(z_0, n+1)^2 \\ - a(n+1)p_-(z_0, n)^2)/p_-(z_0, n)p_-(z_0, n+1). \end{aligned}$$

Since λ_0 is not in the convex hull of the support of $\rho(\lambda)$, $p_-(z_0, n) \neq 0$. Thus the convergence of the series is guaranteed by (2.22), the fact that $p_-(z, n) = p_+(1/z, n)$, Lemma 3, and (3.2). □

THEOREM 11. *Given* $d\rho(\lambda)$ *and*

$$d\hat{\rho}(\lambda) = \frac{d\rho(\lambda)}{(\lambda - A)^2 + B^2} = \frac{d\rho(\lambda)}{(\lambda_1 - \lambda)(\lambda_1^* - \lambda)}, \qquad \lambda_1 = A + iB,$$

with $\lambda_1 = (z_1 + 1/z_1)/2$, $0 < |z_1| < 1$ (λ_1 *on the physical sheet*). *If* $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4a(i)^2| + |b(i-1)|\} < \infty$, *then* $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4\hat{a}(i)^2| + |\hat{b}(i+1)|\} < \infty$ *and* $\hat{f}_+(z) = ((1 - zz_1)(1 - zz_1^*)/2|z_1|)f_+(z)$.

Proof. Expanding $\hat{p}(\lambda, n)$ in terms of $p(\lambda, n)$ gives

$$\begin{aligned}\hat{p}(\lambda, n) = &\frac{\hat{k}(n)}{k(n)} p(\lambda, n) + C(n, n-1) p(\lambda, n-1) \\ &+ \frac{k(n-2)}{\hat{k}(n)} p(\lambda, n-2), \qquad n \geq 2,\end{aligned} \tag{3.26}$$

where $\hat{k}(n)$ and $k(n)$ are the leading coefficients of $\hat{p}(\lambda, n)$ and $p(\lambda, n)$, respectively. To determine $C(n, n-1)$ multiply (3.26) by $\lambda_1 - \lambda$ and integrate with respect $d\hat{\rho}(\lambda)$, then use (2.32) to find

$$C(n, n-1) = \frac{-\hat{k}(n)}{k(n)} \frac{p_+(z_1^*, n)}{p_+(z_1^*, n-1)} - \frac{k(n-2)}{\hat{k}(n)} \frac{p_+(z_1^*, n-2)}{p_+(z_1^*, n-1)}, \qquad n \geq 2.$$

Multiplying (3.26) by $(\lambda_1^* - \lambda) d\hat{\rho}(\lambda)$ and integrating, then using the above equation yields

$$\hat{k}(n)^2 = k(n) k(n-2) \frac{\Delta(n-1)}{\Delta(n)}, \tag{3.27}$$

where $\Delta(n) = [p_+(z_1, n) p_+(z_1^*, n-1) - p_+(z_1^*, n) p_+(z_1, n-1)]$. Therefore

$$\hat{a}(n+1)^2 = a(n+1) a(n-1) \left[\frac{\Delta(n-1)\Delta(n+1)}{\Delta(n)^2} \right]. \tag{3.28}$$

Likewise

$$\hat{b}(n) = \frac{\hat{k}(n, n-1)}{\hat{k}(n)} - \frac{\hat{k}(n+1, n)}{\hat{k}(n+1)},$$

where $\hat{k}(n, n-1)$ is the coefficients of λ^{n-1} in $\hat{p}(\lambda, n)$. From (3.26) we find

$$\frac{\hat{k}(n, n-1)}{\hat{k}(n)} = \frac{k(n, n-1)}{k(n)} - a(n) \frac{p_+(z_1^*, n)}{p_+(z_1^*, n-1)} - \frac{k(n-1)k(n-2)}{\hat{k}(n)^2} \frac{p_+(z_1^*, n-2)}{p_+(z_1^*, n-1)},$$

which, using (3.27), becomes

$$\frac{\hat{k}(n, n-1)}{\hat{k}(n)} = \frac{k(n, n-1)}{k(n)} - a(n) \frac{p_+(z_1^*, n)}{p_+(z_1^*, n-1)} - a(n) \frac{p_+(z_1^*, n-2)}{p_+(z_1^*, n-1)} \frac{\Delta(n)}{\Delta(n-1)}.$$

Therefore,

$$\begin{aligned}\hat{b}(n) = b(n) &- a(n) \frac{p_+(z_1^*, n)}{p_+(z_1^*, n-1)} + a(n+1) \frac{p_+(z_1^*, n+1)}{p_+(z_1^*, n)} \\ &- a(n) \frac{p_+(z_1^*, n-2)}{p_+(z_1^*, n-1)} \frac{\Delta(n)}{\Delta(n-1)} + a(n+1) \frac{p_+(z_1^*, n-1)}{p_+(z_1^*, n)} \frac{\Delta(n+1)}{\Delta(n)}.\end{aligned}$$

Using Lemma 3 we see that the result follows if it can be shown that

$$\sum_{n=2}^{\infty} n \nu(2n) \left| \frac{\Delta(n)^2 - \Delta(n+1)\Delta(n-1)}{\Delta(n)^2} \right| < \infty. \tag{3.29}$$

Since $p_+(z,n)$ satisfies (1.1), it follows that

$$a(n+1)\Delta(n+1) = (\lambda_1 - \lambda_1^*)|p_+(z_1,n)|^2 + a(n)\Delta(n). \tag{3.30}$$

Therefore $\Delta(n) \neq 0$ for finite n and for large n goes as

$$\Delta(n) = O\left(|z_1|^{2n}\left(\frac{1}{z_1^*} - \frac{1}{z_1}\right)\right).$$

Substituting (3.30) in (3.29) we find that the series converges if it can be shown that the following is true:

$$\sum_{n=2}^{\infty} n\nu(2n)|z_1|^{-4n}\big|(\lambda_1 - \lambda_1^*)|p_+(z_1,n)|^2|p_+(z_1,n-1)|^2 - a(n)\Delta(n)(|p_+(z_1,n)|^2 - |p_+(z_1,n-1)|^2)\big| < \infty.$$

If we eliminate $\lambda_1^* p_+(z_1^*,n)$ and $\lambda_1 p_+(z_1,n)$ using (1.1), then the result follows from Lemma 3. □

THEOREM 12. *Given* $d\rho(\lambda)$ *and*

$$d\rho^*(\lambda) = ((\lambda - A)^2 + B^2)d\rho(\lambda) = (\lambda_1 - \lambda)(\lambda_1^* - \lambda)d\rho,(\lambda), \qquad \lambda_1 = A + iB,$$

with $\lambda_1 = (z_1 + 1/z_1)/2$, $1/R \leq |z_1| < 1$.

If $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4a(i)^2| + |b(i-1)|\} < \infty$ and

$$f_+(1/z_1) = 0, \tag{3.31}$$

then $\sum_{i=1}^{\infty} i\nu(2i)\{|1 - 4a^*(i)^2| + |b^*(i-1)|\} < \infty$ and

$$f_+^*(z) = 2|z_1|\frac{f_+(z)}{(1 - zz_1)(1 - zz_1^*)}.$$

Proof. Expanding $(\lambda_1 - \lambda)(\lambda_1^* - \lambda)p^*(\lambda,n)$ in terms of $p(\lambda,n)$ gives

$$(\lambda_1 - \lambda)(\lambda_1^* - \lambda)p^*(\lambda,n) = \frac{k^*(n)}{k(n+2)}p(\lambda,n+2) + C(n+1,n)p(\lambda,n+1) + \frac{k(n)}{k^*(n)}p(\lambda,n). \tag{3.32}$$

Therefore setting $\lambda = \lambda_1^*$ and using (3.31) and (2.24) yields

$$C(n+1,n) = -\frac{k^*(n)}{k(n+2)}\frac{p_-(z_1^*,n+2)}{p_-(z_1^*,n+1)} - \frac{k(n)}{k^*(n)}\frac{p_-(z_1^*,n)}{p_-(z_1^*,n+1)}.$$

Now setting $\lambda = \lambda_1$ in (3.32) and using the above equation gives

$$k^*(n)^2 = k(n+2)k(n)\frac{\tilde{\Delta}(n+1)}{\tilde{\Delta}(n+2)},$$

where $\tilde{\Delta}(n) = [p_-(z_1,n)p_-(z_1^*,n-1) - p_-(z_1^*,n)p_-(z_1,n-1)]$. Therefore

$$a^*(n+1)^2 = a(n+3)a(n+1)\left[\frac{\tilde{\Delta}(n+1)\tilde{\Delta}(n+2)}{\tilde{\Delta}(n+2)^2}\right].$$

Equating coefficients of λ^{n+1} in (3.31) yields

$$\frac{k^*(n,n-1)}{k^*(n)} - (\lambda_1 + \lambda_1^*) = \frac{k(n+2,n+1)}{k(n+2)} - \frac{k(n+1)}{k(n+2)}\frac{p_-(z_1^*,n+2)}{p_-(z_1^*,n+1)} - \frac{k(n)}{k^*(n)^2}\frac{p_-(z_1^*,n)}{p_-(z_1^*,n+1)}.$$

Thus

$$\hat{b}(n) = b(n+2) - a(n+2)\frac{p_-(z_1^*,n+2)}{p_-(z_1^*,n+1)} + a(n+2)\frac{p_-(z_1^*,n+3)}{p_-(z_1^*,n+2)} - a(n+2)\frac{p_-(z_z^*,n)}{p_-(z_1^*,n+1)}\frac{\tilde{\Delta}(n+2)}{\tilde{\Delta}(n+1)} + a(n+3)\frac{p_-(z_1^*,n+1)}{p_-(z_1^*,n+2)}\frac{\tilde{\Delta}(n+3)}{\tilde{\Delta}(n+1)}.$$

The result now follows using manipulations similar to those used in the preceding theorem. □

4. The absolutely continuous part. In this section the relation between the decay of coefficients in the recurrence formula and the decay of the Fourier coefficients of the absolutely continuous part of the measure will be investigated. We begin by recalling a result of Baxter.

Let $\sigma(\theta)$ be a real nonnegative periodic function integrable on $[-\pi,\pi]$. Let

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-in\theta}\sigma(\theta)d\theta, \qquad c_0 \neq 0,$$

and define

$$D_n(\sigma) = \begin{vmatrix} c_0 & c_{-1} & \cdots & c_{-n} \\ c_1 & c_0 & \cdots & c_{-n+1} \\ \vdots & \vdots & & \vdots \\ c_n & c_{n-1} & \cdots & c_0 \end{vmatrix}, \qquad n = 0,1,2,\ldots.$$

(Note that $D_n(\sigma) > 0$.) Let

$$\gamma(n) = \frac{(-1)^n}{D_{n-1}(\sigma)}\begin{vmatrix} c_1 & c_0 & \cdots & c_{-n+2} \\ c_2 & c_1 & \cdots & \\ \vdots & \vdots & & \vdots \\ c_n & c_{n-1} & \cdots & c_1 \end{vmatrix}.$$

LEMMA 4. *Given $\sigma(\theta)$ as above assume (abusing notation) $|\sigma(z)| < \infty$ for $1/R \leq |z| \leq R$, $R \geq 1$. Then $\ln\sigma(z) \in A_\nu$ if and only if $\sum_{n=1}^{\infty}\nu(n)|\gamma(n)| < \infty$. Here $\nu(n)$ is given in* (1.4).

Remark. Baxter proved this theorem for $R = 1$ (see Baxter [3], [4]), and his proof with slight modifications can be carried over to this case.

COROLLARY 5. *Let $\{\phi(z,n)\}$ be the orthonormal polynomials associated with $\sigma(\theta)$ (see* Geronimus [12]). *If $\ln\sigma(z) \in A_\nu$, then $\{z^n\bar{\phi}(1/z,n)\}$ is a Cauchy sequence in A_ν and converges in norm to a nonzero function.*

Proof. See Baxter [3], [4]. □

THEOREM 13. *Let $\rho(\lambda)$ be a bounded, nondecreasing, absolutely continuous function on $[-1,1]$ with*

$$d\rho(\lambda) = \sigma(\theta)d\lambda, \qquad \lambda = \cos\theta, \qquad 0 \le \theta \le \pi.$$

Furthermore, let $\sigma(\theta)/\sin\theta = \sigma(-\theta)/\sin(-\theta)$ and $\ln\sigma(z)/(z-1/z) \in A_\nu$. Suppose $R > 1$ in (1.4). Then

$$\sum_{n=1}^{\infty} n\nu(2n)\{|1-4a(n)^2| + |b(n-1)|\} < \infty \tag{4.1}$$

if and only if

$$\sum_{n=1}^{\infty} n\nu(n)|g(n)| < \infty. \tag{4.2}$$

Here $\sigma(z)/(z-1/z) = \sum_{m=-\infty}^{\infty} g(m)z^m$, $1/R \le |z| \le R$, with $g(m) = g(-m)$.

Proof. Without loss of generality assume $\int_{-1}^{1} d\rho(\lambda) = 1$. From Theorems 1 and 4 we find that (4.1) implies that

$$\frac{\sigma(z)}{z-1/z} = \frac{i}{\pi f_+(z)f_+(1/z)}, \qquad z = e^{i\theta}, \tag{4.3}$$

with $zf_+(z) \in A_\nu^+$. Since $\ln\sigma(z)/(z-1/z) \in A_\nu$, $(zf_+(z))^{-1} \in A_\nu$, which implies that (4.3) holds for $1/R \le |z| \le R$. Differentiating (4.3) with respect to z yields

$$\sum_{m=-\infty}^{\infty} mq(m)z^{m-1} = -\frac{i}{\pi}\left[\frac{(zf_+(z))'}{(zf_+(z))^2 1/zf_+(1/z)} + \frac{(1/zf_+(1/z))'}{zf_+(z)(1/zf_+(1/z))^2}\right].$$

From Lemma 1 and Theorem 1(v) we find that $(zf_+(z))'$ and $(1/zf_+(1/z))' \in A_\nu$ since $R > 1$. Hence (4.1) implies (4.2).

To prove sufficiency note that $\ln\sigma(z)/(z-1/z) \in A_\nu$ and $\sigma(\theta)/\sin\theta = \sigma(-\theta)/\sin(-\theta)$ imply that we can construct $f_+(z)$ such that $zf_+(z) \in A_\nu^+$, $zf_+(z) \ne 0$, $|z| \le R$, $zf_+(z)|_{z=0} > 0$, and

$$\frac{2i\sigma(z)}{z-1/z} = \frac{1}{2\pi z f_+(z) 1/zf_+(1/z)}, \qquad 1/R \le |z| \le R.$$

Differentiating the above equation with respect to z, multiplying by $zf_+(z)(1/z)f_+(1/z)$, applying E_+, the operator that projects A_ν onto A_ν^+, then multiplying by $zf_+(z)$ gives

$$\begin{aligned}(zf_+(z))' = &-zf_+(z)E_+\left\{\frac{(1/zf_+(1/z))'}{1/zf_+(1/z)}\right\} \\ &- 2izf_+(z)E_+\left\{zf_+(z)\frac{1}{z}f_+(1/z)\left(\frac{2i\sigma(z)}{z-1/z}\right)'\right\}.\end{aligned}$$

Thus it follows from (4.2) that $(zf_+(z))' \in A_\nu$. Defining

$$s(z) = \frac{f_+(1/z)}{f_+(z)},$$

we find from above that $(s(z))' \in A_\nu$. The polynomials $\{p(\lambda,n)\}$ orthonormal with respect to $\sigma(\theta)$ can be written as (see Szegö [22])

$$(z-1/z)p(\lambda,n) = \sqrt{\frac{2}{\pi}}\left(1-\frac{\phi(0,2n+2)}{k(2n+2)}\right)^{-1/2} \times [z^{-n-1}\phi(z,2n+2) - z^{n+1}\phi(1/z,2n+2)], \qquad \lambda = \frac{1}{2}(z+1/z),$$

where $\{\phi(z,n)\}$ are polynomials on the unit circle orthonormal with respect to $\sigma(\theta)/\sin\theta$ with leading coefficient $k(n)$. By Corollary 5, $(z-1/z)p(\lambda,n) \in A$ for all n which implies through (2.32) that

$$p_+(z,n) = A(n,n)z^n\left(1+\sum_{i=1}^{\infty}\alpha(n,i)z^i\right) \in A \quad \text{for all } n,$$

since

$$\int \frac{p(\lambda',n)\sigma(\lambda')}{\lambda-\lambda'}d\lambda' = -\frac{1}{2i}\int_{-\pi}^{\pi} p(\cos\theta',n)\sigma(\theta')\left(\frac{e^{i\theta'}+z}{e^{i\theta'}-z}\right)d\theta'.$$

It is known that the Fourier coefficients of $p_+(z,n)$ satisfy the discrete analog of the Marchenko equation [5], [8], [10]:

$$w(2n+m)+\alpha(n,m)+\sum_{i=1}^{\infty}\alpha(n,i)w(i+2n+m) = 0, \qquad m \geq 1, n \geq 0$$

with

$$w(n) = \frac{1}{2\pi i}\oint (1-s(z))z^n\frac{dz}{z}.$$

Therefore

$$\begin{aligned} & nv(2n+2)|\alpha(n,m)-\alpha(n+1,m)| \\ & \quad \leq nv(2n+2)|w(2n+m+2)-w(2n+m)| \\ & \quad\quad + \sum_{i=1}^{\infty} nv(2n+2)|w(2n+i+m+2)-w(2n+m+i)||\alpha(n+1,i)| \\ & \quad\quad + \sum_{i=1}^{\infty} nv(2n+2)|\alpha(n+1,i)-\alpha(n,i)||w(i+2n+m)|. \end{aligned} \tag{4.4}$$

Since $\sum_{n=1}^{\infty}|w(n)| < \infty$ there exists $w_1(n)$ and $\hat{w}(n)$ such that

$$w_1(n) = \begin{cases} w(n)-\hat{w}(n), & w_1(n),\hat{w}(n) \neq 0, \quad n \leq N, \\ w(n), & n > N, \end{cases}$$

and $\sum_{n=1}^{\infty}|w_1(n)| < 1$. Replace $w(n)$ by $w_1(n)+\hat{w}(n)$ in the third term on the right-hand side of (4.4), iterate, then sum n from 1 to infinity. Setting $\gamma = \sup_m \sum_{n=1}^{\infty} nv(2n+2)|\alpha(n,m)-\alpha(n+1,m)|$, and using the fact that $s'(z) \in A_v$ we find

$$\gamma \leq \frac{2c\sup_m \sum_{n=1}^{\infty} nv(2n+2)\{|w(2n+m+2)-w(2n+m)|\}}{1-\sum_{i=1}^{\infty}|w_1(i)|} + \frac{\{\sup_{0\leq m\leq N}\sum_{n=1}^{N} nv(2n+2)|\alpha(n+1,m)-\alpha(n,m)|\}\sum_{i=1}^{N}|\hat{w}(i)|}{1-\sum_{i=1}^{\infty}|w_1(i)|} < \infty.$$

In the above equation $c = \max(\sup_i \sum_{n=1}^{\infty} |\alpha(n,i)| < \infty, 1)$. Thus $\sum_{n=1}^{\infty} nv(2n+2)|\alpha(n,m) - \alpha(n+1,m)| < \infty$ for all m, which implies the result since $\alpha(n,1) - \alpha(n-1,1) = 2b(n)$ and $\alpha(n,2) - \alpha(n-1,\gamma \to 2) = 1 - 4a(n+1)^2 + 2b(n)\alpha(n,n+1)$. □

Remark. We note that the case $R = 1$ does not follow from the above theorem. For $R = 1$, (4.2) needs to be replaced by $\sum_{n=1}^{\infty} nv(n)|g(n) - g(n+2)| < \infty$ (Geronimo[8]* and Guseinov [13]). This is because Lemma 1 does not hold for $z = \pm 1$ in this case; consequently, part (v) in Theorem 1 cannot be strengthened.

Unfortunately, without further hypotheses an analog of Theorem 1 in [11] is unavailable at this time. This is due to the fact that we do not have control on the number of zeros of $zf_+(z)$ for $1 < |z| < R$. Also, although it is clear that one can add or remove a finite number of zeros from $zf_+(z)$ and not decrease the rate of convergence of the coefficients in the recurrence formula, it is not clear whether an infinite number can be added or removed without decreasing the rate of convergence. Some statements about the zeros can be made. First let H_R^2 be the Hilbert space of functions analytic inside and square integrable on the boundary of the disk of radius R, i.e., $g(z) \in H_R^2$ if $g(z) = \sum_{n=0}^{\infty} c(n)z^n$, $|z| < R$ and $(1/2\pi)\int_{-\pi}^{\pi} |g(\mathrm{Re}^{i\theta})|^2 d\theta = \sum_{n=0}^{\infty} |c(n)|^2 R^{2n} < \infty$. Since $zf_+(z)$ is an element of H_R^2 its zeros satisfy the Blaschke condition $\sum_{i=1}^{\infty}(1 - |z_i|/R) < \infty$.

The results of §3, Theorem 13 in §4, and Theorem 6 in [11] can be combined to give the following.

THEOREM 14. *Let $\rho(\lambda)$ be a positive measure with absolutely continuous part $\sigma(\theta), \lambda = \cos\theta, 0 < \theta < \pi$. Set $\sigma(-\theta)/\sin(-\theta) = \sigma(\theta)/\sin(\theta)$, $0 < \theta < \pi$, and suppose $\sigma(z)$ has a meromorphic extension to $\frac{1}{R} < z < R > 1$ such that $|\sigma(z)| < \infty$ for $|z| = R$ and $|z| = 1/R$. Let $\{z_i\}$ denote the singularties of $\sigma(z)/(z - 1/z)$, including multiplicities. Then $\sum_{n=1}^{\infty} n\nu(2n)\{|1 - 4a(n)^2| + |b(n-1)|\} < \infty$ if and only if*

$$d\rho(\lambda) = \begin{cases} \sigma(\theta)d\lambda, & \lambda = \cos\theta, \quad 0 \le \theta \le \pi, \\ \sum_{i=1}^{N} \rho_i \delta(\lambda - \lambda_i)d\lambda, & \lambda \notin [-1,1], \quad N < \infty, \end{cases}$$

$\ln(\sigma(z)d(z)/(z - 1/z)) \in A_v$, and $\sum_{n=-\infty}^{\infty} n\nu(n)|g(n)| < \infty$, where $d(z) = \prod_{i=1}^{M}(z - z_i)((1/z) - z_i), |z_i| \le 1, M < \infty$, with $d(e^{i\theta}) = \overline{d(e^{i\theta})}$, and $\sigma(z)d(z)/(z - 1/z) = \sum_{n=-\infty}^{\infty} g(n)z^n$.

5. Examples—the Askey–Wilson polynomials. We now apply the above theorems to the Askey–Wilson polynomials. These give rise to the q-analogs of many of the classical polynomials [1], [2], [15]. In the examples below the recurrence coefficients $a(n)^2$ and $b(n)$ have the form

$$4a(n)^2 = 1 + \sum_{i=1}^{\infty} q^{ni}\gamma_i(q) \tag{5.1}$$

and

$$b(n) = \sum_{i=1}^{\infty} q^{ni}\beta_i(q), \tag{5.2}$$

*We wish to point out that Theorem 3.1 and its consequence, Lemma 3.1 in Geronimo [8], are incorrect. The correct proof of the sufficiency part of Theorem 1 in [8] is given in Appendix A of [8] (see also Theorem 13 above).

where $\gamma_i(q)$ and $\beta_i(q)$ are rational functions of q independent of n and $0 \le q < 1$. In order for (5.1) and (5.2) to converge we will assume the $\limsup |\gamma_i(q)|^{1/i} < 1/q$ and $\limsup |\beta_i(q)|^{1/i} < 1$. Coefficients having the form (5.1) and (5.2) have analogs in the scattering theory literature of physics and are called Yukawa type potentials [20]. We wish to analyze the structure of $p_+(z,n)$ with coefficients of the form (5.1) and (5.2). To this end let $\hat{p}_+(z,n) = z^{-n}p_+(z,n)/\gamma(n+1)$ with $\gamma(n+1)$ given in the proof of Theorem 2. In this case (1.1) becomes

$$z^2 4a(n+1)^2\hat{p}_+(z,n+1) + 2b(n)z\hat{p}_+(z,n) + \hat{p}_+(z,n-1) = 2z\lambda\hat{p}_+(z,n). \tag{5.3}$$

Using an idea of Martin [18], we make the ansatz $\hat{p}_+(z,n) = 1 + \sum_{i=1}^{\infty} h_i(q,z)q^{ni}$ and substitute this into (5.3). Let $x = q^n$ and note that by varying n [15] we may consider x as an independent variable. Equating powers of x^k in (5.3) we find that

$$h_k = \frac{-zq^k(\gamma_k q^k z + 2\beta_k + \sum_{i=1}^{k-1}(z\gamma_i q^k + 2\beta_i)h_{k-i})}{(1-z^2q^k)(1-q^k)}, \tag{5.4}$$

which implies that $h_k(z)$ is a rational function of z with possible poles located only at the zeros of the function $(z^2q : q)_{k-1}$, where

$$(a : q)_k = \begin{cases} (1-a)(1-aq)\cdots(1-aq^{k-1}), & k \ge 1, \\ 1, & k = 0. \end{cases}$$

In order to show that the equation for $p_+(z,n)$ given above is valid, define $\hat{h}_k = (z^2q : q)_{k-1}(q : q)_{k-1}h_k$; then (5.4) reads as

$$\begin{aligned} \hat{h}_k = -zq^k \Bigg[& (\gamma_k q^k z + 2\beta_k)(z^2q : q)_{k-2}(q : q)_{k-2} \\ & + \sum_{i=1}^{k-1}(z\gamma_{k-i}q^k + 2\beta_{k-i})\frac{(z^2q : q)_{k-1}(q : q)_{k-1}\hat{h}_i}{(z^2q : q)_{i-1}(q : q)_{i-1}} \Bigg]. \end{aligned} \tag{5.5}$$

Let S be a compact subset of the complex plane. Since $|(z^2q : q)_k| \le e^{\sum_{i=1}^{k}|z^2|q^i}$ there exists a constant c depending only upon the set S such that $|(z^2q : q)_k| < c$ and

$$\left|\frac{(z^2q : q)_{k-1}(q : q)_{k-1}}{(z^2q : q)_{i-1}(q : q)_{i-1}}\right| < c$$

for all $k > 0$ and $0 \le i \le k$. Consequently, (5.5) becomes

$$|\hat{h}_k| \le c\left(w_k + \sum_{i=1}^{k-1} w_{k-i}q^i|\hat{h}_i|\right),$$

where $w_k = |z|q^k(|\gamma_k z| + 2|\beta_k|)$. Therefore,

$$\sum_{k=1}^{\infty}|\hat{h}_k| \le c\left(\sum_{k=1}^{\infty} w_k + \sum_{i=1}^{\infty}|\hat{h}_i|q^i\sum_{k=0}^{\infty} w_{k+1}\right).$$

Since $\sum_{k=1}^{\infty}|w_k|<\infty$ there exists an N depending only upon S so that $q^n\sum_{i=1}^{\infty}|w_i|c<1$ for all $n\geq N$. Consequently,

$$\sum_{i=N}^{\infty}|\hat{h}_i|q^i\sum_{k=1}^{\infty}w_k\leq\sum_{i=N}^{\infty}|\hat{h}_i|q^N\sum_{k=1}^{\infty}w_k,$$

which leads to

$$\sum_{k=N}^{\infty}|\hat{h}_k|\leq\frac{c\left(\sum_{k=1}^{\infty}w_k+\sum_{k=1}^{N-1}|\hat{h}_k|+\sum_{k=1}^{N}|\hat{h}_k|q^k\sum_{j=1}^{\infty}w_j\right)}{1-q^Nc\sum_{k=1}^{\infty}|w_k|}<\infty$$

since the above is true for every compact subset of the complex plane, and since $\{\hat{h}_k\}$ is a sequence of polynomials in z we have shown the following.

THEOREM 15. *Suppose that $a(n)$ and $b(n-1)$ have the form* (5.1) *and* (5.2), *respectively, for $n\geq 1$, with $\limsup|\gamma_i|^{1/i}<1/q$ and $\limsup|\beta_i|^{1/i}<1$. Then for each $n\geq -1$, $p_+(z,n)$ is a meromorphic function of z having its possible poles located only at the zeros of $(z^2q:q)_\infty$.*

Remark. It is left as an exercise to show that the above theorem can be extended to the difference equation

$$z^24a(n+1)^2\hat{p}_+(z,n+1)+2b(n)z\hat{p}_+(z,n)+d(n)\hat{p}_+(z,n-1)=2z\lambda e(n)\hat{p}_+(z,n),\quad n\geq 0,$$

where $d(n)=1+\sum_{i=1}^{\infty}d_i(q)q^{ni}$ and $e(n)=1+\sum_{i=1}^{\infty}e_i(q)q^{ni}$ with $\limsup|d_i(q)|^{1/i}<1$ and $\limsup|e_i(q)|^{1/i}<1$.

We now apply our results to the Askey–Wilson polynomials [1], [2], [14], [15], [19]. In this case $a_n^2=\frac{1}{4}A_{n-1}C_n$, $n=1,2,\ldots$, and $b_n=\frac{1}{2}(a+1/a-A_n-C_n)$, $n=0,1,2,\ldots$, where

$$A_n=\frac{a^{-1}(1-abq^n)(1-acq^n)(1-adq^n)(1-abcdq^{n-1})}{(1-abcdq^{2n-1}(1-abcdq^{2n-2})},\qquad n=0,1,2,\ldots,$$

and

$$C_n=\frac{a(1-bcq^{n-1})(1-bdq^{n-1})(1-cdq^{n-1})(1-q^n)}{(1-abcdq^{2n-1}(1-abcdq^{2n-2})},\qquad n=0,1,2,\ldots,$$

where a,b,c,d are chosen so that A_n and C_n are real and $A_{n-1}C_n>0$. Since $1/a-A_n=O(q^n)$, $a-C_n=O(q^n)$ and $1-A_{n-1}C_n=O(q^n)$ the results of the previous section apply when $|q|<1$. Thus the distribution function $\rho(\lambda)$ with respect to which of the polynomials associated with the above recurrence coefficient are orthogonal, satisfies Theorem 4. Furthermore, from Theorem 15 and the remark below it we find that $p_+(z,n)$ is meromorphic with its possible poles located at the zeros of $(z^2q:q)_\infty$. The equation $p_+(z,n)/f_+(z)$ has been computed for these systems by Rahman [19] and Ismail and Rahman [15]. When a,b,c, and d are real or come in complex conjugate pairs and if $\max(|a|,|b|,|c|,|d|)<1$, then Askey and Wilson [2] have shown that $d\rho(\lambda)=\sigma(\lambda)d\lambda$, where

$$\frac{\sigma(\lambda)}{\sin\theta}=\frac{1}{2\pi}\frac{1}{|f_+e^{i\phi}|},\qquad \lambda=\cos\theta,$$

with

$$zf_+(z) = \frac{(az,q)_\infty (bz,q)_\infty (cz,q)_\infty (dz,q)_\infty}{(z^2q:q)_\infty} k(q),$$

where

$$k(q)^2 = \frac{1}{4} \frac{(abcd:q)_\infty}{(ab:q)_\infty (ac:q)_\infty (ad:q)_\infty (bc:q)_\infty (bd:q)_\infty (cd:q)_\infty (q:q)_\infty}.$$

If any of the parameters a, b, c or d become greater than one, then $\rho(\lambda)$ will have an absolutely continuous part and some mass points. This case has been considered by Askey and Wilson [2, Thm. 2.5].

Finally, we note that the error term for the zeros of the polynomials given by Theorem 3 can be improved; in particular, $\theta = k\pi/(n+1) + 1/(n+1) \arg(e^{i\theta} f_+(e^{i\theta})) + O(q^n/n)$. Successive iteration of this equation beginning with $\theta_0 = k\pi/(n+1)$ yields improving estimates for θ. To see this suppose for convenience that a, b, c, and d are real and less than one in magnitude. Then with $\theta_0 = \pi(n-m)/(n+1) = \pi(1 - (m+1)/(n+1))$ we find $\arg(1 - ae^{i\theta_0} q^k) = -a\pi((m+1)/(n+1))(q^k/(1+aq^k)) + O(q^k/n^3)$. Consequently, $\arg(ae^{i\theta_0}, q)_\infty = -a\pi(m+1)/(n+1) \sum_{k=0}^{\infty} q^k/(1+aq^k) + O(1/n^3)$. Set

$$c(a,q) = \sum_{k=0}^{\infty} \frac{aq^k}{1+aq^k}.$$

Applying similar methods to the other factors in the formula for $zf_+(z)$ yields

$$\arg e^{i\theta_0} f_+(e^{i\theta_0}) = \pi \Big[c(a,q) + c(b,q) + c(c,q) + c(d,q) - 2c(q,q) \Big] \frac{m+1}{n+1} + O\left(\frac{1}{n^3}\right).$$

Thus we have proved the following.

THEOREM 16. *Suppose a, b, c, and d are all real and have magnitudes less than one. Then the zeros of $p_n(\cos\theta)$ are given by*

$$\theta = \pi \frac{(n-m)}{n+1} + \pi \Big[c(a,q) + c(b,q) + c(c,q) + c(d,q) - 2c(q,q) \Big] \frac{m+1}{(n+1)^2} + O\left(\frac{1}{n^4}\right).$$

Acknowledgments. The author would like to thank W. Van Assche; without his strong encouragement, this paper may never have seen the light of day.

REFERENCES

[1] R. ASKEY AND M. E.-H. ISMAIL, *A generalization of ultra-spherical polynomials,* in Studies in Pure Mathematics, P. Erdös, ed., Birkhäuser, Basel, Switzerland, 1983, pp. 55–78.

[2] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials,* Mem. Amer. Math. Soc., 319 (1985).

[3] G. BAXTER, *A convergence equivalence related to polynomials orthogonal on the unit circle,* Trans. Amer. Math. Soc., 99 (1961), pp. 471–487.

[4] ———, *A norm inequality for a finite-section Wiener–Hopf equation,* Ill. J. Math., 7 (1963), pp. 97–103.

[5] K. M. CASE, *Orthogonal polynomials from the viewpoint of scattering theory,* J. Math. Phys., 15 (1974), pp. 2166–2174.

[6] T. S. CHIHARA AND P. G. NEVAI, *Orthogonal polynomials and measures with finitely many point masses,* J. Approx. Theory, 35 (1982), pp. 370–380.

[7] J. FAVARD, *Sur les polynomes de Tchebycheff,* C. R. Acad. Sci. Paris, 200 (1935), pp. 2052–2053.

[8] J. S. GERONIMO, *A relation between the coefficients in the recurrence formula and the spectral function for orthogonal polynomials,* Trans. Amer. Math. Soc., 260 (1980), pp. 65–82.

[9] ———, *On the spectrum of infinite dimensional Jacobi matrices,* J. Approx. Theory, 53 (1988), pp. 251–265.

[10] J. S. GERONIMO AND K. M. CASE, *Scattering theory and polynomials orthogonal on the unit circle,* Trans. Amer. Math. Soc., 258 (1980), pp. 467–494.

[11] J. S. GERONIMO AND P. G. NEVAI, *Necessary and sufficient conditions relating the coefficients in the recurrence formula to the spectral function for orthogonal polynomials,* SIAM J. Math. Anal., 14 (1983), pp. 622–637.

[12] Y. L. GERONIMUS, *Polynomials Orthogonal on a Circle and Interval,* I. N. Sneddon, ed., Pergamon Press, New York, 1960.

[13] G. S. GUSEINOV, *The determination of an infinite Jacobi matrix from the scattering data,* Soviet Math. Dokl., 17 (1976), pp. 596–600.

[14] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q–Jacobi and* ${}_4\phi_3$ *polynomials,* J. Approx. Theory, 36 (1982), pp. 43–54.

[15] M. E. H. ISMAIL AND M. RAHMAN, *The associated Askey–Wilson polynomials,* Trans. Amer. Math. Soc., 328 (1991), pp. 201–237.

[16] M. G. KREIN, *Integral equations on a half-line with kernels depending upon the difference of the arguments,* Amer. Math. Soc. Trans., 22 (1962), p. 163.

[17] P. G. NEVAI, *Orthogonal polynomials,* Mem. Amer. Math. Soc., 18 (1979).

[18] A. MARTIN, *On the analytic properties of partial wave scattering amplitudes obtained from the Schrödinger equation,* Nuovo Cimento, 14 (1959), pp. 403–425.

[19] M. RAHMAN, *q-Wilson functions of the second kind,* SIAM J. Math. Anal., 17 (1986), pp. 1280–1286.

[20] V. DE ALFARO AND T. REGGE, *Potential Scattering,* North-Holland, Amsterdam, 1965.

[21] W. VAN ASSCHE, *Asymptotics for orthogonal polynomials and three-term recurrence,* in Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., NATO ASI Series, Kluwer Acedemic, Dordrecht, 1990, pp. 435–462.

[22] G. SZEGÖ, *Orthogonal polynomials,* Amer. Math. Soc. Coll. Publ., 23 (1975).

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 420–428, March 1994

011

ON THE ASYMPTOTICS OF THE TRICOMI–CARLITZ POLYNOMIALS AND THEIR ZERO DISTRIBUTION (I)*

WILLIAM M. Y. GOH† AND JET WIMP‡

Abstract. The asymptotic behavior of the Tricomi–Carlitz polynomials in the complex plane is established.

Key words. orthogonal polynomials, zero distribution, Stieltjes transform, uniform convergence, theorem of Grommer and Hamburger

AMS subject classifications. 33A65, 41A60

1. Introduction.

Tricomi [11] studied the polynomials

$$t_n^{(\alpha)}(x) = \sum_{k=0}^{n} (-1)^k \binom{x-\alpha}{k} \frac{x^{n-k}}{(n-k)!}, \tag{1}$$

which satisfy the recurrence

$$\begin{aligned} (n+1)t_{n+1}^{(\alpha)}(x) - (n+\alpha)t_n^{(\alpha)}(x) + xt_{n-1}^{(\alpha)}(x) = 0, \qquad & n \geq 1, \\ t_0^{(\alpha)}(x) = 1, \qquad t_1^{(\alpha)}(x) = \alpha. & \end{aligned} \tag{2}$$

(For a brief treatment of these polynomials, see Chihara's book [3].)

We observe, as did Tricomi himself, that $\{t_n^{(\alpha)}(x)\}$ is not a system of orthogonal polynomials, the recurrence relation failing to have the required form. However, Carlitz [2] discovered that if one sets

$$f_n(x) = x^n t_n^{(\alpha)}(x^{-2}), \tag{3}$$

then $\{f_n(x)\}$ satisfies

$$\begin{aligned} (n+1)f_{n+1}(x) - (n+\alpha)xf_n(x) + f_{n-1}(x) = 0, \qquad & n \geq 1, \\ f_0(x) = 1, \qquad f_1(x) = \alpha x. & \end{aligned} \tag{4}$$

Carlitz proved that for $\alpha > 0$, $\{f_n(x)\}$ satisfies the orthogonality relation

$$\int_{-\infty}^{\infty} f_m(x) f_n(x) d\psi^{(\alpha)}(x) = \frac{2e^{\alpha}}{(n+\alpha)n!}\delta_{mn}, \tag{5}$$

where $\psi^{(\alpha)}(x)$ is a step function whose jumps are

$$d\psi^{(\alpha)}(x) = \frac{(k+\alpha)^{k-1}e^{-k}}{k!} \quad \text{at } x = \pm(k+\alpha)^{-1/2}, \quad k = 0, 1, 2, \ldots. \tag{6}$$

The generating function

$$\exp\left\{\frac{w}{x} + \frac{1-\alpha x^2}{x^2}\log(1-wx)\right\} = \sum_{n=0}^{\infty} f_n(x) w^n \tag{7}$$

* Received by the editors April 13, 1992; accepted for publication (in revised form) May 24, 1993.

† Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104. This author's research was partially supported by National Science Foundation grant DMS 9101753.

‡ Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104. This author's research was partially supported by National Science Foundation grant DMS 8901610.

follows easily from the recurrence.[1] This series converges for $|w| < |x|$, whenever $x \neq 0$.

For a generalization of the Tricomi–Carlitz polynomials the reader is referred to [1] and [4].

Note that the support of the orthogonality measure is just the set $\{0\} \cup \{\pm(k+\alpha)^{-1/2} : k = 0, 1, 2, \ldots\}$, which has a single accumulation point at 0. To investigate the asymptotic behavior of the zeros of a set of orthogonal polynomials $P_n(x)$ defined on a compact set it is useful to introduce the measures $\{v_n : n = 0, 1, 2, \ldots\}$,

$$\begin{aligned} &v_n(\{x_{j,n}\}) = 1/n, \qquad j = 1, 2, \ldots, n \\ &v_n(A) = 0, \qquad A \text{ contains no zeros of } P_n(x), \end{aligned} \tag{8}$$

where $\{x_{j,n}\}$ are the zeros of $P_n(x)$. Note the measure v_n assigns an equal weight to each zero. The weak limit of v_n (if it exists) is called the distribution of the zeros of $P_n(x)$.

The asymptotic behavior of v_n gives substantial information about how the zeros are distributed in the interval of orthogonality. For an extensive discussion of applications and consequences of this approach, we refer the reader to [6], [7], and [12].

We define the Stieltjes transform of the measure μ:

$$S(\mu; z) = \int_{-\infty}^{\infty} \frac{d\mu}{z - x}. \tag{9}$$

This transform will prove to be a very useful tool in our investigation.

The distribution of the zeros of the Tricomi–Carlitz polynomials can be easily obtained by the following argument:[2] The zeros of f_n are the eigenvalues of the Jacobi matrix J_n (see [13]). Let a_n's be the recurrence coefficients of the orthonormal polynomial $p_n(x) = [(n+\alpha)n!/(2e^{\alpha})]^{1/2} f_n(x)$, i.e.,

$$x p_n(x) = a_{n+1} p_{n+1}(x) + a_n p_{n-1}(x). \tag{10}$$

Here,

$$a_n = \left(\frac{n}{(n+\alpha)(n+\alpha-1)} \right)^{1/2}. \tag{11}$$

We have

$$S(v_n; x) = n^{-1} \operatorname{trace}(xI - J_n)^{-1},$$

whose moments are $\int x^n dv_n = n^{-1} \operatorname{trace} J_n^k$.

Thus the second moment is

$$\frac{2}{n} \sum_{1}^{n-1} a_j^2 = \frac{2}{n} \sum_{1}^{n-1} \frac{j}{(j+\alpha)(j+\alpha-1)} \sim \frac{2 \ln n}{n}. \tag{12}$$

We see that v_n tends towards the δ-distribution as $n \to \infty$.

[1] There is a misprint in this equation in [3].

[2] This procedure was suggested by a referee.

Although the distribution can be worked out with the above approach, it gives no information about the asymptotics of the orthogonal polynomials in the complex plane. Our approach provides this more general information (e.g., see Corollary 2). Note the Jacobi matrix J_n is a compact operator (because $a_n = O(1/n^{1/2})$ here). We mention a theorem by Schwarz [8], which says that for a trace class Jacobi matrix (one for which $\sum_{n=1}^{\infty} a_n < \infty$) there exists a function f analytic in $\mathbf{C}\backslash\{0\}$ such that

$$\lim_{n\to\infty} \frac{p_n(x)}{x^n} = f(x). \tag{13}$$

The above convergence holds uniformly on compact subsets of $\mathbf{C}\backslash\{0\}$. However, for the Tricomi–Carlitz polynomials the Jacobi matrix is not trace class. Therefore it would be interesting to see how the corresponding asymptotic behavior of $p_n(x)$ deviates from what is stated in (13). This question[3] is answered in Proposition 2.

Finally, since v_n converges weakly to the δ function, which is degenerate, we would like to obtain further information about the zero distribution, for example, the "shape" of the δ function. To this end, let $g_n := f_n(\alpha^{-1/2}x)$ so that all zeros are now in $[-1,1]$. We denote them by $r_{j,n}$, $j \le n$. Let X_n be the random variable so that

$$\text{Prob}(X_n = r_{j,n}) = 1/n \quad \text{for } 1 \le j \le n. \tag{14}$$

In order to get substantial information we must normalize correctly. Thus we define

$$\tilde{X}_n := \frac{X_n}{(\alpha/n)^{1/2}}. \tag{15}$$

One can show that the random variable $\tilde{X}_n$ converges in distribution to the distribution function whose density $p(x)$ is defined below:

$$p(x) = \begin{cases} \dfrac{1}{\pi}\left(\dfrac{4\tan^{-1}(x/\sqrt{4-x^2})}{x^3} - \dfrac{\sqrt{4-x^2}}{x^2}\right), & -2 \le x \le 2, \\ \dfrac{2}{|x|^3}, & |x| \ge 2. \end{cases} \tag{16}$$

It is interesting that the graph of $p(x)$ is volcano-like, i.e., with sloping sides and a crater. The proof of (16) requires that the asymptotics of the orthogonal polynomials is in the form $g_n(\sqrt{\frac{\alpha}{n}}z)$. Since it is a continuation of the present work, we intend to publish it elsewhere.

Now let the normalized measure for the zeros of g_n be $\tilde{v}_n$, i.e.,

$$\begin{aligned} &\tilde{v}_n(\{r_{j,n}\}) = 1/n, \qquad j = 1,2,\ldots,n, \\ &\tilde{v}_n(A) = 0, \qquad A \text{ contains no zeros of } g_n(x). \end{aligned} \tag{17}$$

The following proposition is easily proved.

PROPOSITION 1. *Let x in a compact set $K \subseteq \mathbf{C}\backslash[-1,1]$; then*

$$S(\tilde{v}_n : x) = \frac{1}{n}\frac{g_n'(x)}{g_n(x)} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{x - r_{j,n}}, \tag{18}$$

[3] This question was suggested by a referee.

where $r_{j,n}$'s are the zeros of $g_n(x)$.

Note. This proposition allows us to focus on the asymptotics of $g_n(x)$ for $x \in K \subseteq \mathbf{C}\backslash[-1,1]$.

By Cauchy's theorem we have, for fixed nonzero x,

$$f_n(x) = \frac{1}{2\pi i}\oint_{|w|=\rho} \exp\left\{\frac{w}{x} + \frac{1-\alpha x^2}{x^2}\log(1-wx)\right\} w^{-n-1}dw, \quad \text{where } \rho < |1/x|$$

$$= \frac{1}{2\pi i}\oint_{|w|=|x\rho|} \exp\left\{\frac{w}{x^2} + \frac{1-\alpha x^2}{x^2}\log(1-w)\right\} (w/x)^{-n-1}\frac{dw}{x}, \quad \text{where } |x\rho| < 1.$$

Thus

$$\frac{f_n(x)}{x^n} = \frac{1}{2\pi i}\oint_C \exp\left\{\frac{w}{x^2} + \frac{1-\alpha x^2}{x^2}\log(1-w)\right\} w^{-n-1}dw. \tag{19}$$

Equation (19) holds for all $x \neq 0$, and the integration contour C is any simple closed contour in the open unit disk encircling the origin.

Recall

$$g_n(x) := f_n(\alpha^{-1/2}x). \tag{20}$$

Now all zeros of $g_n(x)$ are in $[-1,1]$; so (19) becomes

$$\frac{g_n(x)}{x^n} = \frac{1}{2\pi i}\oint_C \exp\alpha\left\{\frac{w}{x^2} + \frac{1-x^2}{x^2}\log(1-w)\right\} w^{-n-1}dw. \tag{21}$$

Equation (21) holds for all $x \neq 0$.

The asymptotics of (21) are fairly simple:

$$\frac{g_n(x)}{x^n} = [w^n]\left(\exp\left(\frac{\alpha w}{x^2}\right)(1-w)^{\alpha(1-x^2)/x^2}\right). \tag{22}$$

We emphasize that although one can employ Darboux's method to get the asymptotics of (22) for each *fixed* x (see [16, p. 116]), uniformity is, unfortunately, an issue here, and the Darboux approach does not guarantee uniformity.

Rather, we tackle the problem by using an elementary approach.

PROPOSITION 2.

$$\frac{g_n(x)}{x^n n^{-\alpha(1-x^2)/x^2-1}} \to \frac{e^{\alpha/x^2}}{\Gamma(-\alpha(1-x^2)/x^2)} \quad \text{as } n \to \infty \tag{23}$$

uniformly for all x in a compact set $K \subseteq \mathbf{C}\backslash[-1,1]$.

Proof. To simplify notation we introduce

$$A(x) := \frac{\alpha(1-x^2)}{x^2} \quad \text{and} \quad B(x) := \frac{\alpha}{x^2}. \tag{24}$$

Write the Taylor expansion

$$(1-w)^{A(x)} = \sum_{n=0}^{\infty} a_n w^n, \tag{25}$$

$$(26)\qquad e^{wB(x)} = \sum_{n=0}^{\infty} b_n w^n, \quad \text{where } a_n = (-1)^n \binom{A(x)}{n} \quad \text{and} \quad b_n = \frac{B^n(x)}{n!}.$$

We form the Cauchy product

$$(27)\qquad [w^n]\left(\exp\left(\frac{\alpha w}{x^2}\right)(1-w)^{\alpha(1-x^2)/x^2}\right) = \sum_{j=0}^{n} b_j a_{n-j}.$$

Hence

$$(28)\qquad \frac{g_n(x)}{x^n n^{-A(x)-1}} = \sum_{j=0}^{n} \frac{B^j(x)}{j!} \frac{(-1)^{n-j}}{n^{-A(x)-1}} \binom{A(x)}{n-j}.$$

Using

$$(-1)^{n-j}\binom{A(x)}{n-j} = \frac{\Gamma(n-j-A(x))}{\Gamma(-A(x))\Gamma(n-j+1)},$$

we have

$$(29)\qquad \frac{g_n(x)}{x^n n^{-A(x)-1}} = \frac{1}{\Gamma(-A(x))} \sum_{j=0}^{n} \frac{B^j(x)}{j!} \frac{1}{n^{-A(x)-1}} \frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)}.$$

Write

$$(30)\qquad \frac{e^{B(x)}}{\Gamma(-A(x))} = \frac{1}{\Gamma(-A(x))} \sum_{j=0}^{\infty} \frac{B^j(x)}{j!}.$$

Combining (29) and (30) gives

(31)

$$\frac{g_n(x)}{x^n n^{-A(x)-1}} - \frac{e^{B(x)}}{\Gamma(-A(x))} = \frac{1}{\Gamma(-A(x))} \sum_{j=0}^{n} \frac{B^j(x)}{j!} \left(\frac{1}{n^{-A(x)-1}} \frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)} - 1\right) - \frac{1}{\Gamma(-A(x))} \sum_{j=n+1}^{\infty} \frac{B^j(x)}{j!}.$$

We now decompose the above sum as follows:

$$(32)\qquad \frac{g_n(x)}{x^n n^{-A(x)-1}} - \frac{e^{B(x)}}{\Gamma(-A(x))} = S_1 + S_2 + S_3,$$

where

$$S_1 = \frac{1}{\Gamma(-A(x))} \sum_{j=0}^{n/\log\log n} \frac{B^j(x)}{j!} \left(\frac{1}{n^{-A(x)-1}} \frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)} - 1\right),$$

$$S_2 = \frac{1}{\Gamma(-A(x))} \sum_{j=n/\log\log n}^{n} \frac{B^j(x)}{j!} \left(\frac{1}{n^{-A(x)-1}} \frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)} - 1\right),$$

$$S_3 = \frac{-1}{\Gamma(-A(x))} \sum_{j=n+1}^{\infty} \frac{B^j(x)}{j!}. \tag{33}$$

In order to estimate S_i, $i = 1, 2, 3$, we must interrupt our proof for four short lemmas.

LEMMA 1. *Let $x \in K \subseteq \mathbf{C}\backslash[-1,1]$, K compact. Then $A(x)$ is never ≥ 0, and there exists $M > 0$ such that for all $x \in K$ we have $|A(x)| \leq M$.*

Proof. Solving $\alpha(1-x^2)/x^2 = \beta$ for x we have $x = \pm\sqrt{\alpha/(\alpha+\beta)}$. If $\beta \geq 0$, then x would be real and $|x| \leq 1$. This is a contradiction. The second part of the statement is trivial. □

LEMMA 2. *There exists $M > 0$ such that for all $x \in K$ we have $|1/\Gamma(-A(x))| \leq M$, and $|B(x)| \leq M$.*

Proof. The proof is trivial. □

LEMMA 3. *Let j be $\leq n/\log\log n$, and $x \in K \subseteq \mathbf{C}\backslash[-1,1]$. Then as $n \to \infty$ we have*

$$\left(\frac{1}{n^{-A(x)-1}} \frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)} - 1\right) = O\left(\frac{1}{\log\log n}\right), \tag{34}$$

where the implicit constant holds uniformly.

Proof. By Lemma 1 there exists $M_1 > 0$ (depending on K) such that for all $x \in K \subseteq \mathbf{C}\backslash[-1,1]$ we have

$$|\mathrm{Im}(n-j-A(x))| \leq M_1. \tag{35}$$

The assumption of the lemma implies $|\mathrm{Re}(n-j-A(x))| \to \infty$, as $n \to \infty$. Hence there exists a $\delta > 0$ (depending on K) such that for all $x \in K$ $|\mathrm{Arg}(n-j-A(x))| \leq \pi/2 - \delta$, provided n is large. Using the asymptotics of $\log\Gamma(z)$ (e.g., [14]), which hold uniformly for all z such that $|\mathrm{Arg}\, z| \leq \pi/2 - \delta$, we have

$$\begin{aligned} &\left(\frac{1}{n^{-A(x)-1}} \frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)}\right) \\ &= \left(1-\frac{j}{n}\right)^{-A(x)-1} \frac{\left(1-\frac{A(x)}{n-j}\right)^{-A(x)-1/2}}{\left(1+\frac{1}{n-j}\right)^{1/2}} \left(1-\frac{A(x)+1}{n-j+1}\right)^{n-j} \\ &\quad e^{A(x)+1}\left(!+O\left(\frac{1}{n-j}\right)\right). \end{aligned} \tag{36}$$

By Lemma 1, $|A(x)|$ is bounded. Hence we may estimate each of the above factors as follows:

$$\left(1-\frac{j}{n}\right)^{-A(x)-1} = 1 + O\left(\frac{j}{n}\right) = 1 + O\left(\frac{1}{\log\log n}\right), \tag{37}$$

$$\left(1-\frac{A(x)}{n-j}\right)^{-A(x)-1/2} = 1 + O\left(\frac{|A(x)|}{n-j}\right), \tag{38}$$

$$\left(1-\frac{A(x)+1}{n-j+1}\right)^{n-j} = e^{-A(x)+1}\left(1+O\left(\frac{1}{n-j}\right)\right). \tag{39}$$

The implicit constants in (36)–(39) hold uniformly for $x \in K$ and for j such that $j \le n/\log\log n$.

Putting (37), (38), and (39) into (36) we get

$$\frac{1}{n^{-A(x)-1}}\,\frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)} - 1 = \left(1+O\left(\frac{1}{\log\log n}\right)\right)\left(1+O\left(\frac{1}{n-j}\right)\right) - 1$$

$$= O\left(\frac{1}{\log\log n}\right). \qquad \square$$

LEMMA 4. *Let $n/\log\log n < j \le n$ and $x \in K \subseteq \mathbf{C}\backslash[-1,1]$. There exists a constant c depending only on K such that*

$$\left|\frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)}\right| \le c^n. \tag{40}$$

Proof. Let $n-j=l$. Thus l is in the interval $[0, n(1-(1/\log\log n))]$.

Now

$$\begin{aligned}\left|\frac{\Gamma(l-A(x))}{\Gamma(l+1)}\right| &= |\Gamma(-A(x))|\left|\binom{A(x)}{l}\right| \\ &= |\Gamma(-A(x))|\left|\frac{A(x)(1-A(x))\cdots(l-1-A(x))}{l!}\right| \\ &= |\Gamma(-A(x))||A(x)|\left|1-\frac{A(x)}{1}\right|\left|1-\frac{A(x)}{2}\right|\cdots\left|1-\frac{A(x)}{l-1}\right|\frac{1}{l}.\end{aligned} \tag{41}$$

By Lemma 1, $A(x)$ is never nonnegative. Hence

$$|\Gamma(-A(x))| \le M \quad \text{(bounded)}.$$

Since $|A(x)|$ is bounded, there is an n_0 depending on K such that for all $j \ge n_0$ we have $|A(x)/j| \le \frac{1}{2}$, and as a consequence

$$\left|\frac{\Gamma(l-A(x))}{\Gamma(l+1)}\right| = M^{n_0}\left(\frac{3}{2}\right)^{l-n_0}. \tag{42}$$

Thus there exists a constant c such that

$$\left|\frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)}\right| \le c^n. \qquad \square$$

With Lemmas (1)–(4) at our disposal we are prepared to estimate S_1, S_2, and S_3 in (32). Thus

$$|S_1| \le \left|\frac{1}{\Gamma(-A(x))}\right| \sum_{j=0}^{n/\log\log n}\left|\frac{B^j(x)}{j!}\left(\frac{1}{n^{-A(x)-1}}\,\frac{\Gamma(n-j-A(x))}{\Gamma(n-j+1)}-1\right)\right|. \tag{43}$$

Using Lemma 3 we have

$$|S_1| \le MO\left(\frac{1}{\log\log n}\right)\sum_{j=0}^{n/\log\log n}\frac{|B^j(x)|}{j!}. \tag{44}$$

The converging series $\sum_{j=0}^{n/\log\log n} |B^j(x)|/j!$ is obviously uniformly bounded. Hence $S_1 \to 0$ uniformly for $x \in K$ as $n \to \infty$.

Next, by Lemma 4 we see that

$$|S_2| \le \left|\frac{1}{\Gamma(-A(x))}\right| \sum_{j=n/\log\log n}^{n} \frac{|B^j(x)|}{j!}\left(\frac{c^n}{n^{-|A(x)|-1}}+1\right). \tag{45}$$

Now

$$j! = \Gamma(j+1) \ge \Gamma\left(\frac{n}{\log\log n}+1\right) \ge K_1 \exp\left(\frac{n\log n}{\log\log n} - \frac{n}{\log\log n}\right). \tag{46}$$

Using (46) in (45) we get

$$|S_2| \le M\left(n - \frac{n}{\log\log n}\right) \frac{|B(x)|^n n^{1+|A(x)|} c^n}{K_1 \exp\left(\frac{n\log n}{\log\log n} - \frac{n}{\log\log n}\right)}. \tag{47}$$

Since $\exp(n\log n/\log\log n)$ is dominant, we have $S_2 \to 0$ uniformly for $x \in K$ as $n \to \infty$. Finally, we observe that the tail of a uniformly converging series $\sum_{j=n+1}^{\infty}(B^j(x)/j!) \to 0$ uniformly for $x \in K$ as $n \to \infty$. Thus $S_3 \to 0$ uniformly. □

PROPOSITION 3.

$$\frac{g_n'(x)}{g_n(x)} - \left(\frac{n}{x} + \frac{2\alpha\log n}{x^3}\right) \to \frac{-2\alpha}{x^3}\left(1 + \frac{\Gamma'(-\alpha(1-x^2)/x^2)}{\Gamma(-\alpha(1-x^2)/x^2)}\right), \tag{48}$$

uniformly for $x \in K \subseteq \mathbf{C}\backslash[-1,1]$.

Proof. Since the convergence in (23) is uniform and both functions

$$\frac{g_n(x)}{x^n} n^{-\alpha(1-x^2)/x^2-1} \quad \text{and} \quad \frac{e^{\alpha/x^2}}{\Gamma(-\alpha(1-x^2)/x^2)}$$

are analytic on $\mathbf{C}\backslash[-1,1]$, by a classical theorem of complex analysis on uniform convergence we can differentiate both sides of (23) with respect to x, still maintaining the uniform convergence. The result follows after a little simplification. □

COROLLARY 1. $\tilde{v}_n \to \delta$ *weakly as* $n \to \infty$.

Proof. Propositions 1 and 3 imply that

$$S(v_n; x) \to \frac{1}{x} \quad \text{uniformly for } x \in K \subseteq \mathbf{C}\backslash[-1,1]. \tag{49}$$

By a theorem of Grommer and Hamburger (see, e.g., [12] and [15]) we conclude that $\tilde{v}_n$ converges weakly to $\delta(x)$, the δ-function at $x = 0$. □

Of course, Corollary 1 implies that for any continuous function $h(x)$ on $[-1,1]$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} h(r_{j,n}) = h(0). \tag{50}$$

To have an idea of the rate of convergence of (50) we have the following.

COROLLARY 2. *Let $h(z)$ be analytic in an open set containing $[-1,1]$. Then*

$$n\left(\frac{1}{n}\sum_{j=1}^{n}h(r_{j,n})-h(0)\right) + \alpha h''(0)\ln n \to \frac{1}{2\pi i}\oint_C \frac{-2\alpha}{x^3}\left(1+\frac{\Gamma'(-\alpha(1-x^2)/x^2)}{\Gamma(-\alpha(1-x^2)/x^2)}\right)h(x)dx, \tag{51}$$

where C is a contour in the open set encircling the segment $[-1,1]$.

Proof. Use Proposition 3 and Cauchy's integral formula. □

The existence of a "residual term," $\alpha h''(0)\ln n$ in the above expression (see Theorems 1.11 and 1.14 in [12]), is rather interesting.

REFERENCES

[1] R. ASKEY AND M. E. H. ISMAIL, *Recurrence relations, continued fractions, and orthogonal polynomials*, Mem. Amer. Math. Soc., 300 (1984), pp. 47–68.

[2] L. CARLITZ, *On some polynomials of Tricomi*, Boll. Un. Mat. Ital., 13 (1958), pp. 58–64.

[3] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[4] T. S. CHIHARA AND M. E. H. ISMAIL, *Orthogonal polynomials suggested by a queuing model*, Adv. Appl. Math., 3 (1982), pp. 441–462.

[5] P. ERDÖS AND P. TURAN, *On interpolation*, III, Ann. Math., 41 (1940), pp. 510–555.

[6] P. NEVAI, *Orthogonal polynomials*, Mem. Amer. Math. Soc., 213 (1979), pp. 49–56.

[7] P. NEVAI AND J. S. DEHESA, *On asymptotic average properties of zeros of orthogonal polynomials*, SIAM J. Math. Anal., 10 (1979), pp. 1184–1192.

[8] H. M. SCHWARZ, *A class of continued fractions*, Duke Math. J., 6 (1940), pp. 48–65.

[9] H. STAHL AND V. TOTIK, *General Orthogonal Polynomials*, Cambridge University Press, Cambridge, UK, 1992.

[10] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, 4th ed., Providence, RI, 1975.

[11] F. TRICOMI, *Differential Equations*, Blackie and Son, Limited, Torino, 1961.

[12] W. VAN ASSCHE, *Asymptotics for Orthogonal Polynomials*, Lecture Notes in Math. 1265, Springer-Verlag, New York, 1980.

[13] ———, *Asymptotics for orthogonal polynomials and three-term recurrences*, in Orthogonal Polynomials: Theory and Practice, NATO ASI C 294, P. Nevai, ed., Kluwer, Dordrecht, Germany, 1990, pp. 435–462.

[14] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, UK, 1952.

[15] A. WINTNER, *Spektraltheorie der Unendlichen Matrizen*, Hirzel, Leipzig, Germany, 1929.

[16] R. WONG, *Asymptotic Approximations of Integrals*, Academic Press, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 429–440, March 1994

012

WATSON'S BASIC ANALOGUE OF RAMANUJAN'S ENTRY 40 AND ITS GENERALIZATION*

DHARMA P. GUPTA† AND DAVID R. MASSON†

Abstract. The authors generalize Watson's q-analogue of Ramanujan's Entry 40 continued fraction by deriving solutions to a ${}_{10}\phi_9$-series contiguous relation and applying Pincherle's theorem. Watson's result is recovered as a special terminating case, while a limit case yields a new continued fraction associated with an ${}_8\phi_7$-series contiguous relation.

Key words. contiguous relations, continued fractions, Pincherle's theorem, basic hypergeometric series

AMS subject classifications. 33D15, 39A10, 40A15

1. Introduction.

Contiguous relations for hypergeometric functions are an important source for obtaining explicit results for difference equations, continued fractions, Jacobi matrices and their corresponding orthogonal polynomials. At the top of the Askey–Wilson chart of classical orthogonal polynomials [1] one has the ${}_4F_3$ Wilson polynomials. However, the ${}_4F_3$ label is misleading since the properties of these polynomials and their associated case are revealed by two contiguous relations for very well poised ${}_7F_6$ series [8], [14]. These, in turn, can be derived as limits of a contiguous relation for a terminating, very well poised, two-balanced ${}_9F_8$-series [8], [18]. This ${}_9F_8$ contiguous relation is thus fundamental for the classical hypergeometric polynomials. In a previous publication [15] it was shown how this ${}_9F_8$ contiguous relation was also related to Ramanujan's famous Entry 40 continued fraction [3], [16].

All of the above are $q \to 1$ limits of basic hypergeometric analogues. Thus the ${}_4\phi_3$ Askey-Wilson polynomials should be viewed in the light of very well poised ${}_8\phi_7$ series [9] which are limits of terminating, very well poised, balanced ${}_{10}\phi_9$'s. The analogous contiguous relation for ${}_{10}\phi_9$'s is thus fundamental to the whole scheme of classical and basic hypergeometric orthogonal polynomials. In this paper we derive this important contiguous relation and a corresponding continued fraction. A special terminating version of this continued fraction yields the following result of Watson [17], which is the q-analogue of Ramanujan's Entry 40 [3], [16].

THEOREM A (Watson [17]). *Denoting the base by q^2 (instead of more usual q), let*

$$\frac{1}{G(x)} = \prod_{m=0}^{\infty} (1 - xq^{2m+1}), \qquad (|q| < 1),$$

$$\mathcal{P} = G(\alpha\beta\gamma\delta\epsilon)G\Big(\frac{\alpha\beta\gamma}{\delta\epsilon}\Big)G\Big(\frac{\alpha\beta\delta}{\gamma\epsilon}\Big)G\Big(\frac{\alpha\gamma\delta}{\beta\epsilon}\Big)G\Big(\frac{\alpha\beta\epsilon}{\gamma\delta}\Big)G\Big(\frac{\alpha\gamma\epsilon}{\beta\delta}\Big)G\Big(\frac{\alpha\delta\epsilon}{\beta\gamma}\Big)G\Big(\frac{\alpha}{\beta\gamma\delta\epsilon}\Big)$$

$$\mathcal{Q} = G\Big(\frac{\alpha\beta\gamma\delta}{\epsilon}\Big)G\Big(\frac{\alpha\beta\gamma\epsilon}{\delta}\Big)G\Big(\frac{\alpha\beta\delta\epsilon}{\gamma}\Big)G\Big(\frac{\alpha\gamma\delta\epsilon}{\beta}\Big)G\Big(\frac{\alpha\beta}{\gamma\delta\epsilon}\Big)G\Big(\frac{\alpha\gamma}{\beta\delta\epsilon}\Big)G\Big(\frac{\alpha\delta}{\beta\gamma\epsilon}\Big)G\Big(\frac{\alpha\epsilon}{\beta\gamma\delta}\Big).$$

* Received by the editors January 11, 1993; accepted for publication February 9, 1993.

† Department of Mathematics, University of Toronto, Toronto, Canada, M5S 1A1. The work of the second author was partially supported by the Natural Science and Engineering Research Council of Canada.

Then, provided that one of the numbers $\beta, \gamma, \delta, \epsilon$ is of the form $q^{\pm n} (n = 1, 2, \dots)$,

$$\frac{\mathcal{P} - \mathcal{Q}}{\mathcal{P} + \mathcal{Q}} = \frac{A_0}{\beta_0} \frac{\alpha_1}{+\beta_1} \frac{\alpha_2}{+\beta_2} + \cdots,$$

where

$$\begin{aligned}
A_0 &= (q + q^{-1})\Pi(\alpha - \alpha^{-1}) \\
\alpha_m &= (q^{m+1} + q^{-m-1})(q^{m-1} + q^{1-m})\Pi(\alpha^2 + \alpha^{-2} - q^{2m} - q^{-2m}), \\
\beta_m &= (q^{2m+1} - q^{-2m-1}) \\
&\quad \cdot \Big\{ (q^m + q^{-m})(q^{m+1} + q^{-m-1})(\Sigma\alpha^2 + \Sigma\alpha^{-2} + 2) - \Pi(\alpha + \alpha^{-1}) \\
&\qquad - (q + q^{-1})(q^m + q^{-m})(q^{m+1} + q^{-m-1})(q^{2m+1} + q^{-2m-1}) \Big\}
\end{aligned}$$

with the products and sums ranging over the numbers $\alpha, \beta, \gamma, \delta, \epsilon$.

A second special terminating version of the continued fraction obtained here will give the basic analogue of Masson's Proposition 1 in [15], which is described as a "missing companion" of Ramanujan's Entry 40. For the sake of completeness we state Masson's result.

THEOREM B (Masson [15]). *Let $P' = \Pi\Gamma\big((3 + \alpha \pm \beta \pm \gamma \pm \delta \pm \epsilon)/4\big)$ (0, 2 or 4 minus signs) and $Q' = \Pi\Gamma\big((1 + \alpha \pm \beta \pm \gamma \pm \delta \pm \epsilon)/4\big)$ (1 or 3 minus signs). Then if one of the parameters $\beta, \gamma, \delta, \epsilon$ is an odd integer,*

$$\frac{Q'}{P'} = \frac{-1}{a_0} \frac{2b_1}{-a_1} \frac{b_2}{-a_2} \frac{b_3}{-a_3} - \cdots,$$

where

$$\begin{aligned}
b_n &= \Big(\Pi\big((2n-1)^2 - \alpha^2\big)\Big) \Big/ (16)^3 (2n-1)^2, \\
a_n &= \Big\{2n^6 + n^4(5 - \Sigma\alpha^2)/4 + n^2\big(-26 + (1 + \Sigma\alpha^2)^2 - 2\Sigma\alpha^4\big)/64 - a_0\Big\} \Big/ (4n^2 - 1), \\
a_0 &= \big\{2(1 - \Sigma\alpha^4) + (1 - \Sigma\alpha^2)^2 - 8\Pi\alpha\big\} \Big/ (16)^2,
\end{aligned}$$

with these products and sums ranging over the parameters $\alpha, \beta, \gamma, \delta, \epsilon$.

Masson [15] also gave the nonterminating versions of Ramanujan's Entry 40 and Theorem B.

The object of the present study is to obtain the nonterminating versions of Watson's theorem and the q-analogue of Masson's theorem. They are given in §4 by Corollaries 7 and 8, respectively. Our approach is similar to that in several recent papers [5], [6], [12], [13] on the subject where Pincherle's theorem [11] has been used to bring out the connection between several of Ramanujan's Chapter 12 entries and the general theory of hypergeometric orthogonal functions (Askey and Wilson [1], Wilson [18]). For other approaches to explaining some of Ramanujan's continued fraction entries see [3], [10], [19].

2. Contiguous relation. We consider a terminating, very well poised, balanced

${}_{10}\phi_9$ basic hypergeometric function:

$$\begin{aligned}\phi =&\phi(a;b,c,d,e,f,g,h)\\ :=&{}_{10}\phi_9\left(\begin{matrix}a, q\sqrt{a}, -q\sqrt{a}, b, c, d, e, f, g, h;\ q, q\\ \sqrt{a}, -\sqrt{a}, \frac{aq}{b}, \frac{aq}{c}, \frac{aq}{d}, \frac{aq}{e}, \frac{aq}{f}, \frac{aq}{g}, \frac{aq}{h}\end{matrix}\right),\\ a^3q^2 =&bcdefgh\,, \qquad |q|<1,\end{aligned} \tag{2.1}$$

with, say, $h = q^{-n}$, $n = 0, 1, \ldots$, $g = sq^{n-1}$, $s := a^3q^3/(bcdef)$. We follow the usual notation for variations of ϕ with respect to the parameters. For example $\phi(b+, c-)$ represents the ϕ with b and c replaced by bq and c/q respectively. ϕ_+ denotes the ${}_{10}\phi_9$ obtained by replacing a by aq^2 and b, c, d, e, f, g, h by $bq, cq, dq, eq, fq, gq, hq$, respectively.

We need a contiguous relation basic analogue to the contiguous relation derived by Wilson [18] for the ${}_9F_8$ hypergeometric function. To work out this contiguous relation, we shall use Wilson's method [18] of using the basic hypergeometric analogues of the relevant formulas.

LEMMA 1. *Let ϕ be given by* (2.1) *(not necessarily terminating). Then*

$$\begin{aligned}&\phi(b-,c+)-\phi\\ &=\frac{\frac{aq}{c}(1-\frac{cq}{b})(1-\frac{bc}{aq})(1-aq)(1-aq^2)(1-d)(1-e)(1-f)(1-g)(1-h)}{(1-\frac{aq}{b})(1-\frac{aq^2}{b})(1-\frac{a}{c})(1-\frac{aq}{c})(1-\frac{aq}{d})(1-\frac{aq}{e})(1-\frac{aq}{f})(1-\frac{aq}{g})(1-\frac{aq}{h})}\\ &\quad\times\phi_+(b-)\,.\end{aligned} \tag{2.2}$$

Proof. A straightforward term by term subtraction on the left side of (2.2) leads to the result. □

LEMMA 2. *If ϕ (given by* (2.1)*) is terminating, then*

$$\begin{aligned}&\frac{b^2(1-h)(1-\frac{aq}{bc})(1-\frac{aq}{bd})(1-\frac{aq}{be})(1-\frac{aq}{bf})(1-\frac{aq}{bg})}{(1-\frac{aq}{b})(1-\frac{aq^2}{b})}\phi_+(b-)\\ &-\frac{h^2(1-b)(1-\frac{aq}{ch})(1-\frac{aq}{dh})(1-\frac{aq}{eh})(1-\frac{aq}{fh})(1-\frac{aq}{gh})}{(1-\frac{aq}{h})(1-\frac{aq^2}{h})}\phi_+(h-)\\ &-\frac{b(1-\frac{h}{b})(1-\frac{aq}{c})(1-\frac{aq}{d})(1-\frac{aq}{e})(1-\frac{aq}{f})(1-\frac{aq}{g})}{(1-aq)(1-aq^2)}\phi=0.\end{aligned} \tag{2.3}$$

Proof. By eliminating $\phi_+(b-)$ from (2.2) and another similar relation written for $\phi(b-, d+) - \phi$ we obtain

$$\begin{aligned}&c(1-c)\left(1-\frac{a}{c}\right)\left(1-\frac{dq}{b}\right)\left(1-\frac{bd}{aq}\right)\phi(b-,c+)\\ &-d(1-d)\left(1-\frac{a}{d}\right)\left(1-\frac{cq}{b}\right)\left(1-\frac{bc}{aq}\right)\phi(b-,d+)\\ &+d\left(1-\frac{b}{q}\right)\left(1-\frac{c}{d}\right)\left(1-\frac{aq}{b}\right)\left(1-\frac{cd}{a}\right)\phi=0.\end{aligned} \tag{2.4}$$

With, say, $h = q^{-n}$, we can apply an iterate of Bailey's transformation 8.5(1) [2, p. 68] to ϕ, $\phi_+(b-)$ and $\phi_+(h-)$ (the transformation [4, ex. 2.19, p. 53] with b, e, g replaced by g, b, e, respectively). The three transformed series are related by (2.4). Reversing the transformations in this relation we arrive at (2.3). □

THEOREM 3. *If ϕ (given by (2.1)) is terminating, then*

$$
\begin{aligned}
&\frac{g(1-h)(1-\frac{a}{h})(1-\frac{aq}{h})(1-\frac{aq}{gb})(1-\frac{aq}{gc})(1-\frac{aq}{gd})(1-\frac{aq}{ge})(1-\frac{aq}{gf})}{(1-\frac{hq}{g})} \\
&\qquad \times [\phi(g-,h+)-\phi] \\
&-\frac{h(1-g)(1-\frac{a}{g})(1-\frac{aq}{g})(1-\frac{aq}{hb})(1-\frac{aq}{hc})(1-\frac{aq}{hd})(1-\frac{aq}{he})(1-\frac{aq}{hf})}{(1-\frac{gq}{h})} \\
&\qquad \times [\phi(h-,g+)-\phi] \\
&-\frac{aq}{h}\left(1-\frac{h}{g}\right)\left(1-\frac{gh}{aq}\right)(1-b)(1-c)(1-d)(1-e)(1-f)\phi = 0.
\end{aligned} \tag{2.5}
$$

Proof. We eliminate $\phi_+(b-)$ and $\phi_+(c-)$ from (2.2), (2.2) with $b \leftrightarrow c$ in (2.2), and with $c \leftrightarrow h$ in (2.3). A final interchange of parameters $b \leftrightarrow g$, $c \leftrightarrow h$ yields the desired result. □

Substituting $h = q^{-n}$, $g = sq^{n-1}$, and renormalizing, the contiguous relation (2.5) becomes the linear second-order difference equation

$$X_{n+1} - a_n X_n + b_n X_{n-1} = 0, \qquad n \geq 0, \tag{2.6}$$

(2.7)

$$
\begin{aligned}
a_n = \Bigg[&\frac{q^{-n+\frac{1}{2}}}{\sqrt{s}}(1-sq^{n-1})(1-\frac{s}{a}q^{n-1})(1-\frac{s}{a}q^{n-2}) \\
&\times \frac{(1-\frac{a}{b}q^{n+1})(1-\frac{a}{c}q^{n+1})(1-\frac{a}{d}q^{n+1})(1-\frac{a}{e}q^{n+1})(1-\frac{a}{f}q^{n+1})}{(1-sq^{2n})} \\
&+ \frac{q^{-n+\frac{3}{2}}}{\sqrt{s}}(1-q^n)(1-aq^n)(1-aq^{n+1}) \\
&\times \frac{(1-\frac{bs}{a}q^{n-2})(1-\frac{cs}{a}q^{n-2})(1-\frac{ds}{a}q^{n-2})(1-\frac{es}{a}q^{n-2})(1-\frac{fs}{a}q^{n-2})}{(1-sq^{2n-2})} \\
&+ \frac{\sqrt{s}}{a}q^{n-\frac{1}{2}}(1-sq^{2n-1})\left(1-\frac{s}{aq^2}\right)(1-b)(1-c)(1-d)(1-e)(1-f)\Bigg] \\
&\Big/\left[(1-sq^{2n-1})(1-\frac{s}{a}q^{n-2})(1-aq^{n+1})\right],
\end{aligned} \tag{2.7}
$$

(2.8)

$$
\begin{aligned}
b_n = &\frac{q^{-2n+3}}{s}(1-q^n)(1-sq^{n-2})\left(1-\frac{a}{b}q^n\right)\left(1-\frac{a}{c}q^n\right)\left(1-\frac{a}{d}q^n\right)\left(1-\frac{a}{e}q^n\right)\left(1-\frac{a}{f}q^n\right) \\
&\times \frac{(1-\frac{bs}{a}q^{n-2})(1-\frac{cs}{a}q^{n-2})(1-\frac{ds}{a}q^{n-2})(1-\frac{es}{a}q^{n-2})(1-\frac{fs}{a}q^{n-2})}{(1-sq^{2n-1})(1-sq^{2n-2})^2(1-sq^{2n-3})},
\end{aligned} \tag{2.8}
$$

$$s = \frac{a^3q^3}{bcdef}$$

with the solution

$$X_n^{(1)} = \frac{q^{-\frac{n^2}{2}+n}}{s^{\frac{n}{2}}} \frac{(sq^{2n-1})_\infty (aq^{n+1})_\infty}{(sq^{n-1})_\infty (\frac{s}{a}q^{n-1})_\infty (\frac{aq^{n+1}}{b}, \frac{aq^{n+1}}{c}, \frac{aq^{n+1}}{d}, \frac{aq^{n+1}}{e}, \frac{aq^{n+1}}{f})_\infty} \times \phi(a; b, c, d, e, f, sq^{n-1}, q^{-n}). \tag{2.9}$$

Here the infinite product $(a)_\infty$ means

$$(a)_\infty = (a; q)_\infty = (1-a)(1-aq)(1-aq^2)\cdots$$

and

$$(a, b, \ldots, k)_\infty = (a)_\infty (b)_\infty \ldots (k)_\infty.$$

For the exceptional values $s = q, q^2$, the a_n, b_n and b_{n+1} in (2.7) and (2.8), and the $X_n^{(1)}, X_{n-1}^{(1)}$ in (2.9) are indeterminate at $n = 0$. We resolve this indeterminancy by taking limits as $n \to 0$.

Next, we proceed to find a second linearly independent solution to the second-order difference equation (2.6). This can be obtained by using a q-analogue of [15]. Thus from (2.6), (2.9), and a symmetry relation ((2.11), which follows) we are able to obtain a second terminating ${}_{10}\phi_9$ solution for the special values $s = q, q^2, \ldots$. For general values of s the second solution will be an appropriate combination of two nonterminating ${}_{10}\phi_9$'s which satisfy a four-term transformation (Gasper and Rahman [4], formula III.39, p. 247). We will consider the case of general s in future work.

Observe that with the replacement

$$(a, b, c, d, e, f, sq^{n-1}, q^{-n}) \to \left(\frac{q}{a}, \frac{q}{b}, \frac{q}{c}, \frac{q}{d}, \frac{q}{e}, \frac{q}{f}, \frac{q^{-n+2}}{s}, q^{n+1}\right) \tag{2.10}$$

we have

$$(a_n, b_n) \to (a_n, b_{n+1}). \tag{2.11}$$

It is easy to check that $b_n \to b_{n+1}$. To check $a_n \to a_n$ we used the "Maple" software on the computer. This meant verifying a polynomial identity in $x = q^{-n}$ of degree 14.

Applying the transformation (2.10) to (2.6) and (2.9) and renormalizing, we obtain the second solution

$$X_n^{(2)} = \frac{q^{-\frac{n^2}{2}+n}}{s^{\frac{n}{2}}} \frac{(\frac{s}{a}q^n)_\infty (sq^{2n-1})_\infty}{(q^{n+1})_\infty (aq^n)_\infty (\frac{bs}{a}q^{n-1}, \frac{cs}{a}q^{n-1}, \frac{ds}{a}q^{n-1}, \frac{es}{a}q^{n-1}, \frac{fs}{a}q^{n-1})_\infty} \times \phi\left(\frac{q}{a}; \frac{q}{b}, \frac{q}{c}, \frac{q}{d}, \frac{q}{e}, \frac{q}{f}, \frac{q^{-n+2}}{s}, q^{n+1}\right), \qquad s = q, q^2, \ldots. \tag{2.12}$$

Note that ϕ is terminating in (2.12) because of the parameter q^{-n+2}/s.

3. Asymptotics and Pincherle's theorem. To obtain a minimal (subdominant) solution for (2.6) we need the large n asymptotics of (2.9) and (2.12). Applying Tannery's theorem to the ${}_{10}\phi_9$'s on the right side of (2.9) and (2.12), we have, as $n \to \infty$,

$$X_n^{(1)} \sim \frac{q^{-\frac{n^2}{2}+n}}{s^{\frac{n}{2}}} W(a; b, c, d, e, f), \qquad \left|\frac{s}{aq}\right| < 1 \tag{3.1}$$

and

$$X_n^{(2)} \sim \frac{q^{-\frac{n^2}{2}+n}}{s^{\frac{n}{2}}} W\left(\frac{q}{a}; \frac{q}{b}, \frac{q}{c}, \frac{q}{d}, \frac{q}{e}, \frac{q}{f}\right), \qquad \left|\frac{aq^2}{s}\right| < 1, \tag{3.2}$$

where

$$W(a; b, c, d, e, f) := {}_8\phi_7\left(\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b, c, d, e, f \\ \sqrt{a}, -\sqrt{a}, \frac{qa}{b}, \frac{qa}{c}, \frac{qa}{d}, \frac{qa}{e}, \frac{qa}{f} \end{matrix}; q, \frac{a^2q^2}{bcdef}\right).$$

We write

$$W_1 := W(a; b, c, d, e, f), \qquad |s| < |qa|$$

and its analytic continuation otherwise, and we write

$$W_2 := W\left(\frac{q}{a}; \frac{q}{b}, \frac{q}{c}, \frac{q}{d}, \frac{q}{e}, \frac{q}{f}\right), \qquad |s| > |aq^2|$$

and its analytic continuation otherwise.

Now taking

$$X_n^{(3)} := W_2 X_n^{(1)} - W_1 X_n^{(2)}, \tag{3.3}$$

it follows from (3.1) and (3.2) that

$$\lim_{n\to\infty} \frac{X_n^{(3)}}{X_n^{(1)}} = 0, \qquad \left|\frac{s}{q}\right| < |a| < \left|\frac{s}{q^2}\right|. \tag{3.4}$$

This establishes that $X_n^{(3)}$ is a minimal solution of (2.6). An application of Pincherle's theorem [11] then leads to the following result.

THEOREM 4. *Let* $s = q, q^2, \ldots$. *Then*

$$\frac{1}{a_0} \; \frac{b_1}{-a_1} \; \frac{b_2}{-a_2} \; -\cdots = \lim_{n\to 0} \frac{W_2 X_n^{(1)} - W_1 X_n^{(2)}}{b_n(W_2 X_{n-1}^{(1)} - W_1 X_{n-1}^{(2)})}. \tag{3.5}$$

Proof. From Pincherle's theorem (3.5) is true for $|s/q| < |a| < |s/q^2|$. For other values of a the result follows by analytic continuation. To the left side of (3.5) we can apply the "parabola theorem" (see Jones and Thron [11, p. 99] and Jacobsen [10]) since from (2.6), $b_n/(a_n a_{n-1}) = (q^3/(1+q)^2)(1+O(q^n))$. Hence the left side of (3.5) is a meromorphic function of a. The right side of (3.5) involves convergent infinite products and ${}_8\phi_7$'s that are each expressible in terms of convergent infinite products and convergent ${}_4\phi_3$'s (Gasper and Rahman [4, eq. (2.10.10), p. 43]). Consequently, the right side of (3.5) is also a meromorphic function of a and (3.5) follows by analytic continuation to all values of a. Note that the exceptional cases $s = q, q^2$ that cause indeterminancy are taken care of by the limit $n \to 0$ on the right side of (3.5). □

For the exceptional values $s = q^2, q$ Theorem 4 gives, respectively, the nonterminating versions of Theorem A (Watson[17]) and the basic analogue of Theorem B (Masson [15]). We now demonstrate how to derive the terminating versions of Theorem 4. We shall need to express the ratio W_1/W_2 in terms of infinite products when $b/a = q^N$, N being an integer. We write

$$\widetilde{W}(a; b, c, d, e, f) := \left(\frac{aq}{b}, \frac{aq}{c}, \frac{aq}{d}, \frac{aq}{e}, \frac{aq}{f}\right)_\infty W(a; b, c, d, e, f)$$

and

$$U(a;b,c,d,e,f) := \frac{\widetilde{W}(a;b,c,d,e,f)}{(aq,b,c,d,e,f)_\infty}.$$

LEMMA 5. *If $b/a = q^N$, where N is an integer, then*

$$U(a;b,c,d,e,f) = \left(\frac{s}{aq}\right)^N U\left(\frac{b^2}{a};b,\frac{bc}{a},\frac{bd}{a},\frac{be}{a},\frac{bf}{a}\right). \tag{3.6}$$

Proof. Refer to Bailey's three-term ${}_8\phi_7$ transformation (Gasper and Rahman [4, formula III.37, p. 246]). If we apply the condition $b/a = q^N$, $N = 0, \pm1, \pm2, \dots$ we obtain the desired result. □

LEMMA 6. *If $s = a^3q^3/(bcdef) = q^M$ and $b/a = q^N$, M and N being integers, then*

$$\begin{aligned}&\frac{\widetilde{W}(a;b,c,d,e,f)}{\widetilde{W}\left(\frac{a}{a};\frac{a}{b},\frac{a}{c},\frac{a}{d},\frac{a}{e},\frac{a}{f}\right)}\\ &\quad = \lambda\frac{(aq,c,d,e,f,\frac{aq^2}{s},\frac{aq}{ef},\frac{aq}{df},\frac{aq}{de})_\infty}{(\frac{bc}{a},\frac{bd}{a},\frac{be}{a},\frac{bf}{a},\frac{q^2}{a},\frac{q}{b},\frac{cd}{a},\frac{ce}{a},\frac{cf}{a})_\infty},\end{aligned} \tag{3.7}$$

where

$$\lambda = (-1)^{n+1}\left(\frac{s}{aq}\right)^N\left(\frac{c}{b}\right)^{n+1}q^{n(n+1)/2} \quad \text{for} \quad \frac{aq^3}{bs} = q^{-n},\ n = 0,1,2,\dots, \tag{3.8}$$

and

$$\lambda = (-1)^{n+1}\left(\frac{s}{aq}\right)^N\left(\frac{b}{c}\right)^{n+1}q^{(n+1)(n+2)/2} \quad \text{for} \quad \frac{bs}{aq} = q^{-n},\ n = -1,0,1,2,\dots. \tag{3.9}$$

Proof. We express the left side of (3.7) in terms of appropriate ${}_4\phi_3$'s. To the numerator W in (3.7) we first apply the identity (3.6) and then the three-term transformation [4, formula III.36, p. 246]. To the denominator W we first apply the ${}_8\phi_7$ transformation [4, formula III.24, p. 243] and then formula III.36 [4]. We also make use of the relation

$$\begin{aligned}\lim_{e\to q^{-n}}(e)_\infty\,{}_4\phi_3\left(\begin{matrix}a,b,c,d\\ e,f,g\end{matrix};q,q\right) &= \frac{(a,b,c,d,fq^{n+1},gq^{n+1},q^{n+2})_\infty}{(aq^{n+1},bq^{n+1},cq^{n+1},dq^{n+1},f,g)_\infty}\\ &\quad\times q^{n+1}\,{}_4\phi_3\left(\begin{matrix}aq^{n+1},bq^{n+1},cq^{n+1},dq^{n+1}\\ q^{n+2},fq^{n+1},gq^{n+1}\end{matrix};q,q\right).\end{aligned} \tag{3.10}$$

All this enables us to recognize and cancel a common linear combination of ${}_4\phi_3$'s from the numerator and the denominator yielding the desired result. We note that in the case $bs/aq = q$, $s = q$, the limit (3.10) is not required and there is an exact cancellation. □

4. Exceptional values s = q, q². We now restate Theorem 4 for the exceptional values $s = q, q^2$ and the form they take when the continued fraction terminates:

COROLLARY 7. *If $s = q^2$, then* (3.5) *can be rewritten as*

$$\frac{1}{a_0}\ -\frac{b_1}{a_1}\ -\frac{b_2}{a_2}\ -\cdots = \frac{2a(1-q)}{q^{3/2}(1-\alpha)(1-\beta)(1-\gamma)(1-\delta)(1-\epsilon)}\left(\frac{1-V}{1+V}\right), \tag{4.1}$$

$$
\begin{aligned}
a_n = &\Big[q^{\frac{1}{2}}\Pi(\alpha^{\frac{1}{2}}+\alpha^{-\frac{1}{2}}) + q^{\frac{1}{2}}(q^{\frac{1}{2}}+q^{-\frac{1}{2}})(q^{n+\frac{1}{2}}+q^{-n-\frac{1}{2}}) \\
&\times (q^{\frac{n}{2}}+q^{-\frac{n}{2}})(q^{\frac{n+1}{2}}+q^{-\frac{n+1}{2}}) \\
&- q^{\frac{1}{2}}(\Sigma\alpha+\Sigma\alpha^{-1}+2)(q^{\frac{n}{2}}+q^{-\frac{n}{2}})(q^{\frac{n+1}{2}}+q^{-\frac{n+1}{2}})\Big] \\
&\Big/(q^{\frac{n}{2}}+q^{-\frac{n}{2}})(q^{\frac{n+1}{2}}+q^{-\frac{n+1}{2}}) \,, \qquad (4.2)\\
b_n = &\frac{-q\Pi(\alpha+\alpha^{-1}-q^n-q^{-n})}{(q^{-\frac{n}{2}}+q^{\frac{n}{2}})^2(q^{-n-\frac{1}{2}}-q^{n+\frac{1}{2}})(q^{-n+\frac{1}{2}}-q^{n-\frac{1}{2}})}, \\
V = &\frac{(\frac{q}{a})_\infty(\frac{q^2}{a})_\infty}{(a)_\infty(aq)_\infty}\,\frac{\widetilde{W}_1}{\widetilde{W}_2}\,, \qquad (4.3)\\
a = &(q\alpha\beta\gamma\delta\epsilon)^{\frac{1}{2}},
\end{aligned}
$$

and product Π *and summation* Σ *are taken over the parameters* $\alpha,\beta,\gamma,\delta,\epsilon$. *If one of the parameters* $\beta,\gamma,\delta,\epsilon = q^N$, N *integer,* $(\alpha,\beta,\gamma,\delta,\epsilon) \to (\alpha^2,\beta^2,\gamma^2,\delta^2,\epsilon^2)$ *and the base* q *is changed to* q^2, *then the right side of* (4.1) *becomes*

$$
\frac{2(q^{-1}-q)}{q\Pi(\alpha-\alpha^{-1})}\,\frac{\mathcal{P}-\mathcal{Q}}{\mathcal{P}+\mathcal{Q}}, \qquad (4.4)
$$

with

$$
\begin{aligned}
\frac{1}{\mathcal{P}} &= \Big(\alpha\beta\gamma\delta\epsilon, \frac{\alpha}{\beta\gamma\delta\epsilon}, \frac{\alpha\beta\gamma}{\delta\epsilon}, \frac{\alpha\beta\delta}{\gamma\epsilon}, \frac{\alpha\gamma\delta}{\beta\epsilon}, \frac{\alpha\beta\epsilon}{\gamma\delta}, \frac{\alpha\gamma\epsilon}{\beta\delta}, \frac{\alpha\delta\epsilon}{\beta\gamma}; q^2\Big)_\infty, \\
\frac{1}{\mathcal{Q}} &= \Big(\frac{\alpha\beta\gamma\delta}{\epsilon}, \frac{\alpha\beta\gamma\epsilon}{\delta}, \frac{\alpha\beta\delta\epsilon}{\gamma}, \frac{\alpha\gamma\delta\epsilon}{\beta}, \frac{\alpha\beta}{\gamma\delta\epsilon}, \frac{\alpha\gamma}{\beta\delta\epsilon}, \frac{\alpha\delta}{\beta\gamma\epsilon}, \frac{\alpha\epsilon}{\beta\gamma\delta}; q^2\Big)_\infty,
\end{aligned}
$$

and a_n, b_n *modified accordingly. This reproduces Theorem* A.

Proof. We write $\alpha = a/b$, $\beta = a/c$, $\gamma = a/d$, $\delta = a/e$, $\epsilon = a/f$. After that the proof is straightforward on substituting the values of $X_0^{(1)}$, $X_0^{(2)}$, $b_0X_{-1}^{(1)}$, $b_0X_{-1}^{(2)}$ from (2.9), (2.12) and (2.8) into Theorem 4. A lot of algebra is involved in the simplification. Also, appropriate limits are to be taken whenever indeterminates occur.

To obtain the terminating form (4.4) we need to use Lemma 6 after interchanging, say, b and f. In both cases, viz.,

$$
\frac{aq^3}{fs} = q^{-n}, \quad n = 0,1,2,\ldots \quad \text{and} \quad \frac{fs}{aq} = q^{-n}, \quad n = -1,0,1,2,\ldots
$$

whether the termination is due to one or the other, the result works out to be the same. The above result (4.4) yields Watson's result [17] i.e., Theorem A in §1. □

COROLLARY 8. *If* $s = q$, *then* (3.5) *can be rewritten as*

$$
\frac{1}{a_0} - \frac{b_1}{a_1} - \frac{b_2}{a_2} - \cdots = 2\left(a_0 + \frac{a^2}{q}\,\frac{(\frac{1}{a})_\infty(\frac{1}{a})_\infty}{(aq)_\infty(\frac{a}{q})_\infty}\,\frac{\widetilde{W}_1}{\widetilde{W}_2}\right)^{-1}, \qquad (4.5)
$$

$$
\begin{aligned}
a_n = {} & \frac{q^{\frac{1}{2}}}{(q^{-n-\frac{1}{2}} - q^{n+\frac{1}{2}})(q^{-n+\frac{1}{2}} - q^{n-\frac{1}{2}})} \\
& \cdot \Big\{ (q^{3n} + q^{-3n})(q^{\frac{1}{2}} + q^{-\frac{1}{2}}) \\
& \quad - (q^{2n} + q^{-2n})\Big[q^{\frac{1}{2}} + q^{-\frac{1}{2}} + \Sigma(\alpha^{-\frac{1}{2}} + \alpha^{\frac{1}{2}})\Big] \\
& \quad + (q^n + q^{-n})\Big[- q^{\frac{3}{2}} - q^{-\frac{3}{2}} + (q^{\frac{1}{4}} + q^{-\frac{1}{4}})\Pi(\alpha^{\frac{1}{4}} + \alpha^{-\frac{1}{4}}) \\
& \qquad + (q^{\frac{1}{4}} - q^{-\frac{1}{4}})\Pi(\alpha^{\frac{1}{4}} - \alpha^{-\frac{1}{4}})\Big] \qquad (4.6) \\
& \quad + (q^{-1} + q)(q^{\frac{1}{2}} + q^{-\frac{1}{2}}) + (q^{-1} + q)\Sigma(\alpha^{-\frac{1}{2}} + \alpha^{\frac{1}{2}}) \\
& \quad - (q^{\frac{1}{4}} + q^{-\frac{1}{4}})(q^{\frac{1}{2}} + q^{-\frac{1}{2}})\Pi(\alpha^{\frac{1}{4}} + \alpha^{-\frac{1}{4}}) \\
& \quad + (q^{\frac{1}{4}} - q^{-\frac{1}{4}})(q^{\frac{1}{2}} + q^{-\frac{1}{2}})\Pi(\alpha^{\frac{1}{4}} - \alpha^{-\frac{1}{4}}) \Big\},
\end{aligned}
$$

$$
b_n = q^{\frac{5}{2}} \frac{\Pi(q^{-n} + q^n - \alpha^{\frac{1}{2}}q^{-\frac{1}{2}} - \alpha^{-\frac{1}{2}}q^{\frac{1}{2}})}{(q^{\frac{n}{2}} + q^{-\frac{n}{2}})(q^{\frac{n}{2}-\frac{1}{2}} + q^{-\frac{n}{2}+\frac{1}{2}})(q^{n-\frac{1}{2}} - q^{-n+\frac{1}{2}})^2},
$$

$$
\widetilde{W}_1 = \widetilde{W}\left(a;\ a\sqrt{\frac{q}{\alpha}},\ a\sqrt{\frac{q}{\beta}},\ a\sqrt{\frac{q}{\gamma}},\ a\sqrt{\frac{q}{\delta}},\ a\sqrt{\frac{q}{\epsilon}}\right), \qquad (4.7)
$$

$$
\widetilde{W}_2 = \widetilde{W}\left(\frac{q}{a};\ \frac{\sqrt{\alpha q}}{a},\ \frac{\sqrt{\beta q}}{a},\ \frac{\sqrt{\gamma q}}{a},\ \frac{\sqrt{\delta q}}{a},\ \frac{\sqrt{\epsilon q}}{a}\right),
$$

$$
a = \left(\frac{\alpha\beta\gamma\delta\epsilon}{q}\right)^{\frac{1}{4}},
$$

and product Π *and summation* Σ *are taken over parameters* $\alpha, \beta, \gamma, \delta, \epsilon$. *If one of the parameters* $\beta, \gamma, \delta, \epsilon$ *is* q^N, N *an odd integer,* $(\alpha, \beta, \gamma, \delta, \epsilon) \to (\alpha^4, \beta^4, \gamma^4, \delta^4, \epsilon^4)$ *and the base* q *is changed to* q^4, *then* (4.5) *becomes*

$$
\frac{1}{a_0}\ \frac{b_1}{-a_1}\ \frac{b_2}{-a_2}\ -\cdots = 2\left(a_0 - \frac{q^2}{\alpha^2}\frac{\mathcal{P}'}{\mathcal{Q}'}\right)^{-1}, \qquad (4.8)
$$

where

$$
\begin{aligned}
\frac{1}{\mathcal{Q}'} &= \left(\frac{q\alpha\beta\gamma\delta}{\epsilon}, \frac{q\alpha\gamma\delta\epsilon}{\beta}, \frac{q\alpha\beta\epsilon\delta}{\gamma}, \frac{q\alpha\epsilon\beta\gamma}{\delta}, \frac{q\alpha\beta}{\epsilon\gamma\delta}, \frac{q\alpha\gamma}{\epsilon\beta\delta}, \frac{q\alpha\delta}{\epsilon\beta\gamma}, \frac{q\alpha\epsilon}{\beta\gamma\delta}; q^4\right)_\infty, \qquad (4.9) \\
\frac{1}{\mathcal{P}'} &= \left(q^3\alpha\beta\gamma\delta\epsilon, \frac{q^3\alpha}{\beta\gamma\delta\epsilon}, \frac{q^3\alpha\delta\epsilon}{\beta\gamma}, \frac{q^3\alpha\gamma\epsilon}{\beta\delta}, \frac{q^3\gamma\delta\alpha}{\epsilon\beta}, \frac{q^3\alpha\beta\delta}{\epsilon\gamma}, \frac{q^3\alpha\beta\gamma}{\epsilon\delta}, \frac{q^3\alpha\beta\epsilon}{\gamma\delta}; q^4\right)_\infty,
\end{aligned}
$$

with a_n *and* b_n *modified accordingly.*

Proof. We write $\alpha = a^2q/b^2$, $\beta = a^2q/c^2$, $\gamma = a^2q/d^2$, $\delta = a^2q/e^2$, and $\epsilon = a^2q/f^2$ and make the appropriate substitutions from (2.9), (2.12), and (2.8) into Theorem 4. A considerable amount of algebra is required to reexpress (2.7) as (4.6) for which we used "Maple" software on the computer. We use Lemma 6 after interchanging, say, b and f to arrive at (4.8). The result is the same for either type of termination (3.8) or (3.9). Note that (4.8) can be reexpressed in the form

$$
-\frac{q^2}{\alpha^2}\frac{\mathcal{Q}'}{\mathcal{P}'} = \frac{1}{a_0}\ \frac{2b_1}{-a_1}\ \frac{b_2}{-a_2}\ -\cdots
$$

which is a q-analogue of Theorem B in §1. □

5. Ordinary cases $s = q^3, q^4, \ldots$. By substituting $s = q^m, m$ integer greater than 2, into (3.5) of Theorem 4 we obtain the following corollary.

COROLLARY 9. *For $s = q^m$, $m = 3, 4, \ldots$*

$$\frac{1}{a_0} \; \frac{b_1}{-a_1} \; \frac{b_2}{-a_2} \; -\cdots = \frac{q^{(m-3)/2}(1-q^{m-1})(1-\frac{a}{q})}{(1-\frac{q}{a}^{m-1})(1-\frac{a}{b})(1-\frac{a}{c})(1-\frac{a}{d})(1-\frac{a}{e})(1-\frac{a}{f})} \tag{5.1}$$

$$\times \left[\phi\Big(\frac{q}{a};\frac{q}{b},\frac{q}{c},\frac{q}{d},\frac{q}{e},\frac{q}{f},q^{-m+2},q\Big) - \frac{\widetilde{W}_2}{\widetilde{W}_1}\,\frac{(a)_\infty(aq)_\infty(q)_\infty}{(\frac{q^m}{a})_\infty(\frac{q^{m-1}}{a})_\infty(q^m)_\infty}\,\frac{1}{(1-q^{m-1})}\right.$$

$$\left.\times\,\frac{(\frac{b}{a}q^{m-1},\ \frac{c}{a}q^{m-1},\ \frac{d}{a}q^{m-1},\ \frac{e}{a}q^{m-1},\ \frac{f}{a}q^{m-1})_\infty}{(\frac{bq}{a},\ \frac{cq}{a},\ \frac{dq}{a},\ \frac{eq}{a},\ \frac{fq}{a})_\infty}\right],$$

where a_n, b_n are given by (2.7) *and* (2.8) *for $s = q^m$, $m \geq 3$.*

Note that if we substitute $s = q (m = 1)$ into (5.1) the right side reduces to

$$\frac{q}{a^2}\,\frac{\widetilde{W}_2}{\widetilde{W}_1}\,\frac{(\frac{a}{q})_\infty(aq)_\infty}{(\frac{1}{a})_\infty(\frac{1}{a})_\infty} \tag{5.2}$$

and (5.1) agrees with (4.5) when the indeterminacy in a_0 and b_1 is taken into account, that is,

$$\begin{aligned} \lim_{s\to q} a_0(s) &= \lim_{n\to 0} a_n(s=q), \\ \lim_{s\to q} b_1(s) &= 2\lim_{n\to 1} b_n(s=q). \end{aligned} \tag{5.3}$$

It is the a_0 and b_1 on the left side of (5.3) that should occur in (5.1) with $s = q$, whereas it is $\lim_{n\to 0} a_n(s=q)$ and $\lim_{n\to 1} b_n(s=q)$ that are the a_0 and b_1 that occur in (4.5).

Similarly, for $s = q^2$ the right side of (5.1) becomes

$$-\frac{a}{q^{\frac{3}{2}}}\,\frac{(1-q)}{(1-\frac{a}{b})(1-\frac{a}{c})(1-\frac{a}{d})(1-\frac{a}{e})(1-\frac{a}{f})}\left[1-\frac{(a)_\infty(aq)_\infty}{(\frac{q^2}{a})_\infty(\frac{q}{a})_\infty}\,\frac{\widetilde{W}_2}{\widetilde{W}_1}\right]. \tag{5.4}$$

This agrees with (4.1) since

$$\lim_{s\to q^2} a_0(s) = \lim_{n\to 0} a_n(s=q^2) + \frac{aq^{\frac{1}{2}}(1-\frac{b}{a})(1-\frac{c}{a})(1-\frac{d}{a})(1-\frac{e}{a})(1-\frac{f}{a})}{2(1-q)},$$

$$\lim_{s\to q^2} b_1(s) = \lim_{n\to 1} b_n(s=q^2).$$

We can also consider the terminating case of (5.1) by taking into account Lemma 6.

Remark 1. If in §2 we make the replacements $a \to \lambda a$, $b \to \lambda q/b$, $c \to \lambda q/c$, $d \to \lambda q/d$, $e \to ae^{i\theta}$, $f \to ae^{-i\theta}$, and let $\lambda \to \infty$, then we obtain solutions to the recurrence for Askey-Wilson polynomials [1] with $s = abcd = q^m$, $m = 1, 2, \ldots$ [7]. By applying the above limit to Corollaries 7, 8, and 9, we recover equations (22), (23), and (24), respectively, of Gupta and Masson [7]. Note that [7, eqs. (22) and (23)] give the q-analogue of Ramanujan's Entries 35 and 39 [3], [13], while Corollaries 7 and 8 are the q-analogues of Corollaries 6 and 7 of [15].

Remark 2. The Corollary 8 case $s = q$ is particularly interesting since the approximants of the continued fraction

$$\frac{1}{a_0} \; \frac{2b_1}{-a_1} \; \frac{b_2}{-a_2} \; -\cdots = \frac{q(aq)_\infty(\frac{a}{q})_\infty}{a^2(\frac{1}{a})_\infty(\frac{1}{a})_\infty} \frac{\widetilde{W}_2}{\widetilde{W}_1}$$

are then given explicitly in terms of $X_n^{(1)}$ and $X_n^{(2)}$. To see this we note that the initial conditions

$$\begin{aligned} 2X_1^{(1)} - a_0 X_0^{(1)} &= 0 \\ X_2^{(2)} - a_1 X_1^{(2)} &= 0 \end{aligned}$$

which follow from (2.7), (2.8), (2.9) and (2.12) when $s = q$, together with the recurrence (2.6), imply that for $s = q$

$$\frac{1}{a_0} \; \frac{2b_1}{-a_1} \; \frac{b_2}{-a_2} \; -\cdots - \frac{b_n}{a_n} = \frac{X_{n+1}^{(2)} X_0^{(1)}}{2X_{n+1}^{(1)} X_1^{(2)}}, \qquad n \geq 0 .$$

Remark 3. In the limit as $m \to \infty$ ($s = q^m \to 0$) Corollary 9 yields a new continued fraction result given by

$$\begin{aligned} \frac{1}{c_0} \; \frac{d_1}{-c_1} \; \frac{d_2}{-c_2} \; -\cdots &= \frac{(1-\frac{a}{q})}{q(1-\frac{a}{b})(1-\frac{a}{c})(1-\frac{a}{d})(1-\frac{a}{e})} \\ &\times \left[W\left(\frac{q}{a}; \frac{q}{b}, \frac{q}{c}, \frac{q}{d}, \frac{q}{e}, q\right) - R \right], \end{aligned} \tag{5.5}$$

$$\begin{aligned} R = \frac{(q, a, \frac{q^2}{a}, \frac{de}{a}, \frac{dc}{a}, \frac{ec}{a})_\infty}{(\frac{dq}{a}, \frac{eq}{a}, \frac{cq}{a}, \frac{dec}{aq}, \frac{aq}{b}, b)_\infty} \\ \times \left\{ {}_3\phi_2\left(\begin{matrix} \frac{q}{d}, \frac{q}{e}, \frac{q}{c} \\ q\frac{b}{a}, \frac{q^2 a}{cde} \end{matrix} ; q, b \right) + \frac{(\frac{q}{d}, \frac{q}{e}, \frac{q}{c}, \frac{bcde}{a^2}, \frac{aq}{b}, \frac{b}{a}, \frac{cde}{a^2}, \frac{a^2 q}{cde}, \frac{dec}{aq})_\infty}{(\frac{ec}{a}, \frac{qb}{a}, \frac{qa}{cde}, \frac{q}{a}, a, \frac{bcde}{qa^2}, \frac{a^2 q^2}{bcde}, \frac{de}{a}, \frac{dc}{a})_\infty} \right. \\ \left. \times {}_3\phi_2\left(\begin{matrix} \frac{de}{a}, \frac{dc}{a}, \frac{ec}{a} \\ \frac{bcde}{a^2}, \frac{dec}{a} \end{matrix} ; q, b \right) \right\} \Bigg/ {}_3\phi_2\left(\begin{matrix} \frac{aq}{bc}, \frac{aq}{bd}, \frac{aq}{be} \\ \frac{aq}{b}, \frac{a^2 q^2}{bcde} \end{matrix} ; q, b \right), \end{aligned}$$

where

$$\begin{aligned} c_n = \Bigg[&- \left(1 - \frac{a}{b}q^{n+1}\right)\left(1 - \frac{a}{c}q^{n+1}\right)\left(1 - \frac{a}{d}q^{n+1}\right)\left(1 - \frac{a}{e}q^{n+1}\right) \\ &- q(1-q^n)(1-aq^n)(1-aq^{n+1})\left(1 - \frac{a^2 q^{n+1}}{bcde}\right) \\ &+ a^2 q^{2n+2} \frac{(1-b)(1-c)(1-d)(1-e)}{bcde} \Bigg] \Big/ (1 - aq^{n+1}), \end{aligned} \tag{5.6}$$

$$d_n = q(1-q^n)\left(1 - \frac{a}{b}q^n\right)\left(1 - \frac{a}{c}q^n\right)\left(1 - \frac{a}{d}q^n\right)\left(1 - \frac{a}{e}q^n\right)\left(1 - \frac{a^2 q^{n+1}}{bcde}\right).$$

A direct proof is obtained by applying our methods to the contiguous relation

$$
\begin{aligned}
&q\left(1-\frac{1}{f}\right)\left(1-\frac{a^2q}{bcdef}\right)\left(1-\frac{a}{f}\right)\left(1-\frac{aq}{f}\right)[W(f+)-W] \\
&+\left(1-\frac{aq}{fb}\right)\left(1-\frac{aq}{fc}\right)\left(1-\frac{aq}{fd}\right)\left(1-\frac{aq}{fe}\right)[W(f-)-W] \\
&+\frac{a^2q^2}{bcdef^2}(1-b)(1-c)(1-d)(1-e)W=0,
\end{aligned} \tag{5.7}
$$

where

$$
\begin{aligned}
W &= W(a;\ b,c,d,e,f) \\
&= {}_8\phi_7\left(\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b, c, d, e, f \\ \sqrt{a}, -\sqrt{a}, \frac{aq}{b}, \frac{aq}{c}, \frac{aq}{d}, \frac{aq}{e}, \frac{aq}{f} \end{matrix}\ ;\ q, \frac{a^2q^2}{bcdef}\right).
\end{aligned} \tag{5.8}
$$

The contiguous relation (5.7) is obtained from (2.5) by taking the limit $g \to 0$ with $fg = a^3q^2/(bcdeh)$ and then replacing $h = q^{-n}$ by f. However, the termination of the preceding ${}_8\phi_7$ is not necessary.

REFERENCES

[1] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Memoirs Amer. Math. Soc., 319 (1985), pp. 1–55.

[2] W. N. Bailey, *Generalized Hypergeometric Series*, Cambridge University Press, London, 1935.

[3] B. C. Berndt, R. L. Lamphere, and B. M. Wilson, *Chapter 12 of Ramanujan's second note-book: Continued fractions*, Rocky Mountain J. Math., 15 (1985), pp. 235–310.

[4] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, 1990.

[5] D. P. Gupta, M. E. H. Ismail, and D. R. Masson, *Associated continuous Hahn polynomials*, Canad. J. Math., 43 (1991), pp. 1263–1280.

[6] ———, *Contiguous relations, basic hypergeometric functions and orthogonal polynomials* II, *associated big q-Jacobi polynomials*, J. Math. Anal. Appl., 171 (1992), pp. 477–497.

[7] D. P. Gupta and D. R. Masson, *Exceptional q-Askey-Wilson polynomials and continued fractions*, Proc. Amer. Math. Soc., 112 (1991), pp. 717–727.

[8] M. E. H. Ismail, J. Letessier, G. Valent, and J. Wimp, *Two families of associated Wilson polynomials*, Canad. J. Math., 42 (1990), pp. 659–695.

[9] M. E. H. Ismail and M. Rahman, *Associated Askey-Wilson polynomials*, Trans. Amer. Math. Soc., 328 (1991), pp. 201–239.

[10] L. Jacobsen, *Domains of validity for some of Ramanujan's continued fraction formulas*, J. Math. Anal. Appl., 143 (1989), pp. 412–437.

[11] W. B. Jones and W. J. Thron, *Continued Fractions: Analytic Theory and Applications*, Addison-Wesley, Reading, MA, 1980.

[12] D. R. Masson, *Some continued fractions of Ramanujan and Meixner-Pollaczck polynomials*, Canad. Math. Bull., 32 (1989), pp. 177–181.

[13] ———, *Wilson polynomials and some continued fractions of Ramanujan*, Rocky Mountain J. Math., 21 (1991), pp. 489–499.

[14] ———, *Associated Wilson polynomials*, Constr. Approx., 7 (1991), pp. 521–534.

[15] ———, *A generalization of Ramanujan's best theorem on continued fractions*, Canad. Math. Reports of the Acad. of Sci., 13 (1991), pp. 167–172.

[16] S. Ramanujan, *Notebooks*, Tata Institute, Bombay, India, 1957.

[17] G. N. Watson, *Ramanujan's continued fraction*, Proc. Cambridge Philos. Soc., 31 (1935), pp. 7–17.

[18] J. A. Wilson, *Hypergeometric series, recurrence relations and some new orthogonal polynomials*, Ph.D. thesis, University of Wisconsin-Madison, 1978.

[19] L. C. Zhang, *Ramanujan's continued fractions for products of gamma functions*, J. Math. Anal. Appl., 174 (1993), pp. 22–52.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 441–449, March 1994

013

SOME q-BETA INTEGRALS ON $SU(n)$ AND $Sp(n)$ THAT GENERALIZE THE ASKEY–WILSON AND NASRALLAH–RAHMAN INTEGRALS*

ROBERT A. GUSTAFSON†

Abstract. $SU(n)$ and $Sp(n)$ generalizations of a q-beta integral of Nasrallah–Rahman are evaluated. Selberg's beta integral can be deduced as a special limiting case of the $Sp(n)$ integral. Extensions of the q-Macdonald–Morris constant term identities for the affine root systems of types $S(BC_n)$, $S(B_n)$, $S(B_n)^v$, $S(C_n)$, $SC(C_n)^v$, and $S(D_n)$ can also be obtained from the $Sp(n)$ integral. There are some additional integral evaluations for $Sp(2)$ and $Sp(3)$.

Key words. multivariate beta integrals, q-beta integrals, Selberg's beta integral, Macdonald–Morris conjecture

AMS subject classifications. 33A15, 33A30, 33A65, 33A75

1. Introduction.

Askey and Wilson [1] evaluated an important integral associated to a family of orthogonal polynomials in five parameters. This integral can be viewed as a q-analog of the classical beta integral. Later, Nasrallah and Rahman [8] and Rahman [9] extended the Askey–Wilson integral by introducing an additional parameter.

Let q be a real number, $0 < q < 1$. For any complex number c define

$$[c]_\infty = [c;q]_\infty = \prod_{k=0}^{\infty}(1 - cq^k).$$

THEOREM 1.1 (Nasrallah–Rahman). *Let $a_i \in \mathbb{C}$, $1 \le i \le 5$, with $|a_i| < 1$. Then*

$$(1.2)\quad \frac{1}{2\pi i}\int_T \frac{\left[z\prod_{i=1}^{5}a_i\right]_\infty \left[z^{-1}\prod_{i=1}^{5}a_i\right]_\infty [z^2]_\infty [z^{-2}]_\infty}{\prod_{i=1}^{5}[a_i z]_\infty [a_i z^{-1}]_\infty}\,\frac{dz}{z} = \frac{2\prod_{k=1}^{5}\left[\prod_{\substack{i=1\\ i\neq k}}^{5}a_i\right]_\infty}{[q]_\infty \prod_{1\le i<j\le 5}[a_i a_j]_\infty},$$

where the unit circle T is taken in the positive direction.

Setting $a_5 = 0$ in (1.2), the identity (1.2) reduces to the Askey–Wilson integral.

In this paper we evaluate integrals associated to the groups $Sp(n)$ and $SU(n)$, which generalize the Nasrallah–Rahman integral. In the $Sp(n)$ case the integrals also generalize the Selberg beta integral [10] (see also [2]) and the Askey–Wilson q-Selberg integral of [2]. In the $SU(n)$ case the integral does not generalize the Selberg beta integral but gives a new kind of $SU(n)$ relative of the $Sp(n)$ q-beta integral. In §4 of this paper some integrals particular to $Sp(2)$ and $Sp(3)$ are also evaluated by a method similar to the general $Sp(n)$ integral evaluation.

There is a theory of multivariate basic hypergeometric series corresponding to the integrals in this paper. If we let the parameter q tend to 1 (at least formally), there are also corresponding multivariate Mellin–Barnes-type integrals and ordinary hypergeometric series. These topics are developed in [5].

* Received by the editors September 13, 1992; accepted for publication (in revised form) March 13, 1993.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843. This research was partially supported by National Science Foundation grant DMS-9002342.

Finally, we mention that in the $Sp(n)$ integrals (2.2) that follow the parameters can be specialized to obtain new extensions of the q-Macdonald–Morris constant-term conjectures [6], [7] for the infinite families of affine root systems of types $S(BC_n)$, $S(B_n)$, $S(B_n)^v$, $S(C_n)$, $S(C_n)^v$, and $S(D_n)$, where $n \geq 1$ (when defined) and for arbitrary parameter q. As an illustration of this extension consider the affine root system of type $S(B_n)$ for $n > 1$. Setting $a_3 = -1$, $a_4 = q^{1/2}$, $a_5 = -q^{1/2}$ in (2.2) and using $[c^2]_\infty = [c]_\infty[-c]_\infty[q^{1/2}c]_\infty[-q^{1/2}c]_\infty$ for $c \in \mathbb{C}$, one obtains

$$\begin{aligned}
&\frac{1}{(2\pi i)^n}\int_{T^n}\prod_{1\leq j<k\leq n}\frac{[t_jt_k^{-1}]_\infty[t_j^{-1}t_k]_\infty[t_jt_k]_\infty[t_j^{-1}t_j^{-1}]_\infty}{[bt_jt_k^{-1}]_\infty[bt_j^{-1}t_k]_\infty[bt_jt_k]_\infty[bt_j^{-1}t_k^{-1}]_\infty}\\
&\qquad\cdot\prod_{j=1}^{n}\frac{[t_j]_\infty[t_j^{-1}]_\infty[a_1a_2b^{2n-2}qt_j]_\infty[a_1a_2b^{2n-2}qt_j^{-1}]_\infty}{t_j\prod_{k=1}^{2}[a_kt_j]_\infty[a_kt_j^{-1}]_\infty}dt_j\\
&\qquad=2^nn!\prod_{j=1}^{n}\left\{\frac{[b]_\infty[b^{2n+2j-4}qa_1^2a_2^2]_\infty[qb^{j-1}]_\infty}{[b^j]_\infty[q]_\infty[b^{n+j-2}qa_1a_2]_\infty[a_1a_2b^{j-1}]_\infty[qb^{2j-2}]_\infty}\right.\\
&\qquad\qquad\left.\cdot\prod_{k=1}^{2}\frac{[b^{n+j-2}qa_k]_\infty[a_kb^{j-1}]_\infty}{[a_k^2b^{2j-2}]_\infty}\right\},
\end{aligned}\tag{1.3}$$

where a_1, a_2, $b \in \mathbb{C}$, $\max\{|a_1|, |a_2|, |b|\} < 1$, and T^n is the n-fold direct product of the unit circle traversed in the positive direction. If we set $a_2 = 0$, (1.3) reduces to an identity that is equivalent to the q-Macdonald–Morris constant-term conjecture for B_n (see [2]).

2. A q-beta integral on $Sp(n)$. We give an extension of the q-Selberg integral evaluated in [2]. In the one-dimensional case this reduces to an integral evaluated by Nasrallah and Rahman [8], [9]. The proof of this integral identity is similar to that in [2].

THEOREM 2.1. *Let $n \geq 1$ and $a_1, \ldots, a_5, b, q \in \mathbb{C}$ with $\max\{|a_1|, \ldots, |a_5|, |b|, |q|\} < 1$. Set $A = \prod_{i=1}^{5} a_i$. If T^n is the n-fold direct product of the unit circle $\{t \in \mathbb{C} \mid |t| = 1\}$ traversed in the positive direction, then we have*

$$\begin{aligned}
&\frac{1}{(2\pi i)^n}\int_{T^n}\prod_{1\leq j<k\leq n}\frac{[t_jt_k^{-1}]_\infty[t_j^{-1}t_k]_\infty[t_jt_k]_\infty[t_j^{-1}t_k^{-1}]_\infty}{[bt_jt_k^{-1}]_\infty[bt_j^{-1}t_k]_\infty[bt_jt_k]_\infty[bt_j^{-1}t_k^{-1}]_\infty}\\
&\qquad\cdot\prod_{j=1}^{n}\frac{[t_j^2]_\infty[t_j^{-2}]_\infty[Ab^{2n-2}t_j]_\infty[Ab^{2n-2}t_j^{-1}]_\infty}{\prod_{k=1}^{5}\{[a_kt_j]_\infty[a_kt_j^{-1}]_\infty\}t_j}dt_j\\
&\qquad=2^nn!\prod_{j=1}^{n}\frac{[b]_\infty\prod_{i=1}^{5}[b^{n+j-2}Aa_i^{-1}]_\infty}{[b^j]_\infty[q]_\infty\prod_{1\leq k<\ell\leq 5}[a_ka_\ell b^{j-1}]_\infty}.
\end{aligned}\tag{2.2}$$

Proof. Since the $n = 1$ case of (2.2) is proved in [9], we may assume that $n \geq 2$. Denote the integral on the left-hand side of (2.2) by $I_n(a_1, \ldots, a_5; b; q)$. Let $c_j \in \mathbb{C}$,

$|c_j| < 1$, for $1 \le j \le 2n+3$. Set $C = \prod_{j=1}^{2n+3} c_j$. In [3, Thm. 4.1] we have evaluated the integral

$$\frac{1}{(2\pi i)^n}\int_{T^n} \frac{\prod_{1\le j<k\le n}\{[t_jt_k^{-1}]_\infty[t_j^{-1}t_k]_\infty[t_jt_k]_\infty[t_j^{-1}t_k^{-1}]_\infty\}}{\prod_{i=1}^{2n+3}\prod_{j=1}^{n}[c_it_j]_\infty[c_it_j^{-1}]_\infty}$$

$$\cdot \prod_{j=1}^{n}\left\{[z_j^2]_\infty[z_j^{-2}]_\infty[Cz_j]_\infty[Cz_j^{-1}]_\infty\frac{dz_j}{z_j}\right\} \tag{2.3}$$

$$= \frac{2^n n!\prod_{i=1}^{2n+3}[Cc_i^{-1}]_\infty}{[q]_\infty^n \prod_{1\le j<k\le 2n+3}[c_jc_k]_\infty}.$$

With notation as in the preceding, consider the integral

(2.4)

$$\frac{1}{(2\pi i)^{2n-1}}\int_{T^n}\int_{T^{n-1}} \frac{\prod_{1\le j<k\le n}\{[t_jt_k^{-1}]_\infty[t_j^{-1}t_k]_\infty[t_jt_k]_\infty[t_j^{-1}t_k^{-1}]_\infty\}}{\prod_{j=1}^{n}\prod_{k=1}^{5}[a_kt_j]_\infty[a_kt_j^{-1}]_\infty}$$

$$\cdot \frac{\prod_{j=1}^{n}[t_j^2]_\infty[t_j^{-2}]_\infty \prod_{1\le j<k\le n-1}\{[s_js_k^{-1}]_\infty[s_j^{-1}s_k]_\infty[s_js_k]_\infty[s_j^{-1}s_k^{-1}]_\infty\}}{\prod_{j=1}^{n}\prod_{k=1}^{n-1}\{[b^{1/2}s_kt_j]_\infty[b^{1/2}s_k^{-1}t_j]_\infty[b^{1/2}s_kt_j^{-1}]_\infty[b^{1/2}s_k^{-1}t_j^{-1}]_\infty\}}$$

$$\cdot \prod_{k=1}^{n-1}\frac{[s_k^2]_\infty[s_k^{-2}]_\infty\left[s_kb^{2n-3/2}\prod_{i=1}^{5}a_i\right]_\infty\left[s_k^{-1}b^{2n-3/2}\prod_{i=1}^{5}a_i\right]_\infty}{\left[s_kb^{n-3/2}\prod_{i=1}^{t}a_i\right]_\infty\left[s_k^{-1}b^{n-3/2}\prod_{i=1}^{5}a_i\right]_\infty}\frac{ds_k}{s_k}$$

$$\cdot \prod_{j=1}^{n}\frac{\left[t_jb^{n-1}\prod_{i=1}^{5}a_i\right]_\infty\left[t_j^{-1}b^{n-1}\prod_{i=1}^{5}a_i\right]_\infty}{t_j}dt_j,$$

where $b^{1/2}$ is any fixed square root of b. In the integral (2.4) we may use identity (2.3) to evaluate the interior integral either with respect to the set of variables $\{s_1,\dots,s_{n-1}\}$ or, by changing the order of integration, with respect to the set of variables $\{t_1,\dots,t_n\}$. Equating the resulting integrals, we obtain

$$2^{n-1}(n-1)!\frac{[b^n]_\infty}{[q]_\infty^{n-1}[b]_\infty^n}I_n(a_1,\dots,a_5;b;q)$$

$$= \frac{2^n n!\prod_{i=1}^{5}[b^{n-1}Aa_i^{-1}]_\infty}{[q]_\infty^{n-1}[b]_\infty^{n-1}\prod_{1\le j<k\le 5}[a_ja_k]_\infty}I_{n-1}(a_1b^{1/2},\dots,a_5b^{1/2};b;q). \tag{2.5}$$

We finish the proof of identity (2.2) by doing induction on n, using identity (2.5) and the Nasrallah–Rahman integral for the case $n = 1$. □

3. An analogous q-beta integral on $SU(n)$. We evaluate an $SU(n)$ integral that is analogous to the $Sp(n)$ integral in (2.2). Actually, the $SU(n)$ integral evaluation is slightly different for even n and for odd n, $n > 1$. In the one-dimensional ($SU(2)$) case this integral reduces to the Nasrallah–Rahman integral [8], [9]. In the $SU(3)$ case the integral reduces to one evaluated in [3, Thm. 2.1] (see also [4]).

THEOREM 3.1. *Let $n \ge 1$ and $a, b, c_1, c_2, c_3, d_1, d_2 \in \mathbb{C}$, with $\max\{|a|, |b|, |c_1|, |c_2|, |c_3|, |d_1|, |d_2|, |q|\} < 1$. Then*

(3.2a)

$$\frac{1}{(2\pi i)^{2n-1}}\int_{T^{2n-1}}\frac{\prod_{\substack{i,j=1\\ i\ne j}}^{2n}[z_iz_j^{-1}]_\infty\prod_{j=1}^{2n}\left[z_j(ab)^{2n-2}\prod_{i=1}^{3}c_i\prod_{k=1}^{2}d_k\right]_\infty}{\prod_{1\le i<j\le 2n}[az_iz_j]_\infty[bz_i^{-1}z_j^{-1}]_\infty\prod_{j=1}^{2n}\left\{\prod_{i=1}^{3}[c_iz_j]_\infty\prod_{k=1}^{2}[d_kz_j^{-1}]_\infty\right\}}$$

$$\cdot\frac{dz_1}{z_1}\cdots\frac{dz_{2n-1}}{z_{2n-1}}$$

$$= \frac{(2n)!\prod_{j=1}^{2}\left[a^{2n-2}b^{n-1}d_j\prod_{i=1}^{2}c_i\right]_\infty}{[q]_\infty^{2n-1}[b^{n-1}d_1d_2]_\infty\prod_{1\le i<j\le 3}[a^{n-1}c_ic_j]_\infty[a^n]_\infty[b^n]_\infty}$$

$$\cdot\prod_{j=1}^{n}\left\{\frac{\prod_{1\le i<k\le 3}[(ab)^{n+j-2}c_ic_kd_1d_2]_\infty}{\prod_{i=1}^{3}\prod_{k=1}^{2}[(ab)^{j-1}c_id_k]_\infty}\right\}$$

$$\cdot\prod_{j=1}^{n-1}\left\{\frac{\prod_{k=1}^{2}\left[a^{n+j-2}b^{n+j-1}d_k\prod_{i=1}^{3}c_i\right]_\infty}{[(ab)^j]_\infty[a^jb^{j-1}d_1d_2]_\infty\prod_{1\le i<k\le 3}[a^{j-1}b^jc_ic_k]_\infty}\right\},$$

where $\prod_{j=1}^{2n} z_j = 1$ *and the integral in each variable* $z_1, \dots, z_{2n-1}$ *is over the unit circle* T *taken in the positive direction. We also have*

(3.2b)

$$\frac{1}{(2\pi i)^{2n}} \int_{T^{2n}} \frac{\prod_{\substack{i,j=1 \\ i\neq j}}^{2n+1} [z_i z_j^{-1}]_\infty \prod_{j=1}^{2n+1} \left[z_j (ab)^{2n-1} \prod_{i=1}^{3} c_i \prod_{k=1}^{2} d_k \right]_\infty}{\prod_{1\le i<j\le 2n+1} [a z_i z_j]_\infty [b z_i^{-1} z_j^{-1}]_\infty \prod_{j=1}^{2n+1} \left\{ \prod_{i=1}^{3} [c_i z_j]_\infty \prod_{k=1}^{2} [d_k z_j^{-1}]_\infty \right\}} \cdot \frac{dz_1}{z_1} \cdots \frac{dz_{2n}}{z_{2n}}$$

$$= \frac{(2n+1)! \left[a^{2n-1} b^{n-1} \prod_{i=1}^{3} c_i \prod_{k=1}^{2} d_k \right]_\infty \left[a^{2n-1} b^n \prod_{i=1}^{3} c_i \right]_\infty}{[q]_\infty^{2n} \prod_{i=1}^{3} [a^n c_i]_\infty \prod_{k=1}^{2} [b^n d_k]_\infty \left[a^{n-1} \prod_{i=1}^{3} c_i \right]_\infty}$$

$$\cdot \prod_{j=1}^{n} \left\{ \frac{\prod_{k=1}^{2} \left[a^{n+j-2} b^{n+j-1} d_k \prod_{i=1}^{3} c_i \right]_\infty \prod_{i\le i<k\le 3} [(ab)^{n+j-1} c_i c_k d_1 d_2]_\infty}{[(ab)^j]_\infty [a^j b^{j-1} d_1 d_2]_\infty \prod_{i\le i<k\le 3} [a^{j-1} b^j c_i c_k]_\infty \prod_{i=1}^{3} \prod_{k=1}^{2} [(ab)^{j-1} c_i d_k]_\infty} \right\},$$

where $\prod_{j=1}^{2n+1} z_j = 1$ *and the integral in each variable* $z_1, \dots, z_{2n}$ *is over the unit circle in the positive direction.*

Proof of (3.2a). Since the $n = 1$ case of (3.2a) is proved in [9], we may assume that $n \ge 2$. Let $a_i, b_j \in \mathbb{C}$, $|a_i|, |b_j| < 1$, for $1 \le i \le n$ and $1 \le j \le n+1$. Set $A = \prod_{i=1}^{n} a_i$ and $B = \prod_{j=1}^{n+1} b_j$. Then in [3, Thm. 2.1] we have evaluated the integral

$$\frac{1}{(2\pi i)^{n-1}} \int_{T^{n-1}} \frac{\prod_{k=1}^{n} [ABz_k]_\infty \prod_{\substack{i,j=1 \\ i\neq j}}^{n} [z_i z_j^{-1}]_\infty}{\prod_{k=1}^{n} \left\{ \prod_{i=1}^{n} [a_i z_k^{-1}]_\infty \prod_{j=1}^{n+1} [b_j z_k]_\infty \right\}} \frac{dz_1}{z_1} \cdots \frac{dz_{n-1}}{z_{n-1}}$$

(3.3)

$$= \frac{n! \prod_{j=1}^{n+1} [b_j^{-1} AB]_\infty \prod_{i=1}^{n} [a_i B]_\infty}{[q]_\infty^{n-1} [A]_\infty \prod_{j=1}^{n+1} [b_j^{-1} B]_\infty \prod_{i=1}^{n} \prod_{j=1}^{n+1} [a_i b_i]_\infty},$$

where $\prod_{k=1}^{n} z_k = 1$. With notation as in the preceding, consider the integral

(3.4)

$$\frac{1}{(2\pi i)^{4n-1}}\int_{T^{2n-1}}\int_{T^n}\int_{T^n}\frac{\prod_{\substack{i,j=1\\ i\neq j}}^{2n}[z_iz_j^{-1}]_\infty\prod_{1\le i<j\le n}\{[t_it_j]_\infty[t_it_j^{-1}]_\infty}{\prod_{i=1}^{2n}\prod_{j=1}^{n}\{[a^{1/2}z_it_j]_\infty[a^{1/2}z_it_j^{-1}]_\infty}$$

$$\cdot\frac{[t_i^{-1}t_j]_\infty[t_i^{-1}t_j^{-1}]_\infty[s_is_j]_\infty[s_is_j^{-1}]_\infty[s_i^{-1}s_j]_\infty[s_i^{-1}s_j^{-1}]_\infty\}}{[b^{1/2}z_i^{-1}s_j]_\infty[b^{1/2}z_i^{-1}s_j^{-1}]_\infty\}}$$

$$\cdot\prod_{i=1}^{2n}\frac{\left[z_i(ab)^{2n-2}d_1d_2\prod_{k=1}^{3}c_k\right]_\infty}{\left[z_i(ab)^{n-2}d_1d_2\prod_{k=1}^{3}c_k\right]_\infty}$$

$$\cdot\prod_{j=1}^{n}\left\{\frac{\left[t_ja^{n-3/2}\prod_{k=1}^{3}c_k\right]_\infty}{\prod_{k=1}^{3}[a^{-1/2}c_kt_j]_\infty[a^{-1/2}c_kt_j^{-1}]_\infty}\right.$$

$$\cdot\frac{\left[t_j^{-1}a^{n-3/2}\prod_{k=1}^{3}c_k\right]_\infty\left[s_ja^{n-2}b^{n-3/2}d_1d_2\prod_{k=1}^{3}c_k\right]_\infty}{\left[s_ja^{n-2}b^{-1/2}\prod_{k=1}^{3}c_k\right]_\infty\left[s_j^{-1}a^{n-2}b^{-1/2}\prod_{k=1}^{3}c_k\right]_\infty}$$

$$\left.\cdot\frac{\left[s_j^{-1}a^{n-2}b^{n-3/2}d_1d_2\prod_{k=1}^{3}c_k\right]_\infty[t_j^2]_\infty[t_j^{-2}]_\infty[s_j^2]_\infty[s_j^{-2}]_\infty}{\prod_{i=1}^{2}[b^{-1/2}d_is_j]_\infty[b^{-1/2}d_is_j^{-1}]_\infty}\right\}\frac{dt_jds_j}{t_js_j}\frac{dz_1}{z_1}\cdots\frac{dz_{2n-1}}{z_{2n-1}},$$

where $\prod_{j=1}^{2n}z_j=1$ and $a^{1/2}$, $b^{1/2}$ are any fixed square roots of a, b, respectively. Applying identity (2.3) twice, we evaluate the interior integrals in (3.4) by first integrating

with respect to the set of variables $\{t_1,\dots,t_n\}$ and then integrating with respect to the variables $\{s_1,\dots,s_n\}$. The resulting integral in the variables $\{z_1,\dots,z_{2n-1}\}$ is, up to factors independent of $\{z_1,\dots,z_{2n-1}\}$, just the integral on the left-hand side of (3.2a). On the other hand, by reversing the order of integration we may evaluate (3.4) by applying identity (3.3) to first integrate with respect to the variables $\{z_1,\dots,z_{2n-1}\}$, then applying (2.3) to integrate with respect to the variables $\{s_1,\dots,s_n\}$, and finally applying (2.2) to integrate with respect to the variables $\{t_1,\dots,t_n\}$. The resulting identity is (3.2a). This completes the proof of (3.2a). □

Proof of (3.2b). The proof of identity (3.2b) is entirely similar to that of (3.2a). The only difference is that, instead of the integral (3.4), we consider the integral

(3.5)

$$\frac{1}{(2\pi i)^{4n}}\int_{T^{2n}}\int_{T^n}\int_{T^n}\frac{\displaystyle\prod_{\substack{i,j=1\\ i\neq j}}^{2n+1}[z_iz_j^{-1}]_\infty\prod_{1\le i<j\le n}\{[t_it_j]_\infty[t_it_j^{-1}]_\infty}{\displaystyle\prod_{i=1}^{2n+1}\prod_{j=1}^{n}\{[a^{1/2}z_it_j]_\infty[a^{1/2}z_it_j^{-1}]_\infty}$$

$$\cdot\frac{[t_i^{-1}t_j]_\infty[t_i^{-1}t_j^{-1}]_\infty[s_is_j]_\infty[s_is_j^{-1}]_\infty[s_i^{-1}s_j]_\infty[s_i^{-1}s_j^{-1}]_\infty\}}{[b^{1/2}z_i^{-1}s_j]_\infty[b^{1/2}z_i^{-1}s_j^{-1}]_\infty\}}$$

$$\cdot\prod_{i=1}^{2n+1}\frac{\left[z_j(ab)^{2n-1}d_1d_2\displaystyle\prod_{k=1}^{3}c_k\right]_\infty}{[c_3z_j]_\infty[b^{n-1}d_1d_2z_j]_\infty[a^{n-1}c_1c_2z_j^{-1}]_\infty}$$

$$\cdot\prod_{j=1}^{n}\left\{\frac{[a^nc_1c_2t_j]_\infty[a^nc_1c_2t_j^{-1}]_\infty[b^nd_1d_2s_j]_\infty[b^nd_1d_2s_j^{-1}]_\infty}{\displaystyle\prod_{k=1}^{2}([a^{-1/2}c_kt_j]_\infty[a^{-1/2}c_kt_j^{-1}]_\infty[b^{-1/2}d_ks_j]_\infty[b^{-1/2}d_ks_j^{-1}]_\infty)}\right.$$

$$\left.\cdot\,[t_j^2]_\infty[t_j^{-2}]_\infty[s_j^2]_\infty[s_j^{-2}]_\infty\frac{dt_j}{t_j}\,\frac{ds_j}{s_j}\right\}\frac{dz_1}{z_1}\cdots\frac{dz_{2n}}{z_{2n}},$$

where $\prod_{j=1}^{2n+1}z_j=1$ and the notation is as for (3.4). □

4. Some q-beta integrals on $Sp(2)$ and $Sp(3)$. In $Sp(2)$ and $Sp(3)$ we can evaluate some integrals related to (2.2) that do not generalize to $Sp(n)$. The proofs involve substituting variables in place of the parameters $a_1,\dots,a_4$ in (2.2) and integrating by using Fubini's theorem.

THEOREM 4.1. *With notation as in Theorem* 2.1, *let* $a_1,a_2,b_1,b_2,q\in\mathbb{C}$, *with* $\max\{|a_1|,|a_2|,|b_1|,|b_2|,|q|\}<1$. *Then we have*

$$\frac{1}{(2\pi i)^2}\int_{T^2}\frac{[t_1t_2^{-1}]_\infty[t_1^{-1}t_2]_\infty[t_1t_2]_\infty[t_1^{-1}t_2^{-1}]_\infty}{\prod_{k=1}^{2}[b_kt_1t_2^{-1}]_\infty[b_kt_1^{-1}t_2]_\infty[b_kt_1t_2]_\infty[b_kt_1^{-1}t_2^{-1}]_\infty}$$

$$\cdot\prod_{j=1}^{2}\frac{[t_j^2]_\infty[t_j^{-2}]_\infty dt_j}{\prod_{k=1}^{2}[a_kt_j]_\infty[a_kt_j^{-1}]_\infty t_j} \tag{4.2}$$

$$=\frac{8[(a_1a_2b_1b_2)^2]_\infty[b_1]_\infty[b_2]_\infty}{[q]_\infty^2[a_1a_2]_\infty[b_1b_2]_\infty[a_1a_2b_1b_2]_\infty\prod_{k=1}^{2}[a_k^2b_1b_2]_\infty[a_1a_2b_k]_\infty[b_k^2]_\infty}.$$

Proof. Consider the integral on the left-hand side of (2.2) for $n=2$. Let $a_5=0$, $a_3=b_2^{1/2}z$, $a_4=b_2^{1/2}z^{-1}$, and $b=b_1$. Multiply the integral by $[z^2]_\infty[z^{-2}]_\infty/z$, and integrate with respect to z along the unit circle traversed in the positive direction. We obtain the integral

(4.3)

$$\frac{1}{(2\pi i)^3}\int_T\int_{T^2}\frac{[t_1t_2^{-1}]_\infty[t_1^{-1}t_2]_\infty[t_1t_2]_\infty[t_1^{-1}t_2^{-1}]_\infty}{[b_1t_1t_2^{-1}]_\infty[b_1t_1^{-1}t_2]_\infty[b_1t_1t_2]_\infty[b_1t_1^{-1}t_2^{-1}]_\infty}$$

$$\cdot\prod_{j=1}^{2}\frac{[t_j^2]_\infty[t_j^{-2}]_\infty dt_j}{[b_2^{1/2}zt_j]_\infty[b_2^{1/2}zt_j^{-1}]_\infty[b_2^{1/2}z^{-1}t_j]_\infty[b_2^{1/2}zt_j^{-1}]_\infty\prod_{k=1}^{2}([a_kt_j]_\infty[a_kt_j^{-1}]_\infty)t_j}$$

$$\cdot\frac{[z^2]_\infty[z^{-2}]_\infty}{z}dz.$$

Use identity (2.2) to evaluate the integral (4.3) with respect to the variables t_1 and t_2 first, and then use (2.2) (the Askey–Wilson integral) again to integrate with respect to z. Now reverse the order of integration, and evaluate the integral (4.3) with respect to z first. The resulting identity is (4.2). □

THEOREM 4.4. *With notation as in the preceding, let* $b_1,b_2,b_3,q\in\mathbb{C}$, *with* $\max\{|b_1|,|b_2|,|b_3|,|q|\}<1$. *Then we have*

$$\frac{1}{(2\pi i)^2}\int_{T^2}\frac{[t_1t_2^{-1}]_\infty[t_1^{-1}t_2]_\infty[t_1t_2]_\infty[t_1^{-1}t_2^{-1}]_\infty}{\prod_{k=1}^{3}[b_kt_1t_2^{-1}]_\infty[b_kt_1^{-1}t_2]_\infty[b_kt_1t_2]_\infty[b_kt_1^{-1}t_2^{-1}]_\infty}\prod_{j=1}^{2}\frac{[t_j^2]_\infty[t_j^{-2}]_\infty dt_j}{t_j}$$

$$=\frac{8[qb_1b_2b_3]_\infty}{[q]_\infty^2\prod_{1\le j<k\le 3}[b_jb_k]_\infty}\cdot\prod_{j=1}^{3}\frac{[b_j]_\infty}{[b_j^2]_\infty}. \tag{4.5}$$

Proof. Set $a_1 = b_3^{1/2}z$ and $a_2 = b_3^{1/2}z^{-1}$ in the integral on the left-hand side of (4.2). Multiply the resulting integral by $[z^2]_\infty[z^{-2}]_\infty/z$, and integrate with respect to z along the contour T. Just as before, using (4.2) and (2.2) to evaluate the integrals and reversing the order of integration, we obtain identity (4.5). □

THEOREM 4.6. *With notation as in Theorem* 2.1, *let* $b_1, b_2, q \in \mathbb{C}$ *with* $\max\{|b_1|, |b_2|, |q|\} < 1$. *We have*

(4.7)

$$\frac{1}{(2\pi i)^3}\int_{T^3}\prod_{1\le j<k\le 3}\frac{[t_jt_k^{-1}]_\infty[t_j^{-1}t_k]_\infty[t_jt_k]_\infty[t_j^{-1}t_k^{-1}]_\infty}{\prod_{\ell=1}^{2}[b_\ell t_jt_k^{-1}]_\infty[b_\ell t_j^{-1}t_k]_\infty[b_\ell t_jt_k]_\infty[b_\ell t_j^{-1}t_k^{-1}]_\infty}\prod_{j=1}^{3}\frac{[t_j^2]_\infty[t_j^{-2}]_\infty dt_j}{t_j}$$

$$= 3!2^3\frac{[qb_1^3b_2^3]_\infty}{[b_1b_2]_\infty[b_1b_2^2]_\infty[b_1^2b_2]_\infty[b_1^2b_2^2]_\infty}\prod_{j=1}^{3}\frac{[b_1]_\infty[b_2]_\infty}{[q]_\infty[b_1^j]_\infty[b_2^j]_\infty}.$$

Proof. Set $a_1 = b_2^{1/2}u_1$, $a_2 = b_2^{1/2}u_2$, $a_3 = b_2^{1/2}u_1^{-1}$, $a_4 = b_2^{1/2}u_2^{-1}$, and $a_5 = 0$ in (2.2). Multiply by $[u_1u_2^{-1}]_\infty[u_1^{-1}u_2]_\infty[u_1u_2]_\infty[u_1^{-1}u_2^{-1}]_\infty\prod_{j=1}^{2}[u_j^2]_\infty[u_j^{-2}]_\infty du_j/u_j$, and integrate with respect to the variables u_1 and u_2 along T^2. Using (4.5) and (2.2) to evaluate the integrals and reversing the order of integration, we obtain identity (4.7). □

REFERENCES

[1] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985).

[2] R. A. GUSTAFSON, *A generalization of Selberg's beta integral*, Bull. Amer. Math. Soc., 22 (1990), pp. 97–105.

[3] ———, *Some q-beta and Mellin–Barnes integrals with many parameters associated to the classical groups*, SIAM J. Math. Anal., 23 (1982), pp. 525–551.

[4] ———, *Some q-beta and Mellin–Barnes integrals on compact Lie groups and Lie algebras*, Trans. Amer. Math. Soc., to appear.

[5] ———, *Hypergeometric series well-poised on Sp(n) and SU(n) and Macdonald's hypergeometric functions*, in preparation.

[6] I. G. MACDONALD, *Some conjectures for root systems*, SIAM J. Math. Anal., 13 (1982), pp. 988–1007.

[7] W. G. MORRIS, *Constant Term Identities for Finite and Affine Root Systems: Conjectures and Theorems*, Ph.D. thesis, University of Wisconsin, Madison, WI, 1982.

[8] B. NASRALLAH AND M. RAHMAN, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, SIAM J. Math. Anal., 16 (1985), pp. 186–197.

[9] M. RAHMAN, *An integral representation of a* ${}_{10}\phi_9$ *and continuous bi-orthogonal* ${}_{10}\phi_9$ *rational functions*, Canad. J. Math., 39 (1986), pp. 601–618.

[10] A. SELBERG, *Bemerkinger om et multipelt integral*, Norsk Mat. Tiddskr., 26 (1944), pp. 71–78.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 450–461, March 1994

014

ON THE STRUVE TRANSFORMATION*

P. HEYWOOD† AND P. G. ROONEY‡

Abstract. The Struve or $\mathcal{H}_\nu$ transformation is defined for suitable functions f by

$$(\mathcal{H}_\nu f)(x) = \int_0^\infty (xt)^{1/2}\mathbb{H}_\nu(xt)f(t)dt,$$

where $\mathbb{H}_\nu$ is the Struve function. It is known to be bounded on weighted L_p spaces with weights $t^{p\mu-1}$ if $1 \le p < \infty$, $\mu \ge \gamma(p)$, and $\nu+\frac{1}{2} < \mu < \nu+\frac{5}{2}$. Inversion formulas have been given for $\mathcal{H}_\nu$ but with the added restriction that $\mu > -(\nu+\frac{1}{2})$. Here inversion formulas are given for $\mu \le -(\nu+\frac{1}{2})$.

Key words. Struve function, Mellin transformation, multiplier

AMS subject classification. 44A15

1. Introduction. The spaces $\mathcal{L}_{\mu,p}$ are defined for $1 \le p \le \infty$ to consist of those Lebesgue-measurable functions f on $(0,\infty)$ such that $\|f\|_{\mu,p} < \infty$, where

$$\|f\|_{\mu,p} = \begin{cases} \left[\int_0^\infty |x^\mu f(x)|^p\, dx/x\right]^{1/p}, & 1 \le p < \infty, \\ \operatorname{ess\ sup}_{x>0} |x^\mu f(x)|, & p = \infty. \end{cases}$$

If X and Y are Banach spaces, we denote by $[X,Y]$ the collection of bounded operators from X to Y, $[X,X]$ being abbreviated to $[X]$. Also, we denote by $\mathcal{C}_0$ the collection of functions, continuous and compactly supported on $(0,\infty)$; clearly, $\mathcal{L}_{\mu,p} = L_p((0,\infty), x^{p\mu-1}\,dx)$, and thus, or from [7, Lemma 2.2], $\mathcal{C}_0$ is dense in $\mathcal{L}_{\mu,p}$ for $1 \le p < \infty$. Furthermore, if $1 \le r \le \infty$, the conjugate index r' is defined by $1/r + 1/r' = 1$.

In earlier papers [4], [9] we considered the Struve transformation $\mathcal{H}_\nu$ defined for $f \in \mathcal{C}_0$ by

$$(\mathcal{H}_\nu f)(x) = \int_0^\infty (xt)^{1/2}\mathbb{H}_\nu(xt)f(t)\,dt,$$

where $\mathbb{H}_\nu(z)$ is the Struve function defined by

$$\mathbb{H}_\nu(z) = \sum_{m=0}^\infty (-1)^m (z/2)^{\nu+2m+1}/(\Gamma(m+\tfrac{3}{2})\Gamma(\nu+m+\tfrac{3}{2})).$$

In [9] it is shown that if $1 < p < \infty$, $\nu+\frac{1}{2} < \mu < \nu+\frac{5}{2}$, and $\mu \ge \gamma(p)$ where $\gamma(p) = \max(1/p, 1/p')$, then $\mathcal{H}_\nu \in [\mathcal{L}_{\mu,p}, \mathcal{L}_{1-\mu,q}]$ for all $q \ge p$ such that $q' \ge 1/\mu$, whereas the boundedness on $\mathcal{L}_{\mu,1}$ is given in [4], being essentially the same range of boundedness, with $\gamma(1) = 1$. Note that if $1 \le p \le \infty$, then $\frac{1}{2} \le \gamma(p) \le 1$,

* Received by the editors March 26, 1992; accepted for publication February 12, 1993. This research was supported by Natural Sciences and Engineering Research Council of Canada grant A4048.

† Department of Mathematics, Edinburgh University, JCMB, King's Buildings, Edinburgh, EH93JZ, Scotland(hillary@mathematics.edinburgh.ac.uk).

‡ Department of Mathematics, University of Toronto, Toronto, Ontario M5S 1A1, Canada(rooney@math.toronto.edu).

with equality at the lower endpoint if and only if $p = 2$. Since, for boundedness, $\frac{1}{2} \le \gamma(p) \le \mu < \nu + \frac{5}{2}$, our boundedness result requires that $\nu > -2$.

Inversion formulas for $\mathcal{H}_\nu$ are given in [4] and [9]. The inversion formula given in [4, Thm. 6.1] is valid for $1 \le p < \infty$, $\mu \ge \gamma(p)$, and $\max(\nu+\frac{1}{2}, -(\nu+\frac{1}{2})) < \mu < \nu+\frac{5}{2}$, whereas that given in [9, Thm. 6.3] is valid on a subset of this range of μ and ν. Thus no inversion formulas have been given for $\gamma(p) \le \mu \le -(\nu+\frac{1}{2})$, $\mu < \nu + \frac{5}{2}$, and in this paper we shall find such formulas for this range of μ and ν. Also, when $\mu < 1$ we shall need an inversion formula for $\mathcal{H}_\nu$ that is somewhat different from the one given in [4], and we shall develop it.

It transpires that the case $\mu = -(\nu+\frac{1}{2})$ is a special case. For, as noted in [9, Thm. 5.2], the range of the Struve transformation of order ν on $\mathcal{L}_{\mu,p}$ is the same as that of the Hankel transformation of order $\nu+1$ on $\mathcal{L}_{\mu,p}$, except possibly when $\mu = -(\nu+\frac{1}{2})$, and in order to find inversion formulas for this case we must investigate the range of $\mathcal{H}_\nu$. This causes our program in this paper to divide naturally into three cases. In §2 we deal with the range $\mu \ne -(\nu+\frac{1}{2})$, $\mu < \min(1, \nu+\frac{5}{2})$, and we prove two theorems, one of which is for a part of the range already covered in [4] but which we will need later. Sections 3–5 are devoted to the case $\mu = -(\nu+\frac{1}{2})$. In §3 we show that the range of the Struve transformation in this case is a proper subset of the range of the Hankel transformation of order $\nu+1$, and we characterize that range. In §4 we find an inversion formula for the case $\mu = -(\nu+\frac{1}{2})$, $\mu \ne \frac{1}{2}$, and in §5 we deal with the case $\mu = -(\nu+\frac{1}{2})$, $\mu = \frac{1}{2}$. It seems strange that this last case should be the most elusive since the conditions yield $\nu = -1$, so that the problem is that of inverting $\mathcal{H}_{-1}$ on $\mathcal{L}_{1/2,2} = L^2(0,\infty)$, where one usually expects things to be simpler.

Throughout the paper we use the notation $\int^{\to\infty}$ and $\int_{\to 0}$, which are explained in [11, §1.7]. This causes some of the formulas taken from [1] and [2] to look somewhat different since those formulas are often improper Riemann integrals. In particular, we will have

$$(\mathcal{H}_\nu f)(x) = \int_0^{\to\infty} (xt)^{1/2} \mathbb{H}_\nu(xt) f(t)\, dt$$

whenever this last integral converges. Other notations used include A_ν for

$$\cot(\nu\pi)/\Gamma(\nu+2),$$

and $m_\nu(s)$, which is defined by

$$m_\nu(s) = 2^{s-1/2}\Gamma((\nu+s+\tfrac{1}{2})/2)/\Gamma((\nu-s+\tfrac{3}{2})/2).$$

One of our tools in this work will be the Mellin transformation $\mathcal{M}$, whose properties are summarized in [9, §1].

2. Inversion for $\mu \ne -(\nu+\frac{1}{2})$. In this section we prove two inversion theorems for the range $\mu < \min(1, \nu+\frac{5}{2})$, which are valid when $\mu > \max(-(\nu+\frac{1}{2}), \nu+\frac{3}{2})$ and $\mu < -(\nu+\frac{1}{2})$, respectively. The range of μ and ν in the first theorem is actually within that for which [4, Thm. 6.1] is valid, but the form obtained here is considerably simpler then the form in that theorem, and we shall need the result obtained later.

THEOREM 2.1. *Suppose $f \in \mathcal{L}_{\mu,p}$, where $1 < p < \infty$, $\mu \ge \gamma(p)$, and $\max(-(\nu+\frac{1}{2}), \nu+\frac{3}{2}) < \mu < 1$. Then for almost all $x > 0$*

$$f(x) = x^{-(\nu+1/2)} \frac{d}{dx} x^{\nu+1/2} \int_0^\infty (xt)^{1/2} Y_{\nu+1}(xt)(\mathcal{H}_\nu f)(t) \frac{dt}{t}.$$

Proof. Note that the conditions imply that $-\frac{3}{2} < \nu < -\frac{1}{2}$, and thus $\nu + \frac{5}{2} > 1$, so that $\mu < \nu + \frac{5}{2}$. Fix $x > 0$, and let $g(t) = t^{-1/2}Y_{\nu+1}(xt)$. By [1, eqs. 7.2.1(4) and 7.13.1(4)], $g(t) = O(t^{\nu+1/2}) + O(t^{-\nu-3/2})$ as $t \to 0+$, with a modification when $\nu = -1$, and $g(t) = O(t^{-1})$ as $t \to \infty$. The conditions on μ thus ensure that $g \in \mathcal{L}_{\mu,p'}$, and [9, Thm. 5.3], in which there is a misprint for $\mu \geq \max(\gamma(p), \gamma(q))$, yields

$$\int_0^\infty f(t)(\mathcal{H}_\nu g)(t)\,dt = \int_0^\infty g(t)(\mathcal{H}_\nu f)(t)\,dt \tag{2.1}$$

since $\nu + \frac{3}{2} < \mu < \nu + \frac{5}{2}$.

But then [2, eq. 11.3(2)] gives $\mathcal{H}_\nu g$ since $-\frac{3}{2} < \nu < -\frac{1}{2}$, and by putting this value into (2.1) we have

$$\begin{aligned} -\int_x^\infty t^{\nu+1/2} f(t)\,dt &= x^{\nu+1}\int_0^\infty t^{-1/2}Y_{\nu+1}(xt)(\mathcal{H}_\nu f)(t)\,dt \\ &= x^{\nu+1/2}\int_0^\infty (xt)^{1/2}Y_{\nu+1}(xt)(\mathcal{H}_\nu f)(t)\frac{dt}{t}, \end{aligned} \tag{2.2}$$

and the result follows on differentiation. □

For the inversion when $\mu < -(\nu + \frac{1}{2})$, we first need a lemma.

LEMMA 2.2. *If* $\frac{1}{2} \leq \mu < \min(1, -(\nu + \frac{1}{2}), \nu + \frac{5}{2})$, *then*

$$\begin{aligned} &\lim_{R\to\infty} \frac{1}{2\pi i}\int_{\mu-iR}^{\mu+iR} u^{-s} m_\nu(s) \tan\frac{\pi}{2}(s-\nu-\tfrac{1}{2})/(\nu+\tfrac{3}{2}-s)ds \\ &\quad = u^{-1/2}(Y_{\nu+1}(u) - A_\nu(u/2)^{\nu+1}). \end{aligned}$$

Proof. By integrating around a rectangle with vertices at $\mu \pm iR$ and $\nu + \frac{1}{2} - 2n \pm iR$ and letting R and $n \to \infty$, the limit in the preceding is equal to the sum of the residues of the integrand to the left of the line $\operatorname{Re} s = \mu$. The integrand, fully written out, is

$$\left(u^{-s}2^{s-1/2}\Gamma\left(\left(\nu+s+\tfrac{1}{2}\right)/2\right)\tan\tfrac{\pi}{2}\left(s-\nu-\tfrac{1}{2}\right)\right) / \left(\Gamma\left(\left(\nu+\tfrac{3}{2}-s\right)/2\right)\left(\nu+\tfrac{3}{2}-s\right)\right)$$

$$= \left(u^{-s}2^{s-3/2}\Gamma\left(\left(\nu+s+\tfrac{1}{2}\right)/2\right)\tan\tfrac{\pi}{2}\left(s-\nu-\tfrac{1}{2}\right)\right) / \Gamma\left(\left(\nu+\tfrac{7}{2}-s\right)/2\right).$$

This has two sets of simple poles: at $s_n = -\left(\nu + \frac{1}{2}\right) - 2n$ and at $t_n = \nu + \frac{3}{2} - 2n$, where $n = 0, 1, \ldots$. Note that it follows from the conditions on μ that $-2 < \nu < -1$, from which it follows that the different sets of poles have no members in common. Also, from the hypotheses, $s_0 > \mu$ but $s_n < \mu$, $n = 1, 2, \ldots$, while $t_n < \mu$, $n = 0, 1, \ldots$. The residue at s_n is equal to $u^{-1/2}\cot\pi\nu(-1)^n(u/2)^{\nu+1+2n}/(\Gamma(\nu+n+2)n!)$, and the residue at t_n is equal to $u^{-1/2}\csc\pi\nu(-1)^n(u/2)^{-(\nu+1)+2n}/(\Gamma(n-\nu)n!)$. Hence

$$\begin{aligned} &\lim_{R\to\infty} \frac{1}{2\pi i}\int_{\mu-iR}^{\mu+iR} u^{-s} m_\nu(s) \tan\tfrac{\pi}{2}\left(s-\nu-\tfrac{1}{2}\right) / \left(\nu+\tfrac{3}{2}-s\right)ds \\ &= u^{-1/2}\left(\cot\pi\nu\sum_{n=1}^{\infty}(-1)^n(u/2)^{\nu+1+2n}/(\Gamma(\nu+n+2)n!)\right. \\ &\quad \left. + \csc\pi\nu\sum_{n=0}^{\infty}(-1)^n(u/2)^{-(\nu+1)+2n}/(\Gamma(n-\nu)n!)\right) \\ &= u^{-1/2}(\cot\pi(\nu+1)J_{\nu+1}(u) - A_\nu(u/2)^{\nu+1} - \csc\pi(\nu+1)J_{-(\nu+1)}(u)) \\ &= u^{-1/2}(Y_{\nu+1}(u) - A_\nu(u/2)^{\nu+1}). \qquad \square \end{aligned}$$

THEOREM 2.3. *Suppose that* $f \in \mathcal{L}_{\mu,p}$, *where* $1 < p < \infty$, $\mu \geq \gamma(p)$, *and* $\mu < \min\left(1, -\left(\nu + \frac{1}{2}\right), \nu + \frac{5}{2}\right)$. *Then for almost all* $x > 0$

$$f(x) = x^{-\nu-1/2}\frac{d}{dx}x^{\nu+1/2}\int_0^\infty (xt)^{1/2}\{Y_{\nu+1}(xt) - A_\nu(xt/2)^{\nu+1}\}(\mathcal{H}_\nu f)(t)\frac{dt}{t}.$$

Proof. Note that since $-(\nu + \frac{1}{2}) > \mu \geq \gamma(p) \geq \frac{1}{2}$, $\nu < -1$ and, similarly, $\nu > -2$.

Let $h(u) = u^{-1/2}(Y_{\nu+1}(u) - A_\nu(u/2)^{\nu+1})$. We shall prove that, under the conditions of the theorem,

$$-\int_x^\infty t^{\nu+1/2}f(t)\,dt = x^{\nu+3/2}\int_0^\infty h_\nu(xt)(\mathcal{H}_\nu f)(t)\,dt, \tag{2.3}$$

which on differentiating yields the desired result, and we prove this first on $\mathcal{L}_{1/2,2} = L_2(0,\infty)$.

Suppose then that $f \in \mathcal{L}_{1/2,2}$. Note that if

$$H(s) = m_\nu(s)\tan\left(\frac{\pi}{2}\right)\left(s - \nu - \frac{1}{2}\right)\Big/\left(\nu + \frac{3}{2} - s\right),$$

then $H(\frac{1}{2} + it) \in L_2(\mathbb{R})$, and hence by [12, Thm. 71] and Lemma 2.2, $h \in \mathcal{L}_{1/2,2}$ and $(\mathcal{M}h)(s) = H(s)$. Thus if $h_x(t) = h(xt)$ for $x > 0$ and $h_x \in \mathcal{L}_{1/2,2}$, from elementary considerations $(\mathcal{M}h_x)(s) = x^{-s}H(s)$. Also, from [9], $\mathcal{H}_\nu f \in \mathcal{L}_{1/2,2}$. Hence from [12, Thm. 72]

$$\begin{aligned}
x^{\nu+3/2}\int_0^\infty h(xt)(\mathcal{H}_\nu f)(t)\,dt &= \frac{x^{\nu+3/2}}{2\pi i}\int_{1/2-i\infty}^{1/2+i\infty}(\mathcal{M}h_x)(s)(\mathcal{M}\mathcal{H}_\nu f)(1-s)\,ds \\
&= \frac{x^{\nu+3/2}}{2\pi i}\int_{1/2-i\infty}^{1/2+i\infty} x^{-s}H(s)m_\nu(1-s)\tan\frac{\pi}{2}\left(\nu + \frac{3}{2} - s\right)(\mathcal{M}f)(s)\,ds \\
&= \frac{x^{\nu+3/2}}{2\pi i}\int_{1/2-i\infty}^{1/2+i\infty} x^{-s}(\mathcal{M}f)(s)/\left(\nu + \frac{3}{2} - s\right)ds,
\end{aligned}$$

since $m_\nu(s)m_\nu(1-s) = 1$ and $\tan\frac{\pi}{2}\left(\nu + \frac{3}{2} - s\right) = \cot\frac{\pi}{2}\left(s - \nu - \frac{1}{2}\right)$. Also, if for $x > 0$, $k_x(t) = 0$, $0 < t < x$, and $k_x(t) = t^{\nu+1/2}$, $t > x$, then clearly $k \in \mathcal{L}_{1/2,2}$ and $(\mathcal{M}k_x)(s) = x^{s+\nu+1/2}/(s + \nu + \frac{1}{2})$. Hence from [12, Thm. 72]

$$\begin{aligned}
-\int_x^\infty t^{\nu+1/2}f(t)\,dt &= \frac{1}{2\pi i}\int_{1/2-i\infty}^{1/2+i\infty}(\mathcal{M}k_x)(1-s)(\mathcal{M}f)(s)\,ds \\
&= \frac{x^{\nu+3/2}}{2\pi i}\int_{1/2-i\infty}^{1/2+i\infty} x^{-s}(\mathcal{M}f)(s)/(\nu + \tfrac{3}{2} - s)ds,
\end{aligned}$$

and thus (2.3) holds when $p = 2$ and $\mu = \frac{1}{2}$.

However, under the conditions of the theorem, for each $x > 0$ both sides of (2.3) represent bounded linear functionals on $\mathcal{L}_{\mu,p}$. Clearly, $k_x \in \mathcal{L}_{1-\mu,p'}$ since $\nu + \frac{3}{2} < \frac{1}{2}$ and thus the left-hand side of (2.3) represents such a functional. Also, $h(u) = O(u^{\nu+5/2}) + O^{-(\nu+3/2)})$ as $u \to 0^+$, and $h(u) = O(u^{-1}) + O(u^{\nu+1/2})$ as $u \to \infty$, and thus $h \in \mathcal{L}_{\mu,p'}$. Hence

$$\left|\int_0^\infty h(xt)(\mathcal{H}_\nu f)(t)\,dt\right| \le \left[\int_0^\infty |t^\mu h(xt)|^{p'}\frac{dt}{t}\right]^{1/p'} \|\mathcal{H}_\nu f\|_{1-\mu,p}$$

$$\le Kx^{-\mu}\|h\|_{\mu,p'}\|f\|_{\mu,p},$$

where K is a bound for the norm of $\mathcal{H}_\nu$ and thus the right-hand side of (2.2) also represents a bounded linear functional on $\mathcal{L}_{\mu,p}$. Hence (2.3) holds for the values of μ and ν stated in the theorem, and the theorem is proved. □

The reader should note that $\mathbb{H}_{-3/2}(z) = -J_{3/2}(z)$, so that $\mathcal{H}_{-3/2}$ is just minus the Hankel transformation of order $\frac{3}{2}$, and that then, since $Y_{-1/2}(z) = J_{1/2}(z)$ and $A_{-3/2} = 0$, this gives another inversion of the Hankel transformation of order $\frac{3}{2}$.

3. Range of $\mathcal{H}\nu$ when $\mu = -(\nu + \frac{1}{2})$. In this section we discuss the range of $\mathcal{H}_\nu$ when $\mu = -(\nu + \frac{1}{2})$. To this end we use the properties of the even and odd Hilbert transformations H_+ and H_-, which are summarized in [9, §3], and of the Hankel transformation H_ν, $\nu > -1$, and the extended Hankel transformation H_ν, $\nu < -1$, studied in [4], [5], [7]–[9]; also, M_α, where $\alpha \in \mathbb{R}$, denotes the operator on complex-valued functions on $(0,\infty)$ defined by

$$(M_\alpha f)(t) = t^\alpha f(t).$$

Clearly, M_α is an isometric isomorphism of $\mathcal{L}_{\mu,p}$ onto $\mathcal{L}_{\mu-\alpha,p}$.

In each of the lemmas and theorems of this section we assume that $\gamma(p) \le \mu = -(\nu + \frac{1}{2}) < \nu + \frac{5}{2}$, which gives $\nu > -\frac{3}{2}$. It should be noted that this implies that $\nu \le -1$, with $\nu = -1$ only if $\gamma(p) = \frac{1}{2}$, and thus $p = 2$. Also, it implies that $\nu + \frac{1}{2} \le -\frac{1}{2} < \gamma(p) \le \mu$, and thus μ is in the range for boundedness of $\mathcal{H}_\nu$ on $\mathcal{L}_{\mu,p}$. It further implies that $p > 1$, for since $\nu > -\frac{3}{2}$, $-(\nu + \frac{1}{2}) < 1$, and thus $\gamma(p) < 1$ and hence $p > 1$.

We begin by proving a lemma followed by a theorem that shows immediately that the case in which $\mu = -(\nu + \frac{1}{2})$ is fundamentally different from those cases considered in §2 and from which we immediately obtain, as a corollary, that $\mathcal{H}_\nu(\mathcal{L}_{\mu,p})$ is a proper subset of $H_{\nu+1}(\mathcal{L}_{\mu,p})$.

LEMMA 3.1. *Suppose* $1 < p < \infty$ *and* $\gamma(p) \le \mu = -(\nu + \frac{1}{2}) < \nu + \frac{5}{2}$. *Then on* $\mathcal{L}_{\mu,p}$

$$\mathcal{H}_\nu = H_\nu M_{-(\nu+1/2)} H_- M_{\nu+1/2} \tag{3.1}$$

if $-\frac{3}{2} < \nu < -1$, *and*

$$\mathcal{H}_{-1} = M_{-1/2} H_+ M_{1/2} H_1. \tag{3.2}$$

Proof. Note that $-\frac{3}{2} < \nu \le -1$. It is sufficient to prove the relations when $p = 2$ since $\mathcal{L}_{\mu,2}$ is dense in $\mathcal{L}_{\mu,p}$, $\gamma(p) \ge \gamma(2)$, and each side of (3.1) and (3.2) is a bounded linear transformation from $\mathcal{L}_{\mu,p}$ to $\mathcal{L}_{1-\mu,p}$. From [9, Thm. 5.2] the left-hand side of (3.1) is so bounded. Also, as noted, $M_{\nu+1/2}$ maps $\mathcal{L}_{\mu,p}$ boundedly onto $\mathcal{L}_{\mu-(\nu+1/2),p}$, and from [9, Thm. 3.1], H_- maps $\mathcal{L}_{\mu-(\nu+1/2),p}$ boundedly into itself since $1 \le \mu - (\nu + \frac{1}{2}) < 2$. Further, $M_{-(\nu+1/2)}$ maps $\mathcal{L}_{\mu-(\nu+1/2),p}$ boundedly onto $\mathcal{L}_{\mu,p}$, and finally, from [8, Thm. 1], H_ν maps $\mathcal{L}_{\mu,p}$ boundedly into $\mathcal{L}_{1-\mu,p}$ since the integer m of that theorem is one and our hypotheses ensure that $\gamma(p) \le \mu < \nu + \frac{7}{2}$.

Then by [8, Thm. 1] and [9, eqs. 1.10, 2.3, and Thms. 3.1 and 5.2], if $f \in \mathcal{L}_{\mu,2}$ and Re $s = \frac{1}{2}$, then

$$\begin{aligned}(\mathcal{M}H_\nu M_{-(\nu+1/2)}H_- M_{\nu+1/2}f)(s) &= m_\nu(s)(\mathcal{M}M_{-(\nu+1/2)}H_- M_{\nu+1/2}f)(1-s)\\ &= m_\nu(s)(\mathcal{M}H_- M_{\nu+1/2}f)(1-s-(\nu+\tfrac{1}{2}))\\ &= m_\nu(s)\cot\tfrac{\pi}{2}(1-s-(\nu+\tfrac{1}{2}))(\mathcal{M}M_{\nu+1/2}f)(1-s-(\nu+\tfrac{1}{2}))\\ &= m_\nu(s)\cot\tfrac{\pi}{2}(1-s-(\nu+\tfrac{1}{2}))(\mathcal{M}f)(1-s)\\ &= m_\nu(s)\tan\tfrac{\pi}{2}(s+(\nu+\tfrac{1}{2}))(\mathcal{M}f)(1-s)\\ &= (\mathcal{M}\mathcal{H}_\nu f)(s).\end{aligned}$$

Hence (3.1) holds on $\mathcal{L}_{\mu,2}$ and thus on $\mathcal{L}_{\mu,p}$ for the range of values shown. The proof of (3.2) is similar. □

THEOREM 3.2. *Suppose $f \in \mathcal{L}_{\mu,p}$, where $1 < p < \infty$, $\gamma(p) \le \mu = -(\nu+\frac{1}{2}) < \nu+\frac{5}{2}$. Then the integral*

$$\int_{\to 0}^{\to\infty} t^{\nu+1/2}(\mathcal{H}_\nu f)(t)\,dt$$

converges and equals zero.

Proof. If $\nu < -1$, then from (3.1), $\mathcal{H}_\nu f = H_\nu\phi$, where $\phi = M_{-(\nu+1/2)}H_- M_{\nu+1/2}f$. As noted in the proof of Lemma 3.1, $M_{-(\nu+1/2)}H_- M_{\nu+1/2} \in [\mathcal{L}_{\mu,p}]$, and thus $\mathcal{H}_\nu f \in H_\nu(\mathcal{L}_{\mu,p})$. But from [5, Thm. 6.3], with $l = 1$,

$$\int_{\to 0}^{\to\infty} t^{\nu+1/2}(H_\nu\phi)(t)\,dt$$

converges and equals zero, which is the result to be proved in this case.

If $\nu = -1$, then $\mu = \frac{1}{2}$, so that $p = 2$. But then from (3.2), if $f \in \mathcal{L}_{1/2,2} = L_2(0,\infty)$, $\mathcal{H}_{-1}f = M_{-1/2}H_+\psi$, where $\psi = M_{1/2}H_1 f$. Now by [12, Chap. 8], $H_1 \in [L_2(0,\infty)]$ and thus $\psi \in \mathcal{L}_{0,2}$, and thus by [3, Cor. 4.3]

$$\int_{\to 0}^{\to\infty} (H_+\psi)(t)\frac{dt}{t}$$

converges and equals zero, which is again the result to be proved since $(H_+\psi)(t)/t = t^{-1/2}(\mathcal{H}_{-1}f)(t)$. □

COROLLARY 3.3. *Suppose $1 < p < \infty$, $\gamma(p) \le \mu = -(\nu+\frac{1}{2}) < \nu+\frac{5}{2}$. Then $\mathcal{H}_\nu(\mathcal{L}_{\mu,p})$ is a proper subset of $H_{\nu+1}(\mathcal{L}_{\mu,p})$.*

Proof. In [9, Thm. 5.1] $\mathcal{H}_\nu$ is defined to be $H_{\nu+1}S_\nu$, where S_ν is a transformation that, for the values of μ and ν for which $\mathcal{H}_\nu$ is bounded, maps $\mathcal{L}_{\mu,p}$ boundedly onto itself. Hence, $\mathcal{H}_\nu(\mathcal{L}_{\mu,p}) \subseteq H_{\nu+1}(\mathcal{L}_{\mu,p})$. Let $f(x) = x^{\nu+3/2}/(x^2+1)$. Then $f \in \mathcal{L}_{\mu,p}$, and from [2; eq. 8.5(12)], $(H_{\nu+1}f)(x) = x^{1/2}K_{\nu+1}(x)$. From Theorem 3.2, if $\mathcal{H}_\nu(\mathcal{L}_{\mu,p}) = H_{\nu+1}(\mathcal{L}_{\mu,p})$, then

$$\int_{\to 0}^{\to\infty} t^{\nu+1}K_{\nu+1}(x)\,dx = 0,$$

which is impossible since the integrand is positive! □

Since $\mathcal{H}_\nu(\mathcal{L}_{\mu,p})$ is a proper subset of $H_{\nu+1}(\mathcal{L}_{\mu,p})$, it is of interest to find exactly what subset it is, and the following theorem characterizes it.

THEOREM 3.4. *Suppose* $1 < p < \infty$ *and* $\gamma(p) \le \mu = -(\nu + \frac{1}{2}) < \nu + \frac{5}{2}$. *Then* $g \in \mathcal{H}_\nu(\mathcal{L}_{\mu,p})$ *if and only if*

(a) $g \in H_{\nu+1}(\mathcal{L}_{\mu,p})$,

(b) $\int_{\to 0}^{1} t^{\nu+1/2} g(t)\, dt$ *converges, and*

(c) $\varphi \in H_{\nu+1}(\mathcal{L}_{\mu,p})$, *where* $\varphi(x) = x^{-\nu+1/2} \int_{\to 0}^{x} t^{\nu+1/2} g(t)\, dt$.

Proof. Suppose first that $-\frac{3}{2} < \nu < -1$. If $g \in \mathcal{H}_\nu(\mathcal{L}_{\mu,p})$, then $g = \mathcal{H}_\nu f$, some $f \in \mathcal{L}_{\mu,p}$, and thus from (3.1), $g = H_\nu \psi$, where $\psi = M_\mu H_- M_{-\mu} f$. Now $M_{-\mu} f \in \mathcal{L}_{2\mu,p}$, and thus since $1 < 2\mu < 2$, from [9, Thm. 3.1(b)], $H_- M_{-\mu} f \in \mathcal{L}_{2\mu,p}$, and thus $\psi \in \mathcal{L}_{\mu,p}$. Hence $g \in H_\nu(\mathcal{L}_{\mu,p})$, and it follows from [5, Thm. 6.5], since in this case $l = l_\nu = 1$, that $g \in H_{\nu+2}(\mathcal{L}_{\mu,p})$, (b) holds, and $\varphi \in H_{\nu+2}(\mathcal{L}_{\mu,p})$. But from [10, Thm. 1], since $\nu > -\frac{3}{2}$, $H_{\nu+2}(\mathcal{L}_{\mu,p}) = H_{\nu+1}(\mathcal{L}_{\mu,p})$, and thus (a) and (c) also hold.

Conversely, if g satisfies (a), (b), and (c), then, since from [10, Thm. 1], $H_{\nu+1}(\mathcal{L}_{\mu,p}) = H_{\nu+2}(\mathcal{L}_{\mu,p})$ and since $l_\nu = 1$, g satisfies (a), (b), and (c) of [5, Thm. 6.5], and thus there is a $\psi \in \mathcal{L}_{\mu,p}$, so that $g = H_\nu \psi$. Since from [9, Thm. 3.1(b)], $H_- \in [\mathcal{L}_{2\mu,p}]$, and maps $\mathcal{L}_{2\mu,p}$ one-to-one onto itself, from [11, Thm. 42.-H], H_-^{-1} exists and $H_-^{-1} \in [\mathcal{L}_{2\mu,p}]$. Let $f = M_\mu H_-^{-1} M_{-\mu} \psi$. Then $f \in \mathcal{L}_{\mu,p}$ and $\psi = M_\mu H_- M_{-\mu} f$, and thus, by using (3.1), $g = H_\nu M_\mu H_- M_{-\mu} f = \mathcal{H}_\mu f$, so that $g \in \mathcal{H}_\nu(\mathcal{L}_{\mu,p})$.

Suppose next that $\nu = -1$. Then $\mu = \frac{1}{2}$ and $p = 2$, so that $\mathcal{L}_{\mu,p} = \mathcal{L}_{1/2,2}$. Suppose $g \in \mathcal{H}_{-1}(\mathcal{L}_{1/2,2})$. Then from (3.2), $M_{1/2} g = H_+ M_{1/2} H_1 f = H_+ \psi$, where $\psi = M_{1/2} H_1 f$. But from [6, Thm. 1], since $\mathcal{L}_{1/2,2} = L^2(0,\infty)$, $H_1 f \in \mathcal{L}_{1/2,2}$ and thus $\psi \in \mathcal{L}_{0,2}$. But then, from [3, Thm. 4.4], $M_{1/2} g \in \mathcal{L}_{0,2}$, $\int_{\to 0}^{1} (M_{1/2} g)(t)\, dt/t = \int_{\to 0}^{1} t^{-1/2} g(t)\, dt$ converges, that is, (b) holds, and $k \in \mathcal{L}_{0,2}$, where

$$k(x) = \int_{\to 0}^{x} (M_{1/2} g)(t)\, dt/t = \int_{\to 0}^{x} t^{-1/2} g(t)\, dt.$$

Hence $g \in \mathcal{L}_{1/2,2}$, and since, from [6], $H_0(\mathcal{L}_{1/2,2}) = \mathcal{L}_{1/2.2}$, (a) holds. Also, $\varphi = M_{-1/2} k$, so that $\varphi \in \mathcal{L}_{1/2,2} = H_0(\mathcal{L}_{1/2,2})$ and (c) holds.

Conversely, suppose g satisfies (a), (b), and (c). Then, since $H_0(\mathcal{L}_{1/2,2}) = \mathcal{L}_{1/2,2}$, $M_{1/2} g$ satisfies (i), (ii), and (iii) of [3, Thm. 4.4] with $p = 2$ and thus $M_{1/2} g = H_+ \psi$, where $\psi \in \mathcal{L}_{0,2}$. But then $M_{-1/2} \psi \in \mathcal{L}_{1/2,2} = H_1(\mathcal{L}_{1/2,2})$, and thus there is an $f \in \mathcal{L}_{1/2,2}$, so that $M_{-1/2} \psi = H_1 f$. It follows then that $g = M_{-1/2} H_+ \psi = M_{-1/2} H_+ M_{1/2} H_1 f = \mathcal{H}_{-1} f$ and $g \in \mathcal{H}_{-1}(\mathcal{L}_{1/2,2})$. □

4. Inversion when $\mu = -(\nu + \frac{1}{2})$, $\mu \ne \frac{1}{2}$. Theorem 4.2, which follows, and its corollary give two inversion formulas for $\mathcal{H}_\nu$ in this case. However, we first need a lemma.

LEMMA 4.1. *Suppose that* $f \in \mathcal{L}_{\mu,p}$, *where* $1 < p < \infty$, $-\frac{3}{2} < \nu < -1$, *and* $\mu = -(\nu + \frac{1}{2})$. *Then as* $a \to 0_+$

$$a^{\nu+1} \int_1^\infty t^{-1/2} f(t) \mathbb{H}_{\nu+1}(at)\, dt \to 0. \tag{4.1}$$

Proof. Given $\epsilon > 0$, choose R so large that

$$\left\{ \int_R^\infty |t^u f(t)|^p \frac{dt}{t} \right\}^{1/p} < \epsilon.$$

Since $\mathbb{H}_{\nu+1}(t) \sim t^{\nu+2}/(2^{\nu+1}\pi^{1/2}\Gamma(\nu+\frac{5}{2}))$ as $t \to 0^+$, there is a constant K_R such that $|\mathbb{H}_{\nu+1}(t)| \le K_R t^{\nu+2}$ for $0 \le t \le R$. If $0 < a < 1$, we then have

$$\left| a^{\nu+1} \int_1^R t^{-1/2} \mathbb{H}_{\nu+1}(at) f(t)\, dt \right| \le K_R a^{2\nu+3} \int_1^R t^{\nu+3/2} |f(t)|\, dt.$$

Moreover, by using Hölder's inequality, since $\mu = -(\nu + \frac{1}{2})$,

$$\begin{aligned}
&\left| a^{\nu+1} \int_R^\infty t^{-1/2} \mathbb{H}_{\nu+1}(at) f(t)\, dt \right| \\
&\quad \le a^{\nu+1} \left\{ \int_R^\infty |t^\mu f(t)|^p \frac{dt}{t} \right\}^{1/p} \left\{ \int_R^\infty |t^{1/2-\mu} \mathbb{H}_{\nu+1}(at)|^{p'} \frac{dt}{t} \right\}^{1/p'} \\
&\quad < \epsilon a^{\nu+1} \left\{ \int_{aR}^\infty \left| \left(\frac{u}{a}\right)^{\nu+1} \mathbb{H}_{\nu+1}(u) \right|^{p'} \frac{du}{u} \right\}^{1/p'} \\
&\quad \le \epsilon \left\{ \int_0^\infty |y^{\nu+1} \mathbb{H}_{\nu+1}(u)|^{p'} \frac{du}{u} \right\}^{1/p'} = A\epsilon.
\end{aligned}$$

Here A is finite since $u^{\nu+1}\mathbb{H}_{\nu+1}(u) = O(u^{2\nu+3})$ as $u \to 0^+$ and since $\mathbb{H}_{\nu+1}(u) = O(u^{-1/2}) + O(u^\nu) = O(u^{-1/2})$ as $u \to \infty$, so that $u^{\nu+1}\mathbb{H}_{\nu+1}(u) = O(u^{\nu+1/2})$ as $u \to \infty$.

Combining these estimates, we find that

$$\left| a^{\nu+1} \int_1^\infty t^{-1/2} \mathbb{H}_{\nu+1}(at) f(t)\, dt \right| \le A\epsilon + K_R a^{2\nu+3} \int_1^R t^{\nu+3/2} |f(t)|\, dt,$$

Hence as $a \to 0^+$ the upper limit of the left-hand side is less than $A\epsilon$, and since ϵ is arbitrary, (4.1) follows. □

We now can prove the following inversion theorem and its corollary.

THEOREM 4.2. *Suppose that $f \in \mathcal{L}_{\mu,p}$, where $1 < p < \infty$, $\mu = -(\nu + \frac{1}{2}) \ge \gamma(p)$, and $\nu + \frac{3}{2} < \mu < 1$. Then for almost all $x > 0$*

$$f(x) = x^{-\nu-1/2} \frac{d}{dx} x^{\nu+1/2} \int_{\to 0}^\infty (xt)^{1/2} Y_{\nu+1}(xt) (\mathcal{H}_\nu f)(t) \frac{dt}{t}. \tag{4.2}$$

Proof. It is easy to see that $\mu < \nu + \frac{5}{2}$, so that $\mathcal{H}_\nu f$ exists. Suppose first that $\mu > \gamma(p)$, and choose $\epsilon > 0$ but sufficiently small that $\mu - \epsilon > \max(\gamma(p), \nu + \frac{3}{2})$ and $\mu + \epsilon < \min(1, \nu + 5/2)$. Define f_1 and f_2 by $f_1 = f \cdot \chi_{(1,\infty)}$ and $f_2 = f - f_1$. Then $f_2 \in \mathcal{L}_{\mu+\epsilon,p}$ and $f_1 \in \mathcal{L}_{\mu-\epsilon,p}$, and thus we can apply Theorems 2.1 and 2.3, respectively, and it follows that for almost all $x > 0$

$$f_2(x) = x^{-\nu-1/2} \frac{d}{dx} x^{\nu+1/2} \int_0^\infty (xt)^{1/2} Y_{\nu+1}(xt) (\mathcal{H}_\nu f_2)(t) \frac{dt}{t} \tag{4.3}$$

and

$$f_1(x) = x^{-\nu-1/2} \frac{d}{dx} x^{\nu+1/2} \int_0^\infty (xt)^{1/2} \{ Y_{\nu+1}(xt) - A_\nu (xt/2)^{\nu+1} \} \cdot (\mathcal{H}_\nu f_1)(t) \frac{dt}{t}.$$

In view of Theorem 3.2, we can write this last formula as

$$f_1(x) = x^{-\nu-1/2}\frac{d}{dx}x^{\nu+1/2}\int_{\to 0}^{\to\infty}(xt)^{1/2}Y_{\nu+1}(xt)(\mathcal{H}_\nu f_1)(t)\frac{dt}{t}. \tag{4.4}$$

But for fixed $x > 0$, $(xt)^{1/2}Y_{\nu+1}(xt)$ is bounded for $t \geq 1$, and Hölder's inequality gives

$$\int_1^\infty |(\mathcal{H}_\nu f_1)(t)|\frac{dt}{t} \leq \left[\int_1^\infty |t^{1-\mu}(\mathcal{H}_\nu f_1)(t)|^p\frac{dt}{t}\right]^{1/p}\left[\int_1^\infty t^{(\mu-1)p'}\frac{dt}{t}\right]^{1/p'},$$

and the last two integrals are finite since $\mathcal{H}_\nu f_1 \in \mathcal{L}_{\mu,p}$ and $\mu < 1$. Thus we do not need the arrow on the upper limit of integration in (4.4). Then, adding (4.3) and the modified (4.4), we have the result.

If $\mu = \gamma(p)$, (4.3) still holds but (4.4) does not since now $\mu - \epsilon < \gamma(p)$. Now we apply (2.1) to f_1 and g_x, where

$$g_x(t) = t^{-1/2}\left[Y_{\nu+1}(xt) - A_\nu\left(\frac{xt}{2}\right)^{\nu+1}\chi_{(0,a)}(t)\right],$$

x and a being fixed positive numbers. As in the proof of Theorem 2.3, $g_x(t) = O(t^{\nu+5/2}) + O(t^{-\nu-3/2})$ as $t \to 0^+$ and $g_x(t) = O(1/t)$ as $t \to \infty$, and so the conditions that ensure that $g_x \in \mathcal{L}_{\mu,p'}$ are $\mu + \nu + \frac{5}{2} > 0$, $\mu < \nu + \frac{3}{2}$, and $\mu < 1$. The second and third of these are among our hypotheses, while $\mu = -(\nu + \frac{1}{2})$ yields $\mu + \nu + \frac{5}{2} = 2$. Thus (2.1) yields

$$\int_0^\infty (\mathcal{H}_\nu f_1)(t)g_x(t)\,dt = \int_0^\infty f_1(t)(\mathcal{H}_\nu g_x)(t)\,dt. \tag{4.5}$$

In view of Theorem 3.2, we can write the left-hand side of (4.5) in the form

$$\int_{\to 0}^\infty t^{-1/2}Y_{\nu+1}(xt)(\mathcal{H}_\nu f_1)(t)\,dt - A_\nu\left(\frac{x}{2}\right)^{\nu+1}\int_{\to 0}^a t^{\nu+1/2}(\mathcal{H}_\nu f_1)(t)\,dt,$$

and the second term tends to zero as $a \to 0^+$. We can write the right-hand side of (4.5) in the form

$$\int_0^\infty f_1(t)(\mathcal{H}_\nu h_x)(t)\,dt - A_\nu\left(\frac{x}{2}\right)^{\nu+1}\int_0^\infty f_1(t)(\mathcal{H}_\nu k_a)(t)\,dt, \tag{4.6}$$

where $h_x(t) = t^{-1/2}Y_{\nu+1}(xt)$ and $k_a(t) = t^{\nu+1/2}\chi_{(0,a)}(t)$, provided we show that one of the two integrals exists. From our hypotheses, $-\frac{3}{2} < \nu < -1$. Hence from [2, eq. 11.3(2)] the first integral exists and equals

$$-x^{-\nu-1}\int_x^\infty t^{\nu+1/2}f_1(t)\,dt,$$

this integral existing since Hölder's inequality gives

$$\int_x^\infty t^{\nu+1/2}|f_1(t)|\,dt \leq \left[\int_x^\infty |t^\mu f_1(t)|^p\frac{dt}{t}\right]^{1/p}\left[\int_x^\infty t^{(\nu-\mu+3/2)p'}\frac{dt}{t}\right]^{1/p'} < \infty$$

since $\mu > \nu + \frac{3}{2}$.

By [2, eq. 11.2(2)], $(\mathcal{H}_\nu k)(t) = t^{-\nu-3/2}(at)^{\nu+1}\mathbb{H}_{\nu+1}(at)$ since $\nu > -\frac{3}{2}$, and hence the second term in (4.6) reduces to

$$A_\nu \left(\frac{ax}{2}\right)^{\nu+1} \int_1^\infty t^{-1/2} f_1(t)\mathbb{H}_{\nu+1}(at)\, dt,$$

and as $a \to 0^+$ this tends to zero by Lemma 4.1. Thus if we let a tend to zero in (4.5), we find that

$$\int_{\to 0}^\infty t^{-1/2} Y_{\nu+1}(xt)(\mathcal{H}_\nu f_1)(t)\, dt = -x^{-\nu-1} \int_x^\infty t^{\nu+1/2} f_1(t)\, dt,$$

which, on differentiation, yields

$$x^{\nu+1/2} f_1(x) = \frac{d}{dx} x^{\nu+1/2} \int_{\to 0}^\infty (xt)^{1/2} Y_{\nu+1}(xt)(\mathcal{H}_\nu f_1)(t) \frac{dt}{t}$$

almost everywhere on $(0, \infty)$. Multiplying this by $x^{-\nu-1/2}$ and then adding it to (4.3), we obtain (4.2). □

COROLLARY 4.3. *Under the same hypotheses,*

$$(4.7)\quad f(x) = x^{-\nu-1/2} \frac{d}{dx} x^{\nu+1/2} \int_0^{\to\infty} (xt)^{1/2} \left[Y_{\nu+1}(xt) - A_\nu \left(\frac{xt}{2}\right)^{\nu+1} \right] (\mathcal{H}_\nu f)(t) \frac{dt}{t}$$

almost everywhere on $(0, \infty)$.

Proof. Theorem 3.2 shows that the integrals on the right of (4.2) and (4.7) are equal, and it is easy to see that the integral in (4.7) does not need an arrow at its lower limit. □

5. Inversion of $\mathcal{H}_{-1}$ on $L^2(0, \infty)$. We now consider the problem of inverting $\mathcal{H}_\nu$ on $\mathcal{L}_{\mu,p}$ in the case for which $\mu = \nu + \frac{3}{2}$ and for which the singularity condition, $\mu = -(\nu + \frac{1}{2})$, applies. For the two conditions to hold simultaneously we must have $\nu = -1$ and $\mu = \frac{1}{2}$; thus $\gamma(p) = \frac{1}{2}$, so that $p = 2$ and $\mathcal{L}_{\mu,p}$ is $\mathcal{L}_{1/2,2} = L^2(0, \infty)$. First we prove a lemma.

LEMMA 5.1. *If* $f \in L^2(0, \infty)$, *then as* $a \to \infty$

$$\int_0^\infty t^{-1/2} \mathbb{H}_0(at) f(t)\, dt \to 0.$$

Proof. Let K denote $\left\{\int_0^\infty t^{-1} |\mathbb{H}_0(t)|^2 dt\right\}^{1/2}$, which is finite since $\mathbb{H}_0(t) = O(t)$ as $t \to 0$ and $\mathbb{H}_0(t) = O(t^{-1/2})$ as $t \to \infty$. Given $\epsilon > 0$, choose δ so small that

$$\left[\int_0^\delta |f(t)|^2\, dt\right]^{1/2} < \epsilon.$$

Then, using Schwarz's inequality and the substitution $u = at$, we see that

$$\int_0^\delta |t^{-1/2}\mathbb{H}_0(at) f(t)|\, dt \le \left[\int_0^\infty t^{-1} |\mathbb{H}_0(at)|^2\, dt \cdot \int_0^\delta |f(t)|^2 dt\right]^{1/2} < K\epsilon,$$

while

$$\int_\delta^\infty |t^{-1/2}\mathbb{H}_0(at)f(t)|\,dt \le \|f\|_2 \cdot \left[\int_{a\delta}^\infty u^{-1}|\mathbb{H}_0(u)|^2\,du\right\}^{1/2}.$$

Hence

$$\left|\int_0^\infty t^{-1/2}\mathbb{H}_0(at)f(t)\,dt\right| < K\epsilon + \|f\|_2 \cdot \left[\int_{a\delta}^\infty u^{-1}|\mathbb{H}_0(u)|^2\,du\right]^{1/2},$$

from which it follows that

$$\overline{\lim_{a\to\infty}}\left|\int_0^\infty t^{-1/2}\mathbb{H}_0(at)f(t)\,dt\right| \le K\epsilon,$$

and the result follows. □

THEOREM 5.2. *If $f \in L^2(0,\infty)$, then for almost all $x > 0$*

$$f(x) = x^{1/2}\frac{d}{dx}\int_{\to 0}^\infty t^{-1/2}(Y_0(xt) - Y_0(t))(\mathcal{H}_{-1}f)(t)\,dt.$$

Proof. Let a and x be fixed positive numbers, and let

$$g_a(t) = t^{-1/2}(Y_0(xt) - Y_0(t) - 2\pi^{-1}\log x\chi_{(0,a)}(t)).$$

From [1, eq. 7.2.4(33)],

$$Y_0(xt) = 2\pi^{-1}\left(\gamma + \log\left(\frac{xt}{2}\right)\right) + O(t^2\log t)$$

as $t \to 0^+$, where γ is Euler's constant. Hence $g_a(t) = O(t^{3/2}\log t)$ as $t \to 0^+$; also, from [1, eq. 7.13.1(4)], $g_a(t) = O(t^{-1})$ as $t \to \infty$. Hence $g_a(t) \in L^2(0,\infty)$, and thus if $f \in L^2(0,\infty)$, from [9, eq. (5.9)]

$$\int_0^\infty f(t)(\mathcal{H}_{-1}g_a)(t)\,dt = \int_0^\infty g_a(t)(\mathcal{H}_{-1}f)(t)\,dt. \tag{5.1}$$

Now, from [2, eqs. 11.2(2) and 11.3(2)]

$$(\mathcal{H}_{-1}g_a)(y) = -y^{-1/2}(\chi_{(x,\infty)}(y) - \chi_{(1,\infty)}(y) + (2\pi^{-1}\log x)\mathbb{H}_0(ay))$$

$$= y^{-1/2}(\chi_{(1,x)}(y) - (2\pi^{-1}\log x)\mathbb{H}_0(ay)).$$

Each term of this last line is in $L^2(0,\infty)$, and thus we can write the left-hand side of (5.1) in the form

$$\int_1^x y^{-1/2}f(y)\,dy - 2\pi^{-1}\log x\int_0^\infty y^{-1/2}\mathbb{H}_0(ay)f(y)\,dy,$$

and we can write the right-hand side as

$$\int_{\to 0}^\infty t^{-1/2}(Y_0(xt) - Y_0(t))(\mathcal{H}_{-1}f)(t)\,dt - 2\pi^{-1}\log x\int_{\to 0}^a t^{-1/2}(\mathcal{H}_{-1}f)(t)\,dt$$

since the last integral exists by Theorem 3.2. If we now let $a \to \infty$ in (5.1) and use Lemma 5.1 and Theorem 3.2, it follows that

$$\int_1^x y^{-1/2} f(y)\, dy = \int_{\to 0}^{\infty} t^{-1/2}(Y_0(xt) - Y_0(t))(\mathcal{H}_{-1}f)(t)\, dt,$$

and the result follows on differentiating. □

Finally, we have the following corollary, whose proof is similar to that of Corollary 4.3.

COROLLARY 5.3. *If $f \in L^2(0,\infty)$, then for almost all $x > 0$*

$$f(x) = x^{1/2}\frac{d}{dx}\int_0^{\to\infty} t^{-1/2}(Y_0(xt) - Y_0(t) - 2\pi^{-1}\log x)(\mathcal{H}_{-1}f)(t)\, dt.$$

Acknowledgment. P. Heywood wishes to thank the University of Toronto for the provision of research facilities.

REFERENCES

[1] A ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. TRICOMI, *Higher Transcendental Functions* II, McGraw-Hill, New York, 1952.
[2] ———, *Tables of Integral Transforms* I & II, McGraw-Hill, New York, 1954.
[3] P. HEYWOOD AND P. G. ROONEY, *On the inversion of the even and odd Hilbert transformations*, Proc. Roy. Soc. Edinburgh Sect. A, 109 (1988), pp. 201–211.
[4] ———, *On the Hankel and some related transformations*, Canad. J. Math., 40 (1988), pp. 989–1009.
[5] ———, *On the inversion of the extended Hankel transformation*, J. Math. Anal. Appl., 160 (1991), pp. 284–302.
[6] H. KOBER, *Hankelsche Transformationen*, Quart. J. Math., 8 (1937), pp. 186–199.
[7] P. G. ROONEY, *A technique for studying the boundedness and extendability of certain types of operators*, Canad. J. Math., 25 (1973), pp. 1090–1102.
[8] ———, *On the boundedness and range of the extended Hankel transformation*, Canad. Math. Bull., 23 (1980), pp. 321–325.
[9] ———, *On the $\mathcal{Y}_\nu$ and $\mathcal{H}_\nu$ transformations*, Canad. J. Math., 32 (1980), pp. 1021–1044.
[10] ———, *On the range of the Hankel transformation*, Bul. London Math. Soc., 11 (1979), pp. 45–48.
[11] A. E. TAYLOR, *Functional Analysis*, John Wiley, New York, 1958.
[12] E. C. TITCHMARSH, *The Theory of Fourier Integrals*, Oxford University Press, Oxford, England, 1937.

SIAM J. MATH ANAL.
Vol. 25, No. 2, pp. 462–473, March 1994

015

ASYMPTOTICS OF POLLACZEK POLYNOMIALS AND THEIR ZEROS*

MOURAD E. H. ISMAIL†

Abstract. An asymptotic expression for the large n behavior of Pollaczek polynomials of degree n and argument $\cos(\theta/\sqrt{n})$ is derived. This is applied to find the second term in the asymptotic expansion of the zeros of the Pollaczek polynomials and it also solves a problem of Richard Askey. Some new analytic properties of a confluent Horn function are also proved.

Key words. Pollaczek polynomials, zeros, asymptotics, confluent Horn functions

AMS subject classifications. primary 33A65; secondary 30E15, 41A60

1. Introduction.

The Pollaczek polynomials are generated by [7, VI eq. (5.2)]

$$P_{-1}^{\lambda}(x;a,b) := 0, P_0^{\lambda}(x;a,b) := 1, \tag{1.1}$$

$$\begin{aligned}(n+1)P_{n+1}^{\lambda}(x;a,b) &= 2[x(a+\lambda+n)+b]P_n^{\lambda}(x;a,b)\\ &\quad -(n+2\lambda-1)P_{n-1}^{\lambda}(x;a,b), \qquad n \geq 0.\end{aligned} \tag{1.2}$$

The Pollaczek polynomials are orthogonal when either $\lambda > 0$ and $\lambda + a > 0$ hold or $0 > \lambda > -\frac{1}{2}$ and $0 < 1 + \lambda + a < 1$ are valid. The parameter b is assumed to be real. For a discussion of the nature of the orthogonality see [2] and [6]. The Pollaczek (or Pollaczek–Szegö) polynomials have the explicit representation (see [7, Chap. VI eq. (5.6)] or [10, §10.21])

$$P_n^{\lambda}(\cos\theta;a,b) = \frac{(2\lambda)_n}{n!}e^{in\theta}\,{}_2F_1(-n,\lambda+it(\theta);2\lambda;1-e^{-2i\theta}), \tag{1.3}$$

where

$$t(\theta) := (a\cos\theta + b)/\sin\theta. \tag{1.4}$$

When $a = b = 0$ the Pollaczek polynomials reduce to the familiar ultraspherical (or Gegenbauer) polynomials [21, §4.7], [20, Chap. 17]. Novikoff wrote a dissertation [17] on the Pollaczek polynomials, where he found the main term in the asymptotic development of their zeros. Let $\{x_{n,k}(\lambda,a,b) : 1 \leq k \leq n\}$ be the zeros of a Pollaczek polynomial $\{P_n^{\lambda}(x;a,b)\}$ arranged as

$$1 > x_{n,1}(\lambda,a,b) > x_{n,2}(\lambda,a,b) > \cdots > x_{n,n}(\lambda,a,b) > -1. \tag{1.5}$$

It is more convenient to analyze the θ zeros $\{\theta_{n,k}(\lambda,a,b)\}$,

$$x_{n,k}(\lambda,a,b) = \cos(\theta_{n,k}(\lambda,a,b)). \tag{1.6}$$

Novikoff proved that if $\lambda = \frac{1}{2}$ and $(a,b) \neq (0,0)$, then for any fixed k we have

$$\sqrt{n}\,\theta_{n,k}(\tfrac{1}{2},a,b) \to \sqrt{2(a+b)} \quad \text{as } n \to \infty. \tag{1.7}$$

* Received by the editors October 7, 1991; accepted for publication (in revised form) March 13, 1993. This research was partially supported by National Science Foundation grants DMS 8814026 and DMS 8912423.

† Department of Mathematics, University of South Florida, Tampa, Florida 33620.

This is in contrast with the case of ultraspherical polynomials when $a = b = 0$,

$$n\theta_{n,k}(\lambda, 0, 0) \to j_{\lambda-1/2,k} \quad \text{as } n \to \infty,$$

where $j_{\nu,k}$ is the kth positive zero of a Bessell function $J_\nu(z)$. The Pollaczek polynomials are orthogonal with all their zeros in (-1, 1) if and only if $\lambda > 0$ and either $a > |b|$, or $a = b \geq 0$ hold; see Theorems 6.1 and 6.2 in [6]. In both cases the Pollaczek polynomials will be orthogonal with respect to an absolutely continuous measure supported on [-1, 1]. In the other cases of orthogonality we cannot assume that the x zeros can be arranged as in (1.5). This situation will be discussed in §5. One of the reasons for the interest in the Pollaczek polynomials and their zeros is that they do not belong to the Szegö class [2] and as such their θ zeros are not uniform on the semicircle. The spectral properties of the Pollaczek polynomials also proved to be very interesting [2], [6].

It is not clear that Novikoff's methods can be used to extend (1.7) to general $\lambda > 0$. Askey [1] conjectured that the next term in the asymptotic expansion of $\theta_{n,k}$ will involve zeros of a certain transcendental function. The purpose of this paper is to solve Askey's problem. In the process of solving Askey's problem we discovered new properties of a special confluent Horn function ψ_2 [9, p. 225].

Define an even entire transcendental function $F_\lambda(z, c)$ by

$$F_\lambda(z, c) := \int_0^1 (1 - v^2)^{\lambda-1} e^{cv^2} \cos(vz)\, dv. \tag{1.8}$$

THEOREM 1. *Let $\lambda > 0$ and a, b be real and define $\zeta_n > 0$ by*

$$\zeta_n^2 := \frac{2(a+b)}{n} + \sqrt{2(a+b)}\,\xi n^{-3/2} + o(n^{-3/2}). \tag{1.9}$$

Then the asymptotic relationship

$$\begin{aligned} \lim_{n\to\infty} P_n^\lambda(\cos\zeta_n; a, b) \left(\frac{2(a+b)}{n}\right)^{\lambda-1/2} \exp\left(-\pi\sqrt{n(a+b)/2}\right) \\ = \frac{1}{\pi} e^{-a-b} F_\lambda(\zeta, a+b) \end{aligned} \tag{1.10}$$

holds uniformly on compact subsets of the complex ξ plane.

In §2 we will prove Theorem 1. Our proof uses the representation (1.3) and the integral representation [20, §30, p. 47]

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 u^{b-1}(1-u)^{c-b-1}(1-uz)^{-a}\, du, \tag{1.11}$$

provided $\mathrm{Re}(c) > \mathrm{Re}(b) > 0$. If a is a not a negative integer we need the additional assumption $|z| < 1$.

In §3 we shall prove that the zeros of the function $F_\lambda(z, c)$ are real and either simple or double, if $c \geq 0$ and $\lambda > 0$. This will be stated as Theorem 3. We shall also study some analytic properties of the function $F_\lambda(z, c)$ including differential recurrence relations and a differential equation satisfied by $F_\lambda(z, c)$. We conjecture that all the zeros of $F_\lambda(z, c)$ are simple for $c \geq 0$ and $\lambda > 0$, but we have not been able to prove this conjecture. In §4 we shall prove Theorem 2.

THEOREM 2. *Assume that a and b are real such that either $a > |b|$ or $a = b \geq 0$ hold. If $\lambda > 0$ and $(a,b) \neq (0,0)$, then*

$$\theta_{n,k}(\lambda,a,b) = \sqrt{\frac{2(a+b)}{n}}\left\{1 + \frac{\xi_k(\lambda,a+b)}{\sqrt{2(a+b)n}} + O(n^{-1})\right\} \quad \text{as } n \to \infty \tag{1.12}$$

holds for $k = 1,2,\ldots,$ where

$$0 < \xi_1(\lambda,c) \leq \xi_2(\lambda,c) \leq \cdots \leq \xi_k(\lambda,c) \leq \cdots, \tag{1.13}$$

are the positive zeros of the function $F_\lambda(z,c)$.

In §4 we shall also establish results concerning zeros of functions that are finite Fourier cosine transforms.

Theorem 2 solves a problem of Askey [1] and gives a new derivation of Novikoff's unpublished asymptotic results [17] and extends them from $\lambda = \frac{1}{2}$ to $\lambda > 0$. In §5 we shall give an alternate asymptotic formula when the assumptions of Theorem 2 are not valid but the Pollaczek polynomials are orthogonal. We will also mention several remarks.

2. Proof of Theorem 1. Using (1.3), (1.4), and (1.11) we find

$$P_n^\lambda(\cos\theta;a,b)$$
$$= \frac{e^{in\theta}\Gamma(n+2\lambda)/n!}{\Gamma(\lambda+it(\theta))\Gamma(\lambda-it(\theta))}\int_0^1 u^{\lambda+it(\theta)-1}(1-u)^{\lambda-it(\theta)-1}[1-u(1-e^{-2i\theta})]^n du.$$

We used $\Gamma(2\lambda+n)/\Gamma(2\lambda) = (2\lambda)_n$. Set $u = (1+v)/2$ to obtain the more convenient integral representation

$$\frac{n!|\Gamma(\lambda+it(\theta))|^2}{\Gamma(2\lambda+n)}P_n^\lambda(\cos\theta;a,b)$$
$$= 2^{1-2\lambda}e^{in\theta}\int_{-1}^1 (1-v^2)^{\lambda-1}\left(\frac{1+v}{1-v}\right)^{it(\theta)}\left\{\frac{1}{2}(1+e^{-2i\theta}) - \frac{v}{2}(1-e^{-2i\theta})\right\}^n dv$$
$$= 2^{1-2\lambda}\int_{-1}^1 (1-v^2)^{\lambda-1}\left(\frac{1+v}{1-v}\right)^{it(\theta)}(\cos\theta - iv\sin\theta)^n dv.$$

Thus

$$\frac{n!|\Gamma(\lambda+it(\theta))|^2}{\Gamma(2\lambda+n)(\cos\theta)^n}2^{2\lambda-1}P_n^\lambda(\cos\theta;a,b) = A_n(\theta) + A_n(-\theta), \tag{2.1}$$

with

$$A_n(\theta) := \int_0^1 (1-v^2)^{\lambda-1}\left(\frac{1+v}{1-v}\right)^{it(\theta)}(1-iv\tan\theta)^n dv. \tag{2.2}$$

To determine the large n behavior of $A_n(\theta/\sqrt{n})$, write the integrand in (2.2) after replacing θ by $\theta/\sqrt{n}$ in the form

$$(1-v^2)^{\lambda-1}\left(\frac{1+v}{1-v}e^{-2v}\right)^{it(\theta/\sqrt{n})}\exp[2iv\,t(\theta/\sqrt{n}) + n\log(1-iv\tan(\theta/\sqrt{n}))].$$

Recall the definition of $t(\theta/\sqrt{n})$ in (1.4) and note that

$$2iv\,t(\theta/\sqrt{n}) + n\log[1 - iv\tan(\theta/\sqrt{n})] = iv(n^{1/2}/\theta)[2(a+b) - \theta^2] + v^2\theta^2/2 + O(1/\sqrt{n}), \qquad n \to \infty.$$

Thus we have

$$A_n(\theta/\sqrt{n}) := \int_0^1 (1-v^2)^{\lambda-1} g_n(v)\,dv, \tag{2.3}$$

where

$$g_n(v) := \exp\left\{\frac{1}{2}\theta^2 v^2 + \frac{iv\sqrt{n}}{\theta}[2(a+b) - \theta^2] + it(\theta/\sqrt{n})\log\left(e^{-2v}\frac{1+v}{1-v}\right) + O(1/\sqrt{n})\right\}.$$

If θ in (2.3) is allowed to depend on n but tends to a finite limit other than $[2(a+b)]^{1/2}$ as $n \to \infty$, the the right side of (2.3) will tend to zero. This follows from the Riemann–Lebesgue lemma and the asymptotic method of stationary phase [18]. On the other hand, if $\theta^2 = 2(a+b) + \xi[2(a+b)/n]^{1/2} + \eta_n$ and $\eta_n\sqrt{n} \to 0$ as $n \to \infty$, then there is a chance that $A_n(\theta n^{-1/2})$ will tend to a finite limit. Indeed, with ξ_n as in (1.9) we have

$$A_n(\zeta_n) \to F_\lambda(\xi, a+b) + O(1/\sqrt{n}) \quad \text{as } n \to \infty. \tag{2.4}$$

To prove this, consider the integral

$$I_n(\theta) := \int_0^1 (1-v^2)^{\lambda-1} \exp\left\{(a+b)v^2 + i\xi v + it(\theta/\sqrt{n})\log(e^{-2v}(1+v)/(1-v))\right\} dv.$$

Thus

$$I_n(\theta) = F_\lambda(\xi, a+b) + \int_0^1 (1-v^2)^{\lambda-1} e^{(a+b)v^2 + i\xi v}\left\{\exp[it(\theta/\sqrt{n})\log(e^{-2v}(1+v)/(1-v))] - 1\right\} dv.$$

To determine the large n behavior of I_n we need to change variables and integrate by parts. Define functions $u(v)$ and $w(v)$ by

$$w'(v) := -(1-v^2)^{\lambda-1}\exp\left\{(a+b)v^2 + i\xi v\right\}, \qquad w(1) := 0,$$

$$u(v) := -2v + \ln(1+v) - \ln(1-v).$$

Integrate by parts to get

$$I_n(\theta) - F_\lambda(\xi, a+b) = it(\theta/\sqrt{n})\int_0^1 w(v)\exp[it(\theta/\sqrt{n})u(v)]u'(v)dv.$$

Now apply the method of stationary phase [18, §§13.2 and 13.3, pp. 100–101] to see that

$$I_n(\theta) - F_\lambda(\xi, a+b) \to 0 \quad \text{as } n \to \infty.$$

By adding and subtracting $I_n(\theta)$ to $A_n(\zeta_n)$ and keeping track of the error term in the asymptotic forms that are used, one can establish (2.4). Therefore,

$$\frac{\Gamma(n+1)}{\Gamma(n+2\lambda)}|\Gamma(\lambda + it(\zeta_n))|^2(\cos\zeta_n)^{-n}2^{2\lambda-2}P_n^\lambda(\cos\zeta_n) \to F_\lambda(\xi, a+b) \quad \text{as } n \to \infty. \tag{2.5}$$

But as $n \to \infty, n^{1-2\lambda}\Gamma(n+1)/\Gamma(2\lambda+n) \to 1$ [9, §1.18, eq. (4)], and $\log(\cos\zeta_n)^{-n} \approx -n\log[1-(\zeta_n)^2/2] \to a+b$. Futhermore, if x and y are real, then

$$|\Gamma(x+iy)|e^{\pi|y|/2}|y|^{\frac{1}{2}-x} \to \sqrt{2\pi} \quad \text{as } |y| \to \infty$$

[9, §1.18, eq. (6)]. The substitution of the last three asymptotic forms in (2.5) establishes (1.10) and completes the proof of Theorem 1.

3. Properties of $\mathbf{F}_\lambda$. When $\lambda > 0$ and $c \geq 0$ we can introduce Pollaczek parameters a and b as $a = c$, $b = 0$, so the corresponding polynomials will be orthogonal with respect to a measure supported on [-1, 1]; see Szegö [21]. In this case all the θ zeros are real.

THEOREM 3. *If $\lambda > 0$ and $c \geq 0$, then all the zeros of $F_\lambda(z,c)$ are real and simple or double.*

Proof. Let $\zeta_n = \zeta_n(\lambda, c)$ be as in (1.9). On compact subsets of the complex ξ plane, $F_\lambda(\xi, c)$ is the uniform limit of constant multiples Pollaczek polynomials with parameters $\lambda, c, 0$, and argument $\cos\zeta_n$. But the aforementioned Pollaczek polynomials have real and simple θ zeros, hence also have real and simple ξ zeros. By Hurwitz's theorem, Theorem 14.3.4 in [12, p. 205], the limiting function must have only real zeros. We will see that $F_\lambda(z,c)$ satisfies a third-order differential equation; see (3.12). From (3.12) it follows that if $F_\lambda(z,c)$ and its first two derivatives vanish at $z = z_0$, then all the derivatives of the entire function $F_\lambda(z,c)$ will also vanish at z_0, a contradiction. This completes the proof of Theorem 3.

One can identify $F_\lambda(z,c)$ as a confluent Horn function. To see this replace $\exp(cv^2)$ and $\cos vz$ by their Maclaurin series expansions; then set $u = v^2$ and integrate term by term. The result is

$$\begin{aligned} F(z,c) &= \frac{1}{2}\sum_{m,n=0}^{\infty} \frac{c^n}{n!}\frac{(-1)^m z^{2m}}{(2m)!}\int_0^1 (1-u)^{\lambda-1}u^{m+n-1/2}du \\ &= \frac{1}{2}\sum_{m,n=0}^{\infty} \frac{c^n}{n!}\frac{(-1)^m z^{2m}}{(2m)!}\frac{\Gamma(\lambda)\Gamma(m+n+1/2)}{\Gamma(\lambda+m+n+1/2)} \\ &= \frac{\Gamma(\lambda)\Gamma(3/2)}{\Gamma(\lambda+1/2)}\sum_{m,n=0}^{\infty} \frac{(-1)^m c^n (z/2)^{2m}}{n!m!(1/2)_m}\frac{(1/2)_{m+n}}{(\lambda+1/2)_{m+n}}, \end{aligned}$$

where we used $\Gamma(a+n)/\Gamma(a) = (a)_n$ and $(2m)! = 2^{2m}m!(1/2)_m$. Now set $s := m+n$ to get

$$F_\lambda(z,c) = \frac{\Gamma(\lambda)\Gamma(3/2)}{\Gamma(\lambda+1/2)}\sum_{s=0}^{\infty}\frac{1/2_s c^s}{(\lambda+1/2)_s s!}\,{}_1F_1\left(-s; 1/2; \frac{z^2}{4c}\right). \tag{3.1}$$

Apply the Kummer transformation [20, §69, p. 125]

$${}_1F_1(a;b;z) = e^z\,{}_1F_1(b-a;b;-z) \tag{3.2}$$

to the ${}_1F_1$ in (3.1) to obtain

$$F_\lambda(z,c) = \frac{\Gamma(\lambda)\Gamma(3/2)}{\Gamma(\lambda+1/2)}\exp\left(\frac{z^2}{4c}\right)\sum_{s=0}^{\infty}\frac{(1/2)_s c^s}{(\lambda+1/2)_s s!}\sum_{k=0}^{\infty}\frac{(s+1/2)_k}{k!(1/2)_k}\left(-\frac{z^2}{4c}\right)^k. \tag{3.3}$$

Therefore, F_λ is related to a confluent Horn function ψ_2 [9, §5.7.1, p. 225] via

$$F_\lambda(z,c) = [\Gamma(\lambda)\Gamma(\tfrac{3}{2})/\Gamma(\lambda+\tfrac{1}{2})]\exp(z^2/(4c))\psi_2(\tfrac{1}{2},\lambda+\tfrac{1}{2},\tfrac{1}{2},c,-z^2/(4c)). \tag{3.4}$$

Observe that the expansion (3.1) is a Fourier Hermite expansion, since [10, (10.13.2), (10.12.14)]

$$H_{2n}(x) = (-1)^n 2^{2n} n! L_n^{(-1/2)}(x^2) = (-1)^n 2^{2n}(1/2)_n \ {}_1F_1(-n;1/2;x^2).$$

Therefore (3.1) is equivalent to the Fourier–Hermite expansion

$$F_\lambda(2z\sqrt{c},c) = \frac{\Gamma(\lambda)\Gamma(3/2)}{\Gamma(\lambda+1/2)}\sum_{s=0}^{\infty}\frac{(-c/4)^s}{s!(\lambda+1/2)_s}H_{2s}(z). \tag{3.5}$$

Recall the Poisson integral representation [22, §3.3, (3)]

$$\Gamma(v+1/2)J_\nu(z) = \frac{2}{\sqrt{\pi}}\left(\frac{z}{2}\right)^\nu\int_0^1(1-t^2)^{\nu-1/2}\cos(zt)dt. \tag{3.6}$$

Thus

$$F_\lambda(z,0) = \Gamma(3/2)\Gamma(\lambda)(z/2)^{\lambda-1/2}J_{\lambda-1/2}(z).$$

This suggests expanding $F_\lambda(z,c)$ in terms of Bessel functions. We start with (1.8), replace $\exp(cv^2)$ by $e^c\exp[-c(1-v^2)]$, then expand the latter exponential in powers of $1-v^2$. The resulting integrals can be evaluated by (3.6). The final result is

$$F_\lambda(z,c) = e^c\frac{\sqrt{\pi}}{2}\Gamma(\lambda)\left(\frac{2}{z}\right)^{\lambda-1/2}\sum_{n=0}^{\infty}\frac{(\lambda)_n}{n!}\left(-\frac{2c}{z}\right)^n J_{\lambda+n-1/2}(z). \tag{3.7}$$

The functions $F_\lambda(z,c)$ generalize Bessel functions, so we are guided in our study of $F_\lambda(z,c)$ by properties of Bessel functions. As a first step we derive differential recurrence relations for the function $F_\lambda(z,c)$. The corresponding differential operators raise and lower the parameter λ. Differentiate (1.8) twice with respect to x to get the first differential recurrence relation

$$\frac{d^2}{dx^2}F_\lambda(x,c) = F_{\lambda+1}(x,c) - F_\lambda(x,c). \tag{3.8}$$

The second is

$$2c\frac{d^2}{dx^2}F_\lambda(x,c) - x\frac{d}{dx}F_\lambda(x,c) = (2\lambda-1)F_\lambda(x,c) - 2(\lambda-1)F_{\lambda-1}(x,c). \tag{3.9}$$

To prove (3.9) first assume $\lambda > 2$ and replace $(1-v^2)^{\lambda-1}$ by $(1-v^2)(1-v^2)^{\lambda-2}$ in the integrand in (1.8). Thus

$$F_\lambda(x,c) = F_{\lambda-1}(x,c) + \frac{1}{2}\int_0^1(-2v)(1-v^2)^{\lambda-2}[ve^{cv^2}\cos(vx)]dv.$$

An integration by parts gives

$$2(\lambda-1)F_\lambda(x,c) = 2(\lambda-1)F_{\lambda-1}(x,c) - \int_0^1(1-v^2)^{\lambda-1}\frac{d}{dv}[ve^{cv^2}\cos(vx)]dv,$$

and, after simple manipulations we establish (3.9) for $\lambda > 2$. The latter restriction can be weakened to $\lambda > 1$ by analytic continuation. Set

$$D := \frac{d}{dx}, \quad L_1 := I + D^2, \quad L_2 := 2cD^2 - xD - (2\lambda - 1)I, \tag{3.10}$$

where I is the identity operator. The operators L_1 and L_2 raise and lower the parameter λ. It is interesting to note that I, L_1, and L_2 form a Lie algebra. Let L_3 represent multiplication by x^2. The operators I, L_1, L_2, and L_3 form a Lie algebra isomorphic to $su(1, 1)$. Thus $\{I, L_1, L_2, L_3\}$ provide a new realization for $su(1, 1)$. This is a somewhat surprising fact because the group theoretic results known to date dealt with Horn functions as solutions of second-order partial differential equations and as such missed this connection. Furthermore, this suggests that the Pollaczek polynomials are matrix elements of representations of a larger Lie group that degenerates to $SU(1,1)$ by a group contraction.

The next step is to derive an analog of the Bessel differential equation. Differentiate (1.8); then integrate by parts to obtain the differential recurrence relation

$$2\lambda \frac{d}{dx} F_\lambda(x, c) = 2c \frac{d}{dx} F_{\lambda+1}(x, c) - x F_{\lambda+1}(x, c). \tag{3.11}$$

Now eliminate $F_{\lambda+1}$ between (3.8) and (3.11). The result is the differential equation

$$2c \frac{d^3}{dx^3} F_\lambda(x, c) - x \frac{d^2}{dx^2} F_\lambda(x, c) + 2(c - \lambda) \frac{d}{dx} F_\lambda(x, c) - x F_\lambda(x, c) = 0. \tag{3.12}$$

The differential equation satisfied by $u := {}_1F_1(-; \nu + 1, -x^2/4)$ is [20, (3), p. 109]

$$xu'' + (2\nu + 1)u' + xu = 0.$$

It is clear that (3.12) reduces to the above differential equation when $c \to 0$ and $\lambda = \nu + \frac{1}{2}$.

It is important to note that we treat the function $F_\lambda(z, c)$ as a function of one variable, namely, z. Although a confluent Horn function ψ_2 (a, b, c, x, y) satisfies a homogeneous second-order partial differential equation; $F_\lambda(z, c)$, as a function of z, does not seem to satisfy a homogeneous second-order ordinary differential equation.

4. Zeros of finite Fourier cosine transforms. We first outline a proof of Theorem 2.

Proof of Theorem 2. Hurwitz's theorem, Theorem 14.3.4 in Hille [12, p. 205], asserts that if $\{f_n(z)\}$ is a sequence of holomorphic functions in a domain D and $f_n(z)$ converges to $f(z)$ uniformly in D, then for every $\varepsilon > 0$, there is N_ε such that $f_n(z)$, for $n > N_\varepsilon$, has the same number of zeros in the disc $|z - a| < \varepsilon$ as f has. Theorem 2 follows from the latter fact and Theorem 1.

We follow the notation in Pólya [19] and define

$$U(z) := \int_0^1 f(t) \cos(zt) dt, \qquad V(z) := \int_0^1 f(t) \sin(zt) dt. \tag{4.1}$$

We assume that the integrals in (4.1) exist. Langer [14] denotes $U(z)$ and $V(z)$ by $\Phi_c(z)$ and $\Phi_s(z)$, respectively.

THEOREM 4 (Pólya [19]). *Let $f(t)$ be a nonnegative, nondecreasing function on $[0, 1)$. Then zeros $U(z)$ and $V(z)$ are all real and simple and each interval $(n\pi, (n +$*

$1)\pi$) *contains one and only one zeros of* $V(z)$. *If in addition* $f(t)$ *is convex, then* $U(z)$ *has one and only one zero in each interval* $((n-1/2)\pi, n\pi), n = 1, 2, \ldots$.

To prove Theorem 4, Pólya first noted that $U(z)$ and $V(z)$ are uniform limits, on compact sets, of the trigonometric polynomials

$$\frac{1}{n}\sum_{j=0}^{n-1} f(j/n)e^{j/n^2}\cos(jz/n) \text{ and } \frac{1}{n}\sum_{j=0}^{n-1} f(j/n)e^{j/n^2}\sin(jz/n),$$

respectively, as $n \to \infty$ and used a trigonometric version of Kakeya's theorem and Hurwitz's theorem to establish the reality of the zeros of $U(z)$ and $V(z)$. Pólya proved the positivity of the Wronskian of $U(x)$ and $V(x)$, that is,

$$U(x)V'(x) - U'(x)V(x) > 0 \quad \text{for real } x.$$

The above inequality shows that the zeros of $U(z)$and $V(z)$ are simple.

It is worth mentioning that one can show that $U(z)$ has an odd number of zeros in every interval $((n-1/2)\pi, n\pi), n > 0$ as follows. Let $x = (m + a/2)\pi$, where m is a positive integer and $0 \le a \le 1$. Then

$$U\left(\left(m+\frac{1}{2}a\right)\pi\right) = \frac{1}{2m+a}\int_0^{2m+a} f(u/(2m+a))\cos\left(\frac{1}{2}\pi u\right)du.$$

The above integral may be written in the form

$$\frac{1}{2m+a}\left\{\sum_{k=1}^{m}(-1)^k v_k + (-1)^m w_m\right\},$$

where

$$\begin{aligned} v_k &:= (-1)^k \int_{2k-2}^{2k} f(u/(2m+a))\cos\left(\frac{1}{2}\pi u\right)du, w_m \\ &:= (-1)^m \int_{2m}^{2m+a} f(u/(2m+a))\cos\left(\frac{1}{2}\pi u\right)du. \end{aligned}$$

Now use

$$\int_{2k-2}^{2k} = \int_{2k-2}^{2k-1} + \int_{2k-1}^{2k}$$

and set $u = 2k-1-U$ in the first integral and $u = 2k-1+U$ in the second integral to get

$$v_k = \int_0^1 \left\{f\left(\frac{2k-1+U}{2m+a}\right) - f\left(\frac{2k-1-U}{2m+a}\right)\right\}\sin\left(\frac{1}{2}\pi U\right)dU.$$

It is clear that the integrand of the least integral is positive and increases with k. It is also clear that $w_m \ge 0$. Thus sgn $U((m+a/2)\pi) = (-1)^m$ and the proof is complete. This proof parallels a proof of Lommel's theorem given in Watson [22, §15.2]. Watson

attributes the proof of Lommel's theorem to Lommel and is based on an idea that goes back to Bessel. Pólya [19] used a different agrument to prove the same result.

When $f(t) = (1-t^2)^{\nu-1/2}, U(z)$ will be a constant multiple of $z^{-\nu}J_\nu(z)$ and Theorem 4 is applicable when $-\frac{1}{2} < \nu \leq \frac{1}{2}$. This is Lommel's theorem [22, §15.2]. When $f(t) = (1-t^2)^{\lambda-1}\exp(ct^2), 0 < \lambda \leq 1, c \geq 0$, Theorem 4 is also applicable and in this case $U(z) = F_\lambda(z, c)$. The version of Theorem 4 mention in [14] is Theorem 12 on page 234 and contains an obvious misprint, where "non" is deleted from "nondecreasing."

The next theorem was motivated by problem 6 of §17 in Chapter IX of Dieudonné [8]. Its proof requires the following lemma.

LEMMA 1. *The zeros of the function $A\sin z - B(\cos z)/z$ has infinitely many zeros and they are all real and simple provided that $AB > 0$.*

This is the case $\nu = -\frac{1}{2}$ of a result involving zeros of $AJ_\nu(z) + Bz(J_\nu(z))'$; see [10, §7.9, p. 59].

THEOREM 5. *Let $f(t)$ be a real valued twice continuously differentiable function on [0, 1] such that $f'(1)f(1) < 0$, and let $\{x_n : 1 \leq n < \infty\}$ be the positive zeros of*

$$g(z) := f(1)\sin z + f'(1)(\cos z)/z.$$

Define $U(z)$ as in (4.1) and assume $f'(0) = 0$. Then

(i) *The function $zU(z)$ has infinitely many zeros.*

(ii) *For sufficiently large $R, R \neq x_n$ for any $n, g(z)$ and $zU(z)$ have the same number of zeros in $|z| < R$.*

(iii) *The zeros of $zU(z)$ in the right half plane can be indexed as $\{z_n : 1 \leq n < \infty\}$ such that $z_n - x_n \to 0$ as $n \to \infty$.*

Proof. Integrate by parts twice to get

$$zU(z) = f(1)\sin z + f'(1)(\cos z)/z + G(z),$$

where

$$G(z) := -\frac{1}{z}\int_0^1 f''(t)\cos(zt)dt.$$

Clearly the function $f(1)\sin z + f'(1)(\cos z)/z$ is odd and its zeros coincide with the zeros of

$$zf(1)/f'(1) + \cot z.$$

From the graphs of $zf(1)/f'(1)$ and $\cot z$ it is easy to see that $x_{n+1} > n\pi, n \geq 0$, and $x_{n+1} - n\pi \to 0$ as $n \to \infty$. Given r, it is clear that there is a constant C such that

$$|G(z)| \leq Ce^{|\mathrm{Im}\, z|}/|z| \quad \text{and} \quad |f'(1)(\cos z)/z| \leq Ce^{|\mathrm{Im}\, z|}/|z| \quad \text{for } |z| \geq r. \tag{4.2}$$

It is also clear that

$$|\sin z| = \sin^2 x + \sinh^2 y, \qquad z = x + iy. \tag{4.3}$$

Thus if $|\mathrm{Im}\, z| \geq a$, then $|\sin z| \geq e^{-a}(\sinh a)e^{|\mathrm{Im}\, z|}$. This shows that $|g(z)| \geq |G(z)|$ for sufficiently large z if $|\mathrm{Im}\, z|$ is bounded away from zero. If $|\mathrm{Im}\, z|$ is near zero and x is bounded away from x_n and $n\pi$, then (4.2) and (4.3) show that $|g(z)| \geq |G(z)|$ holds on circles centered at $z = 0$ with sufficiently large radii, provided they are bounded away from the set of zeros of $g(z)$ and the integer multiples of π. Parts (i) and (ii)

now follow from Rouchés' theorem [8, §9.17]. Part (iii) also follows from Rouchés' theorem applied to a square contour centered at x_n with vertices $(x_n \pm \varepsilon, \pm i\varepsilon)$, where ε is positive and none of the point $\pi, 2\pi, \ldots$ is inside or on the boundary of the square.

COROLLARY 1. *Under the assumptions of Theorem* 5 *the zeros of* $zU(z)$ *are all simple, except, possibly, for finitely many zeros.*

Theorem 5 and Corollary 1 are applicable when $f(t) = (1 + \varepsilon - t^2)^{\lambda-1} \exp(ct^2)$, if $\varepsilon > 0, \lambda \geq 3$, and $c\varepsilon < \lambda - 1$.

5. Other cases of orthogonal Pollaczek polynomials. The Pollaczek polynomials are orthogonal with respect to a positive measure if and only if

$$\text{(5.1) (i)} \quad \lambda > 0 \quad \text{and} \quad \lambda + a > 0, \quad \text{or} \quad \text{(ii)} \quad -1/2 < \lambda < 0 \quad \text{and} \quad -1 < \lambda + a < 0.$$

It is then natural to extend Theorems 1 and 2 to the full range of parameters in (5.1) and also to study the large n behavior of $\theta_{n,n-k}$ for fixed k. It is easy to see from (1.2) or (1.3) that

$$P_n^{\lambda}(-x; a, b) = (-1)^n P_n^{\lambda}(x; a, -b). \tag{5.2}$$

Therefore,

$$\theta_{n,n-k}(\lambda, a, b) = \pi - \theta_{n,k}(\pi, a, -b) \tag{5.3}$$

and the asymptotics of $\theta_{n,n-k}(\lambda, a, b)$ follows from Theorem 2 if $a > |b|$ and $\lambda > 0$.

Our methods do not seem be able to treat cases with $\lambda < 0$, so we will focus our attention on case (i) in (5.1). In view of (5.2) there is no loss of generality in assuming $b \geq 0$, so will restrict ourselves to

$$\lambda > 0, \qquad b \geq 0. \tag{5.4}$$

Recall that the measure that the Pollaczek polynomials are orthogonal with respect to is absolutely continuous if and only if $\lambda > 0$ and either $a > |b|$ or $a = b \geq 0$; otherwise it will have a nontrivial discrete part. Thus the cases not covered by Theorem 2 all have discrete spectrum. We will follow the notation in [6] and denote the set of mass points by D^*. The set D^* has been characterized in all cases of orthogonality. Define sequences $\{\Delta_n\}, \{x_n\}$, and $\{y_n\}$ by

(5.5)
$$\Delta_n = (n+\lambda)^2 + b^2 - a^2, \quad x_n = \frac{-ab + (n+\lambda)\sqrt{\Delta_n}}{a^2 - (n+\lambda)^2}, \quad y_n = \frac{-ab - (n+\lambda)\sqrt{\Delta_n}}{a^2 - (n+\lambda)^2}.$$

When (5.4) holds the set D^* can be described as follows:

$$\text{(5.6)} \quad \text{Region I*-(i)} := \{(\lambda, a, b) : 0 \leq a < b, b \geq 0, \lambda > 0\}, \quad D^* = \{x_n : n \geq 0\},$$

(5.7)
$$\text{Region II*-(i)} := \{(\lambda, a, b) : -b \leq a < 0, \lambda > 0, \lambda + a > 0, b \geq 0\}, \quad D^* = \{x_n : n \geq 0\},$$

(5.8)
$$\text{Region II*-(ii)} := \{(\lambda, a, b) : a < -b, \lambda > 0, b > 0\}, \quad D^* = \{x_n : n \geq 0\} \cup \{y_n : n \geq 0\}.$$

In all regions x_n and y_n are as in (5.5) and they satisfy $x_n < x_{n+1} < -1, 1 < y_{n+1} < y_n$ for all n. Since the kth smallest zero $x_{n,n-k}(\lambda, a, b)$ will tend to x_k as $n \to \infty$ and, in

region II*-(ii), $x_{n,k} \to y_k$ as $n \to \infty$, then in order to study the limiting behavior of zeros that remain in [-1, 1] we must restrict ourselves to regions I*-(i) and II*-(i).

THEOREM 6. *Assume that* a, b, *and* λ *satisfy* (5.4) *and either* $0 \le a < b$ *or* $-b \le a < b$. *If* $(a,b) \ne (0,0)$, *then*

$$\theta_{n,k}(\lambda,a,b) = \sqrt{\frac{2(a+b)}{n}} + \frac{\xi_k(\lambda,a+b)}{n} + O(n^{-3/2}) \quad \text{as } n \to \infty \tag{5.9}$$

holds for $k = 1, 2, \dots$, *where the* ξ_k*'s are the zeros of* $F_\lambda(z, a+b)$ *ordered as in* (1.13).

The proof of Theorem 6 is similar to the proof we gave for Theorem 2 and will be omitted.

The Liouville–Green or WKB asymptotic method [18] is a powerful tool for determining the asymptotic behavior of solutions of second-order differential equations. The Pollaczek polynomials satisfy the three-term recurrence relation (1.2), which is a second-order difference equation. It is highly desirable to have a discrete analog of Liouville–Green (WKB) method with uniform error bounds, which can be applied to the Pollaczek polynomials. One difficulty is that the limiting function $F_\lambda(z,c)$ does not seem to satisfy a three-term recurrence relation. It is easy to derive a four-term recursion relation for $F_\lambda(z,c)$ from the differential recurrence relations (3.8) and (3.9). The works [5] and [11] formulated discrete analogs of the Liouville–Green but we still do not have uniform asymptotic estimates for these discrete Liouville–Green approximation methods.

The discrete analogs of the functions $U(z)$ and $V(z)$ of (4.1) are exponential sums. The survey article [14] surveys the results on zeros of exponential sums and integrals up to 1931. For later developments see [3], [15], and [16]. In [4], Boas comments on [19] and provides valuable information and references on zeros of exponential sums and integrals.

It is clear that the function $F_\lambda(z,c)$ has a companion function

$$G_\lambda(z,c) = \int_0^1 (1-v^2)^{\lambda-1} e^{cv^2} \sin(vz)\,dv.$$

This suggests a further study of the functions F_λ and G_λ similar to the Polya's investigations of the U and V functions [19].

Acknowledgments. This paper is dedicated to Richard Askey and Frank Olver in appreciation for the help they have given me over the years and for their immeasurable influence on my professional career.

I am grateful to Richard Askey for proposing the topic of this paper and for his constant help and encouragement. I thank Ruiming Zhang for highly recommending Pólya's collected papers where I found [19] and Boas's very informative comments [4].

REFERENCES

[1] R. A. ASKEY, Private communication, January 1989.

[2] R. A. ASKEY AND M. E. H. ISMAIL, *Recurrence relations, continued fractions and orthogonal polynomials*, Mem. Amer. Math. Soc. 300, 1984.

[3] R. P. BOAS, *Entire Functions*, Academic Press, New York, 1954.

[4] ———, *Comments on* [42] *über die Nullstellen gewisser ganzer Funktionen*, in George Pólya: Collected Papers, Vol. 2, MIT Press, Cambridge, MA, 1974, pp. 420–421.

[5] P. A. BRAUN, WKB *method for three term recursions and quasi energies of an anharmonic oscillator*, Teoret. Mat. Fiz., 37 (1978), pp. 355–370, English translation, Theoret. J. Math. Phys., 37 (1978), pp. 1070–1081.

[6] J. CHARRIS AND M. E. H. ISMAIL, *On sieved orthogonal polynomials* V: *Sieved Pollaczek polynomials*, SIAM J. Math. Anal., 18 (1987), pp. 1177–1218.
[7] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.
[8] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1969.
[9] A. ERDÉLYI, W. MAGNUS, R. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.
[10] ———, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.
[11] G. GERONIMO AND D. SMITH, WKB (*Liouville–Green*) *analysis of second order difference equations and applications*, to appear.
[12] E. HILLE, *Analytic Function Theory*, Vol. 2, Ginn, Boston, MA, 1962.
[13] M. E. H. ISMAIL AND M. E. MULDOON, *A discrete approach to monotonicity of zeros of orthogonal polynomials*, Trans. Amer. Math. Soc., 323 (1990), pp. 65–78.
[14] R. E. LANGER, *On the zeros of exponential sums and integrals*, Bull. Amer. Math. Soc., 37 (1931), pp. 213–239.
[15] B. JA. LEVIN, *Distribution of Zeros of Entire Functions*, English translation, American Mathematical Society, Providence, RI, 1964.
[16] N. LEVINSON, *Gap and Density Theorems*, Colloquium Publications, Vol. 26, American Mathematical Society, Providence, RI, 1940.
[17] A. NOVIKOFF, *On a special system of polynomials*, Ph.D. dissertation, Stanford University, Stanford, CA, 1954.
[18] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[19] G. PÓLYA, *Über die nullstellen gewisser ganzer funktionen*, Math. Z., 2 (1918), pp. 352–383; reprinted in Collected Papers, Vol. 2, MIT Press, Cambridge, MA, 1974, pp. 166–197.
[20] E. D. RAINVILLE, *Special Functions*, Chelsea, New York, 1971.
[21] G. SZEGÖ, *Orthogonal Polynomials*, fourth edition, Colloquium Publications, American Mathematical Society, Providence, RI, 1975.
[22] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Second Ed., Cambridge University Press, Cambridge, UK, 1944.

SIAM J. MATH ANAL.
Vol. 25, No. 2, pp. 474–490, March 1994

016

ASYMPTOTIC REMAINDERS*

D. S. JONES†

Abstract. It is shown that the remainders in a variety of asymptotic series can be estimated uniformly by the same integral.

Key words. approximation, uniform asymptotics

AMS subject classification. 41A60

1. Introduction. The determination of estimates for the remainders in asymptotic expansions has been the subject of intensive research in the last few years. It may be said to have originated in the book by Dingle (1973), in which divergent series were summed by Borel integrals. It has come to the fore recently because of the desire to improve the uniform validity of asymptotic formulae and, in particular, to provide smooth transitions across Stokes lines. Berry (1989), following Dingle, suggested that the transition could always be accomplished by a suitable error function but did not furnish a rigorous derivation.

Most attempts to supply a rigorous foundation have centered on specific integral representations. For instance, Boyd (1990) used Stieltjes transforms, Jones (1990) and Olver (1990), (1991a), (1991b) used Laplace transforms, Ursell (1990) discussed Watson's lemma, and Paris (1991), (1992) found profit in Mellin–Barnes integrals. The transition functions that have been discovered to have the widest validity are not error functions themselves but are the incomplete gamma function and related functions. However, all of these functions turn out to behave like an error function near a Stokes line, so that the universality suggested by Berry and Dingle does appear to be justified.

In the meantime, Berry (1990), (1991a), (1991b) continued to find more and more places where the remainder of a series can be estimated by an error function in the neighborhood of a Stokes line. One of his papers (1991a), in particular, considers some highly divergent series. Series that are solutions of certain differential equations have been examined by Paris (1992) (see also Paris and Wood (1985)), but, on the whole there have been few investigations that deal directly with series, as opposed to integral representations, especially those of the type that Berry has looked at.

Therefore, it seemed appropriate to see what progress could be made by starting from a series representation, and some results are given in this paper. The method is flexible and is capable of many generalizations, but here we shall limit the theorems to the more straightforward applications and leave the generalizations to another paper (Jones (1993)).

In some of the proofs simplification is achieved by nonstandard analysis. In this context the symbol $\simeq$ is used between quantities that differ by an infinitesimal. Also, $\mathcal{I}$ will denote a generic infinitesimal and is not necessarily the same whenever it occurs. Thus it is legitimate to write

$$2\mathcal{I} = \mathcal{I}, \quad \ln(1+\mathcal{I}) = \mathcal{I}, \quad b\mathcal{I} = \mathcal{I}$$

* Received by the editors April 1, 1992; accepted for publication (in revised form) February 12, 1993.

† Department of Mathematics and Computer Science, University of Dundee, Dundee DD1 4HN, Scotland, U. K.

when b is limited. However, if ω is unlimited, $\omega\mathcal{I}$ must be retained unaltered since it may be anything from infinitesimal to unlimited.

The nonstandard analysis is based on the set theory of Nelson (1977). Nelson provides an algorithm that reduces a nonstandard statement to a classical one. In essence, the argument goes like this in our context. Suppose that for all infinitesimal ε, $f(\varepsilon) \simeq 0$, where f is a standard function. Then, for any standard $\eta > 0$,

$$(*) \qquad |f(\varepsilon)| < \eta$$

for all infinitesimal ε. Hence for all standard $\eta > 0$ there is $\delta > 0$ such that (*) holds for all $|\varepsilon| \leq \delta$. Since this statement contains no free parameters other than the standard f and η, the transfer principle permits the classical statement that for all $\eta > 0$ there is $\delta > 0$ such that (*) is valid for all $|\varepsilon| \leq \delta$.

For the convenience of the reader the nonstandard results used most frequently in subsequent sections are collected in Appendices A and B.

Estimates of the remainder in a series have been considered for many years, and some early formulae did not involve the error function. Nevertheless, it is shown in §2 that both the early estimate and the error function are different ways of approximating a certain integral. It is conjectured, therefore, that this integral is the real universal character in the assessment of remainders. Supporting evidence is supplied in §§3 and 4, where power series are studied under fairly weak conditions on the coefficients.

Asymptotic series are introduced in §5, and a key theorem concerning them is proved. This permits the demonstration in §6 that many asymptotic power series have remainders that can be expressed in terms of the integral just described. One or two extensions are mentioned, but further details are deferred to another paper.

2. A universal link. The aim of this section is to reconcile what appear, at first sight, to be two totally different ways in which the series $\sum_{n=0}^{\infty} a_n z^n$ performs. A classical result due to Izumi (1927) (see also Titchmarsh (1939)) states that if $a_n/a_{n+1} \to 1$ as $n \to \infty$,

$$\frac{\sum_{m=0}^{n} a_m z^m}{a_n z^n} \to \frac{z}{z-1},$$

provided that $|z| \geq 1+\delta > 1$. Similar behavior was obtained by Olver (1974) for some divergent series when $|z|$ was small. One conjectures that $(z-1)^{-1}$ might represent a characteristic way for power series to behave. On the other hand, the research of Berry suggests that the error function has a universal role. It is not immediately clear how this apparent conflict is resolved.

Our proposal for the reconciliation between these seemingly different points of view is to regard them as aspects of the integral

$$(1) \qquad J(\mu, \varepsilon) = \int_0^{\infty} \frac{t^{\mu} e^{-t}}{1+\varepsilon t} dt$$

and to suggest that J is the proper universal function to adopt. For simplicity, μ will be taken to be real and positive in the following.

When ε is small, the integral in (1) has the following asymptotic representation (see Olver (1991b) and Jones (1990)) for $0 \geq \text{ph}\, \varepsilon \geq -3\pi/2$ $(0 \leq \text{ph}\, (1/\varepsilon) \leq 3\pi/2)$:

$$J(\mu,\varepsilon) \sim \frac{\pi i}{\varepsilon}\left(\frac{e^{-\pi i}}{\varepsilon}\right)^{\mu} e^{1/\varepsilon}\left[2H(-\text{ph }\varepsilon-\pi)-\text{erfc}\left\{ir_0\left(\frac{\mu}{2}\right)^{1/2}\right\}\right]$$
$$+(2\pi\mu)^{1/2}e^{\mu\ln\mu-\mu}\left(\frac{1}{1+\mu\varepsilon}+\frac{1}{\varepsilon\mu r_0}\right), \tag{2}$$

where

$$\text{erfc }(w)=\frac{2}{\pi^{1/2}}\int_w^{\infty} e^{-y^2}\,dy$$

and $H(x)$ is the Heaviside step function. Also,

$$\frac{1}{2}r_0^2=-1-\frac{1}{\mu\varepsilon}-\ln\left(\frac{e^{-\pi i}}{\mu\varepsilon}\right). \tag{3}$$

The sign of r_0 is determined first when ph $\varepsilon=-\pi$, being negative when $\mu|\varepsilon|>1$ and positive when $\mu|\varepsilon|<1$. For other values of ph ε the sign of r_0 is fixed by continuity. Thus when ph ε becomes less than $-\pi$, with $\mu|\varepsilon|>1$, r_0 moves into the second quadrant and Im $(r_0)>0$; in fact, Im $(r_0)>0$ when ph $\varepsilon<-\pi$. Likewise, when ph $\varepsilon>-\pi$, Im $(r_0)<0$. Although (2) refers to one range of ph ε, another can be covered easily by taking a complex conjugate. For example, when $0<\text{ph }\varepsilon\leq 3\pi/2$,

$$J(\mu,\varepsilon) \sim \frac{\pi i}{\varepsilon}\left(\frac{e^{-\pi i}}{\varepsilon}\right)^{\mu} e^{1/\varepsilon}\left[\text{erfc}\left\{-ir_2\left(\frac{\mu}{2}\right)^{1/2}\right\}-2H(\text{ph }\varepsilon-\pi)\right]$$
$$+(2\pi\mu)^{1/2}e^{\mu\ln\mu-\mu}\left(\frac{1}{1+\mu\varepsilon}+\frac{1}{\varepsilon\mu r_2}\right),$$

where

$$\frac{1}{2}r_2^2=-1-\frac{1}{\mu\varepsilon}-\ln\left(\frac{e^{\pi i}}{\mu\varepsilon}\right)$$

and Im $(r_2)>0$ for ph $\varepsilon<\pi$, whereas Im $(r_2)<0$ for ph $\varepsilon>\pi$.

While (2) reveals the relation of (1) to the error function, the connection with the Izumi kind of behavior is provided by the following theorem.

LINK THEOREM. *Let $\varepsilon\simeq 0$ and $|\text{ph }\varepsilon|<\pi-\delta$, where δ is a positive standard number. Then*

$$J(\mu,\varepsilon)=\frac{\mu!(1+\mathcal{I})}{1+\mu\varepsilon}$$

for limited $\mu\varepsilon$.

Before proceeding to the proof, we observe that if ph $\varepsilon>-\pi$ and $|r_0\mu^{1/2}|\gg 1$, the error function in (2) can be replaced by its conventional asymptotics because Im $(r_0)<0$. The net result is that

$$J(\mu,\varepsilon)\sim(2\pi\mu)^{1/2}\frac{e^{\mu\ln\mu-\mu}}{1+\mu\varepsilon}.$$

If Stirling's formula for large μ

$$\mu!=(1+\mathcal{I})(2\pi)^{1/2}\exp\left\{\left(\mu+\frac{1}{2}\right)\ln\mu-\mu\right\}, \tag{4}$$

is invoked, a formula similar to that in the Link Theorem is obtained. Consequently, there are circumstances in which the estimates of J in the Link Theorem and (2) agree. In fact, (2) continues to reproduce the same formula when ph $\varepsilon < -\pi$ because the change in the asymptotic behavior of the error function caused by Im $(r_0) > 0$ is canceled by the contribution from the Heaviside step function. A numerical comparison is provided in Appendix C.

Proof. Consider first the case in which μ is limited. When t is limited, $t\varepsilon$ is infinitesimal and $1 + \varepsilon t \simeq 1$. Moreover, there is a finite K for any t such that

$$\left|\frac{t^\mu e^{-t}}{1+\varepsilon t}\right| < Kt^\mu e^{-t}$$

since $|1 + \varepsilon t| \geq \sin\delta$. It follows from Lemma A.2 that

$$J(\mu,\varepsilon) = \mu!(1+\text{I}) = \frac{\mu!}{1+\mu\varepsilon}(1+\text{I})$$

since $\mu\varepsilon \simeq 0$ when μ is limited.

Next, suppose that μ is unlimited but that $\mu\varepsilon$ is limited. Make the substitution $t = \mu u$, so that the integral becomes

$$\mu^{\mu+1}\int_0^\infty \frac{\exp\{\mu(\ln u - u)\}}{1+\mu\varepsilon u}du.$$

Let $\theta = 1/\mu^{1/4}$, so that θ is, in fact, infinitesimal. For $u \geq 1+\theta$, $-1 + 1/u \leq -\theta/(1+\theta)$ and hence

$$\ln u - u \leq \ln(1+\theta) - 1 - \theta - \frac{\theta(u-1-\theta)}{1+\theta}.$$

Also, because μ is unlimited, Stirling's formula (4) gives

(5)
$$\left|\frac{\mu^{\mu+1}}{\mu!}\int_{1+\theta}^\infty \frac{\exp\{\mu(\ln u - u)\}}{1+\mu\varepsilon u}du\right| \leq \frac{K\mu^{\mu+1}}{\mu!}\exp\{(\mu+1)\ln(1+\theta) - \mu(1+\theta) - \ln(\mu\theta)\}$$
$$\leq K\exp\left(-\frac{1}{2}\ln\mu + \theta - \ln\theta\right) \leq \frac{K}{\mu^{1/4}} \simeq 0,$$

K being adjusted as necessary to absorb finite constants.

For $u \leq 1-\theta$, $-1 + 1/u \geq \theta/(1-\theta)$, so that

$$\ln u - u \leq \ln(1-\theta) - 1 + \theta - \frac{\theta(1-\theta-u)}{1-\theta}.$$

Consequently,

(6)
$$\left|\frac{\mu^{\mu+1}}{\mu!}\int_0^{1-\theta} \frac{\exp\{\mu(\ln u - u)\}}{1+\mu\varepsilon u}du\right| \leq K\exp\left(-\frac{1}{2}\ln\mu - \theta - \ln\theta\right) \simeq 0.$$

From (5) and (6) we can deduce

$$\mu^{\mu+1}\int_0^\infty \frac{\exp\{\mu(\ln u - u)\}}{1+\mu\varepsilon u}du = \mu^{\mu+1}\int_{1-\theta}^{1+\theta} \frac{\exp\{\mu(\ln u - u)\}}{1+\mu\varepsilon u}du + \mu!\,\text{I}.$$

Now, in the integrand on the right $1/(1+\mu\varepsilon u) = (1+\mathcal{I})/(1+\mu\varepsilon)$ because u never differs from unity by more than θ and because

$$\left|\frac{\mu\varepsilon\theta}{1+\mu\varepsilon}\right| < \frac{\theta}{\sin\delta} \simeq 0.$$

Moreover, the interval of integration is limited, and so, by Lemma A.1,

$$\mu^{\mu+1}\int_{1-\theta}^{1+\theta}\frac{\exp\{\mu(\ln u - u)\}}{1+\mu\varepsilon u}du = \frac{1+\mathcal{I}}{1+\mu\varepsilon}\mu^{\mu+1}\int_{1-\theta}^{1+\theta}\exp\{\mu(\ln u - u)\}\,du$$
$$= \frac{1+\mathcal{I}}{1+\mu\varepsilon}\left[\mu^{\mu+1}\int_0^\infty \exp\{\mu(\ln u - u)\}\,du + \mu!\mathcal{I}\right]$$

by repetition of the arguments leading to (5) and (6). The proof of the theorem is concluded. □

It is evident that the proof can be adapted to more general integrands so long as they enjoy properties similar to those of the theorem. Without trying to look for the most general statement it is transparent that the following theorem can be proved by repeating the preceding steps.

THEOREM 1. *If* $|f(\varepsilon t)| < K$ *for all* t, *if* $f(\varepsilon) = (1+\mathcal{I})f(0)$ *for all* $\varepsilon \simeq 0$ *and if*

$$f(\mu\varepsilon u) = (1+\mathcal{I})f(\mu\varepsilon)$$

for $1-\theta \le u \le 1+\theta$, *then*

$$\int_0^\infty f(\varepsilon t)t^\mu e^{-t}dt = \mu! f(\mu\varepsilon)(1+\mathcal{I})$$

for limited $\mu\varepsilon$.

Formula (2) demonstrates that $J(\mu,\varepsilon)$ has the error-function behavior expected by Berry, whereas the Link Theorem shows the sort of function obtained by Izumi. Thus $J(\mu,\varepsilon)$ acts as an intermediary between the two kinds of behavior and has the ability to encompass both. It would appear, therefore, that $J(\mu,\varepsilon)$ offers a better choice as the universal function for estimating asymptotic remainders than does either of the other two. Against that it must be said that the other two are, in general, much easier to compute than is $J(\mu,\varepsilon)$. But, given $J(\mu,\varepsilon)$, there is no reason why it should not be replaced by either of the approximations when that is appropriate. It is also of interest that, in circumstances in which both approximations are sharp, the combination of (2) and the Link Theorem offers a good approximation to the error function.

We turn now to the consideration of series in which $J(\mu,\varepsilon)$ arises naturally in the estimation of remainders.

In all subsequent sections δ will denote a positive standard number, which is not necessarily the same in all places.

3. Behavior of series for lower values of z. The series to be discussed in the first place have the form $\sum_{n=0}^\infty a_n z^n/n!$, where the sequence $\{a_n\}$ is standard. Let

$$s_m(z) = \sum_{n=0}^m \frac{a_n z^n}{n!}, \qquad \sigma_m(z) = \sum_{n=m}^\infty \frac{a_n z^n}{n!}.$$

For our purposes it will be sufficient to assume that

$$\frac{a_{n+1}}{a_n} = 1 + \mathcal{I}/n^{1/2}$$

for unlimited n. This condition is not so restrictive as it might appear at first sight. If z were replaced by z/c it could be changed to

$$\frac{a_{n+1}}{a_n} = c + \mathfrak{I}/n^{1/2}.$$

If, then, c is taken to be e, it is permissible to have a_n growing exponentially by, say, picking $a_n = n^n/n!$

THEOREM 2. *Let $a_{n+1}/a_n = 1 + \mathfrak{I}/n^{1/2}$ when n is unlimited. Let ω be an unlimited positive integer and let $\delta < \operatorname{ph} z < 2\pi - \delta$, where δ is a standard positive number. Then there is an unlimited Δ such that*

$$\sigma_\omega(z) = \frac{1+\mathfrak{I}}{1 - z/\omega} \cdot \frac{a_\omega z^\omega}{\omega!}$$

for $|z| < \omega + \Delta\omega^{1/2}$.

Proof. Write

$$\frac{\omega!\sigma_\omega(z)}{a_\omega z^\omega} = \sum_{n=\omega}^{\infty} \frac{\omega! a_n}{n! a_\omega} z^{n-\omega} = \sum_{n=0}^{\infty} b_n \left(\frac{z}{\omega}\right)^n,$$

where

$$b_n = \frac{\omega!\omega^n a_{\omega+n}}{(\omega+n)! a_\omega}.$$

On account of (4) and Lemma B.1,

$$(7) \qquad b_n = \exp\left[\mathfrak{I}\left\{(\omega+n)^{1/2} - \omega^{1/2}\right\} + n - \left(\omega + n + \frac{1}{2}\right)\ln\left(1 + \frac{n}{\omega}\right) + \mathfrak{I}\right].$$

Suppose now that $|z| \le \omega + d\omega^{1/2}$, where d is positive limited. It is convenient to denote $d/\omega^{1/2}$ by D. The series will be split now into two parts, that in which $n \ge \omega'$ and that in which $n \le \omega' - 1$, where ω' is the next integer above $\omega^{7/12}$.

When $n \ge \omega'$,

$$\left|b_n \left(\frac{z}{\omega}\right)^n\right| \le 2\exp\left\{n + nD - \left(\omega + n + \frac{1}{2}\right)\ln\left(1 + \frac{n}{\omega}\right)\right\}$$

after Lemma B.2 is invoked and an infinitesimal in d is absorbed. The function

$$D - \ln\left(1 + \frac{t}{\omega}\right) - \frac{1/2}{t+\omega}$$

diminishes steadily as t increases from 0; since $\ln(1 + \omega'/\omega) > \omega'/\omega - 1/2(\omega'/\omega)^2$ it is less than $-\omega'/2\omega$ when $t = \omega'$ because of the magnitude of ω' and d being limited. Consequently,

$$\begin{aligned} &n + nD - \left(\omega + n + \frac{1}{2}\right)\ln\left(1 + \frac{n}{\omega}\right) \\ &< -\frac{1}{2}\frac{\omega'}{\omega}(n - \omega') - \left(\omega + \omega' + \frac{1}{2}\right)\ln\left(1 + \frac{\omega'}{\omega}\right) + D\omega' + \omega' < -\frac{\omega' n}{4\omega} \end{aligned}$$

since all other terms are dominated by $\omega' n/4\omega$ on using the logarithmic inequality again. Hence

$$\sum_{n=\omega'}^{\infty} \left| b_n \left(\frac{z}{\omega}\right)^n \right| < \frac{2\exp(-\omega'^2/4\omega)}{1-\exp(-\omega'/4\omega)} < \omega^{5/12} \exp\left(-\frac{\omega^{1/6}}{2}\right) \simeq 0. \tag{8}$$

Now

$$\left(1-\frac{z}{\omega}\right) \sum_{n=0}^{\omega'-1} b_n \left(\frac{z}{\omega}\right)^n = b_0 - b_{\omega'-1}\left(\frac{z}{\omega}\right)^{\omega'} - \sum_{n=0}^{\omega'-2} (b_n - b_{n+1}) \left(\frac{z}{\omega}\right)^{n+1}.$$

Since $b_0 = 1$ and it has been established already that $b'_{\omega'-1}(z/\omega)^{\omega'}$ is infinitesimal,

$$\left(1-\frac{z}{\omega}\right) \sum_{n=0}^{\omega'-1} b_n \left(\frac{z}{\omega}\right)^n = 1 + \mathcal{I} - \sum_{n=0}^{\omega'-2} (b_n - b_{n+1}) \left(\frac{z}{\omega}\right)^{n+1}. \tag{9}$$

To estimate the series on the right of (9) note that

$$b_n - b_{n+1} = b_n \left\{ 1 - \frac{\omega a_{\omega+n+1}}{(\omega+n+1)a_{\omega+n}} \right\} = b_n \left\{ \frac{n+1}{\omega+n+1} + \frac{\mathcal{I}}{(\omega+n)^{1/2}} \right\}.$$

Recalling (7), we have

$$\left| \sum_{n=0}^{\omega'-2} (b_n - b_{n+1}) \left(\frac{z}{\omega}\right)^{n+1} \right| \leq \sum_{n=0}^{\omega'-2} \left(\frac{2}{\omega^{1/2}} + \frac{n}{\omega} \right) \exp\left(\frac{nd}{\omega^{1/2}} - \frac{n^2}{4\omega} \right). \tag{10}$$

From (10) it is clear that when n is limited, a term is infinitesimal. As regards the sum, we take advantage of the definition of the Riemann integral, namely,

$$\int_a^b f(x)dx = \operatorname{st} \sum_{a \leq nh < b} f(nh)h, \tag{11}$$

with h infinitesimal (st means standard part). The selection $h = 1/\omega^{1/2}$ ensures that the sum in (10) is bounded by

$$\int_0^{\infty} (2+t) \exp\left(td - \frac{t^2}{4} \right) dt.$$

Invoking Lemma A.4 we see that the sum of the series on the right of (9) is infinitesimal and that

$$\left(1-\frac{z}{\omega}\right) \sum_{n=0}^{\omega'-1} b_n \left(\frac{z}{\omega}\right)^n = 1 + \mathcal{I}.$$

The combination of this result and (8) demonstrates the theorem when $|z| \leq \omega + d\omega^{1/2}$ with d limited.

Consider now the set of integers $n \in \mathbb{N}$ for which

$$\left| \frac{\omega!(1-z/\omega)\sigma_\omega(z)}{a_\omega z^\omega} - 1 \right| \leq \frac{1}{n}$$

for $(|z| - \omega)/\omega^{1/2} < n$. By what has been proved, this set contains all standard integers n. Since there is no set consisting solely of all the standard integers, there must be an unlimited Δ for which these relations hold with Δ in the place of n. Since $1/\Delta$ is infinitesimal, the proof of the theorem is finished. □

Another way of stating Theorem 2 is

$$\sum_{n=0}^{\infty} \frac{a_n z^n}{n!} = s_{\omega-1}(z) + \frac{1+\mathcal{I}}{1-z/\omega} \cdot \frac{a_\omega z^\omega}{\omega!}. \tag{12}$$

This may also be rewritten, by means of the Link Theorem, as

$$\sum_{n=0}^{\infty} \frac{a_n z^n}{n!} = s_{\omega-1}(z) + (1+\mathcal{I}) \frac{a_\omega z^\omega}{(\omega!)^2} J\left(\omega, \frac{ze^{-\pi i}}{\omega^2}\right). \tag{13}$$

Thus there are two ways of expressing the remainder in the series.

It should also be noted that the condition $\delta < \text{ph } z < 2\pi - \delta$ can be relaxed in Theorem 2. It is required only to prevent z/ω from approaching to within an infinitesimal distance of 1. Therefore, if the condition that $|z/\omega - 1|$ be greater than some positive standard number is imposed, the phase of z can be left unrestricted.

4. Behavior of series for higher z. In this section we consider the series of §3, when $|z| \geq \omega + d\omega^{1/2}$, with d a limited positive number.

THEOREM 3. *With the same conditions on a_n, ω, and* ph z *as in Theorem* 2,

$$s_{\omega-1}(z) = -\frac{1+\mathcal{I}}{1-z/\omega} \cdot \frac{a_\omega z^\omega}{\omega!}$$

for $|z| > \omega + d\omega^{1/2}$, d being standard positive.

Proof. Set

$$\frac{(\omega-1)! s_{\omega-1}(z)}{a_{\omega-1} z^{\omega-1}} = \sum_{n=0}^{\omega-1} C_n \left(\frac{z}{\omega}\right)^{n-\omega+1},$$

where

$$C_n = \frac{(\omega-1)! \omega^{n-\omega+1} a_n}{n! a_{\omega-1}}.$$

When n is a standard integer, (4) and Lemma B.2 supply

$$\left| C_n \left(\frac{z}{\omega}\right)^{n-\omega+1} \right| \leq \frac{|a_n|}{n!} \exp\left\{ \left(n + \frac{1}{2}\right) \ln \omega - \omega + \mathcal{I}\omega^{1/2} \right\} \leq \frac{|a_n|}{n!} \exp\left(-\frac{1}{2}\omega\right) \simeq 0$$

since $|z/\omega| > 1$ and $n + 1 < \omega$. Hence, for standard m

$$\sum_{n=0}^{m} C_n \left(\frac{z}{\omega}\right)^{n-\omega+1} \simeq 0.$$

By Robinson's lemma there is an unlimited M such that $M/\omega \simeq 0$ and

$$\sum_{n=0}^{M} C_n \left(\frac{z}{\omega}\right)^{n-\omega+1} \simeq 0. \tag{14}$$

As in Theorem 2,

$$\left(1-\frac{z}{\omega}\right)\sum_{n=M+1}^{\omega-1}C_n\left(\frac{z}{\omega}\right)^{n-\omega+1}=\not{1}-\frac{z}{\omega}-\sum_{n=M+1}^{\omega-2}(C_n-C_{n+1})\left(\frac{z}{\omega}\right)^{n-\omega+2}$$

since $C_{\omega-1}=1$. Moreover,

$$\sum_{n=M+1}^{\omega-2}(C_n-C_{n+1})\left(\frac{z}{\omega}\right)^{n-\omega+2}=\sum_{m=0}^{\omega-M-3}(C_{\omega-2-m}-C_{\omega-1-m})\left(\frac{z}{\omega}\right)^{-m}. \tag{15}$$

From (4) and Lemma B.1

$$\left|C_{\omega-2-m}\left(\frac{z}{\omega}\right)^{-m}\right|$$
$$\leq \exp\left\{\not{1}\frac{m+1}{(\omega-1)^{1/2}}-\frac{2m}{\omega}-m-\left(\omega-m-\frac{3}{2}\right)\ln\left(1-\frac{m}{\omega}\right)-m\ln\left|\frac{z}{\omega}\right|\right\}.$$

Since $|z/\omega|\geq 1+d/\omega^{1/2}$ and since $\ln(1-x)>-x/(1-x)$, we infer that

$$\left|C_{\omega-2-m}\left(\frac{z}{\omega}\right)^{-m}\right|\leq \exp\left\{-\frac{md}{2\omega^{1/2}}+\frac{3}{2}\ln\left(1-\frac{m}{\omega}\right)\right\}. \tag{16}$$

Furthermore,

$$|C_{\omega-2-m}-C_{\omega-1-m}|\leq |C_{\omega-2-m}|\left(\frac{m}{\omega}+\frac{1}{\omega^{\frac{1}{2}}}\right)\left(\frac{\omega}{\omega-1-m}\right)^{3/2}$$

from Lemma B.1 in Appendix B. Combining this with (16), we have

$$\left|(C_{\omega-2-m}-C_{\omega-1-m})\left(\frac{z}{\omega}\right)^{-m}\right|\leq\left(\frac{m}{\omega}+\frac{1}{\omega^{1/2}}\right)\exp\left(-\frac{md}{4\omega^{1/2}}\right).$$

This shows, on the one hand, that terms in (15) with limited m are infinitesimal and, on the other hand, that the sum of the series is less than

$$\int_0^\infty (1+t)e^{-td/4}\,dt$$

by (11). Lemma A.4 now reveals that the series in (15) is infinitesimal.

Bringing together all the estimates, we obtain

$$\left(1-\frac{z}{\omega}\right)\sum_{n=0}^{\omega-1}C_n\left(\frac{z}{\omega}\right)^{n-\omega+1}=\not{1}-\frac{z}{\omega}.$$

By drawing on $a_\omega\simeq a_{\omega-1}$, the statement of the theorem is confirmed and the proof is complete. □

Again, the constraint on the phase of z can be dropped so long as z/ω differs from unity by a standard positive quantity.

Although both Theorems 2 and 3 indicate a singularity at $z=\omega$, it cannot be inferred that there is a simple pole there. For instance, e^z has $a_n=1$, which meets the conditions of the theorems, but e^z has no singularity at $z=\omega$.

To conclude this section we give a proof of Izumi's result along these lines.

THEOREM 4. *If* $|z| > 1 + \delta$, δ *positive standard,*

$$\sum_{n=0}^{\omega-1} a_n z^n = -\frac{1+\mathcal{I}}{1-z} a_\omega z^\omega.$$

Proof. Let

$$\frac{1}{a_{\omega-1} z^{\omega-1}} \sum_{n=0}^{\omega-1} a_n z^n = \sum_{n=0}^{\omega-1} d_n z^{n-\omega+1},$$

where $d_n = a_n / a_{\omega-1}$. As in Theorem 3, there is an unlimited M such that the terms for $n \leq M$ give an infinitesimal contribution. Also,

$$\sum_{n=M+1}^{\omega-1} d_n z^{n-\omega+1} = \sum_{m=0}^{\omega-M-2} d_{\omega-1-m} z^{-m}.$$

For limited m, $d_{\omega-1-m} \simeq 1$, whereas for any m

$$|d_{\omega-1-m} z^{-m}| \leq \left(\frac{1+\mathcal{I}}{1+\delta}\right)^m.$$

Since $(1+\mathcal{I})/(1+\delta) < 1$, this is a convergent sequence and, by Lemma A.4,

$$\sum_{m=0}^{\omega-M-2} d_{\omega-1-m} z^{-m} = \frac{1+\mathcal{I}}{1-1/z}\left(1 - \frac{1}{z^{\omega-M-1}}\right) = \frac{(1+\mathcal{I})z}{z-1}.$$

The proof of the theorem is now complete. □

5. Asymptotic series. Denote the series $\sum_{m=0}^{\infty} C_m z^m$, C_m being standard, by $F(z)$, and denote a typical remainder by $R_n(z)$, so that

$$R_n(z) = F(z) - \sum_{m=0}^{n} C_m z^m.$$

If

$$R_{n-1}(z) = (C_n + \mathcal{I}) z^n \tag{17}$$

for all standard n when $z \simeq 0$, it will be said that $\sum C_m z^m$ is an asymptotic series for F as $z \to 0$ and the usual notation,

$$F(z) \sim \sum_{m=0}^{\infty} C_m z^m,$$

will be used. Restrictions on ph z may be imposed, and then the series will be asymptotic only for the relevant phases of z.

On this basis, the Maclaurin series of any function with derivatives continuous at the origin is asymptotic. For

$$F(\varepsilon) - \sum_{m=0}^{n} \frac{F^{(m)}(0)}{m!} \varepsilon^m = \frac{F^{(n+1)}(\theta'\varepsilon)}{(n+1)!} \varepsilon^{n+1},$$

with $0 < \theta' < 1$. Since $F^{(n+1)}(\theta'\varepsilon) = (1 + \text{Ł})F^{(n+1)}(0)$, the assertion is confirmed.

THEOREM 5. *Let* $\sum_{m=0}^{\infty} C_m z^m$ *be an asymptotic series as* $z \to 0$ *when* ph $z = \alpha$, *and suppose that*

$$|F(\gamma e^{i\alpha})| < Ae^{\gamma r}$$

with A *and* γ *standard for* $r > 0$. *Then*

$$\sum_{m=0}^{\infty} m! C_m \varepsilon^m$$

is an asymptotic series as $\varepsilon \to 0$ *for* $|\text{ph } \varepsilon - \alpha| \leq \pi/2 - \delta$, *with* δ *standard and positive.*

Proof. Let ε be infinitesimal, and let ph $\varepsilon = \alpha - \beta$, where $|\beta| \leq \pi/2 - \delta$. Define

$$\mathcal{F}(\varepsilon) = \int_0^{\infty e^{i\beta}} e^{-t} F(\varepsilon t) dt.$$

The integral exists by virtue of the hypothesis on the growth of F and $\gamma|\varepsilon|$ being less than $\sin\delta$, because γ is standard, whereas ε is infinitesimal. Furthermore,

$$\mathcal{F}(\varepsilon) = \sum_{m=0}^{n-1} m! C_m \varepsilon^m + \int_0^{\infty e^{i\beta}} e^{-t} R_{n-1}(\varepsilon t) dt. \tag{18}$$

It is necessary to show now that the integral in (18) can be expressed as $(n! C_n + \text{Ł})\varepsilon^n$ when n is standard. For t limited, $\varepsilon t \simeq 0$, and so by (17)

$$\frac{R_{n-1}(\varepsilon t)}{\varepsilon^n} \simeq C_n t^n.$$

The required result now follows from Lemma A.2 provided that the integrand can be bounded suitably.

Set $t = ue^{i\beta}$. When $t\varepsilon \simeq 0$, invoke (17), and then

$$\left| \frac{e^{-t} R_{n-1}(\varepsilon t)}{\varepsilon^n} \right| < (1 + |C_n|) u^n \exp(-u \cos\beta).$$

When $t\varepsilon$ is not infinitesimal u must be unlimited and

$$\begin{aligned} \left| \frac{e^{-t} R_{n-1}(\varepsilon t)}{\varepsilon^n} \right| &= \left| \frac{e^{-t}}{\varepsilon^n} \left\{ F(\varepsilon t) - \sum_{m=0}^{n-1} C_m (\varepsilon t)^m \right\} \right| \\ &\leq (Ae^{\gamma u|\varepsilon|} + A_n u^n) \exp(-u\cos\beta - n \ln|\varepsilon|) \end{aligned}$$

because n is limited. It is now evident that for all u

$$\left| \frac{e^{-t} R_{n-1}(\varepsilon t)}{\varepsilon^n} \right| \leq A' u^n \exp\left(-\frac{1}{4} u \cos\beta \right),$$

and Lemma A.2 can be applied. The proof of the theorem is terminated. □

6. Uniform remainders. The series examined in §§3 and 4 is now discussed further. The formula for the remainder in Theorem 2, coupled with the Link Theorem, suggests that

$$\sigma_\omega(ze^{\pi i}) = \frac{a_\omega(-z)^\omega}{(\omega!)^2} J\left(\omega, \frac{z}{\omega^2}\right) \tag{19}$$

for $|z| < \omega + \Delta\omega^{1/2}$. If (19) is true, it is of wider validity than is the estimate of Theorem 2 for the remainder; it furnishes a transition that is uniform with respect to variations for ph z. Similar remarks can be made for Theorems 3 and 4. However, it is not transparent that (19) is uniformly valid. It is tempting to try verification by analytic continuation, but that course is not open without more knowledge of the analytical properties of the infinitesimals, though some progress might be achieved, within certain limitations, with standard parts. Therefore, uniform remainders will be derived by a different procedure.

Let

$$f(z) = \sum_{n=0}^{\infty} \frac{a_n}{n!}(-z)^n, \qquad F(z) = \sum_{n=0}^{\infty} a_n(-z)^n.$$

The properties of the a_n ensure that $f(z)$ is an entire function and that $F(z)$ has a standard radius of convergence of unity. Indeed, $f(z)$ is of exponential type. More can be said when $F(z)$ is regular at a point of its circle of convergence and can be continued beyond it. For then it is known (see Titchmarsh (1939)) that, if F is regular in crossing the circle of convergence in the direction of $e^{i\alpha}$,

$$|f(te^{i\alpha})| < Me^{\eta t} \tag{20}$$

for any nonnegative t with η standard and $\eta < 1$. One consequence of (20) is that

$$F(z) = \int_0^\infty e^{-t} f(zt)dt \tag{21}$$

for $|z| < 1/\eta$ and ph $z = \alpha$. Let δ_0 be standard positive and such that $1/\eta > 1 + \delta_0$.

THEOREM 6. *Suppose that* $|F(te^{i\alpha})| < Ae^{\gamma t}$ *for all nonnegative* t *and some* α *such that* $|\alpha| \leq \pi - \delta$. *Then, when* $|z| \leq 1/\eta - \delta_0$ *and* ph $z = \alpha$,

$$\begin{aligned} F(z) &= \sum_{n=0}^{\omega-1} a_n(-z)^n + (1+\mathit{I})\frac{a^\omega}{\omega!}(-z)^\omega J\left(\omega, \frac{z}{\omega}\right) \\ &= \sum_{n=0}^{\omega-1} a_n(-z)^n + \frac{1+\mathit{I}}{1+z} a_\omega(-z)^\omega. \end{aligned}$$

Proof. By virtue of (21)

$$F(z) = \sum_{n=0}^{\omega-1} a_n(-z)^n + \int_0^\infty e^{-t} \sum_{n=\omega}^{\infty} \frac{a_n}{n!}(-zt)^n dt.$$

On account of Theorems 2 and 3, the integral can be written as

$$\frac{a_\omega}{\omega!}(-z)^\omega \int_0^\infty \frac{(1+\mathit{I})e^{-t}t^\omega}{1+zt/\omega}dt + \int_{(\omega+\Delta\omega^{1/2})/|z|}^\infty e^{-t} f(zt)dt.$$

The second integral is infinitesimal from (20) and the assumption on $|z|$. The first integral leads to the first statement of the theorem by means of Lemma A.2. The

second statement can be inferred from the Link Theorem. There is nothing more to prove. □

Actually, the exponential growth of F has not been used in the proof; it would have been sufficient to assume that F could be continued analytically across its circle of convergence in the direction of $e^{i\alpha}$. However, for the purposes of the next theorem it is convenient to make just one assumption about F.

THEOREM 7. *If F satisfies the conditions of Theorem 6, then for infinitesimal ε*

$$\sum_{m=0}^{\infty} m!a_m(-\varepsilon)^m = \sum_{m=0}^{n-1} m!a_m(-\varepsilon)^m + a_n(-\varepsilon)^n(1+\text{ɬ}) \cdot \left[J(n,\varepsilon) + \frac{2\pi i}{\varepsilon^{n+1}} e^{1/\varepsilon} \left\{e^{\pi i n} H(\text{ph } \varepsilon - \pi) H(\text{ph } \varepsilon - \alpha) - e^{-\pi i n} H(-\text{ph } \varepsilon - \pi) H(\alpha - \text{ph } \varepsilon)\right\}\right]$$

for every $n \in \mathbb{N}$ and $|\text{ph } \varepsilon - \alpha| \leq \pi/2 - \delta$.

Proof. Define $\mathcal{F}(\varepsilon)$ as in Theorem 5, so that it represents the asymptotic series on the left of the equation in the theorem.

When n is limited, Theorem 5 shows that the remainder is $(n!a_n + \text{ɬ})(-\varepsilon)^n$. Since $n\varepsilon$ is infinitesimal, $J(n,\varepsilon)$ is effectively $n!$ (see §2) and the terms involving the Heaviside step functions are infinitesimal when they are present. Accordingly, there is agreement with the statement of the theorem.

When n is unlimited, Theorem 4 and the second statement of Theorem 6 indicate that

$$\mathcal{F}(\varepsilon) = \sum_{m=0}^{n-1} m!a_m(-\varepsilon)^m + a_n(-\varepsilon)^n(1+\text{ɬ}) \int_0^{\infty e^{i\beta}} \frac{e^{-t}t^n}{1+\varepsilon t} dt + \int_{(1+\delta)e^{i\beta}/|\varepsilon|}^{\infty e^{i\beta}} e^{-t}F(\varepsilon t)dt.$$

The final integral is infinitesimal, as in Theorem 6. In the second term the contour of integration can be deformed into the real axis, possibly capturing a pole in the process. The net result is the formula of the theorem, and the proof has been completed. □

Theorems 6 and 7 reveal how the remainders in certain series may be represented in the same universal manner by means of the function J. In many practical cases one may expect F to comply with the conditions of Theorem 6 for a continuum of values of α, so that the conclusions of the two theorems will hold for wider ranges of ph z and ph ε than they may appear to at first sight. Also, analytic continuation may permit less restriction on β than has been implemented in Theorem 5, with consequent relaxation of the constraint on ph ε.

Other series can give rise to remainders that can be expressed in terms of J. For instance, suppose that

$$F(z) = \sum_{n=0}^{\infty} a_n(-z^2)^n$$

and that F is still of exponential growth at infinity. The argument of Theorem 7 can be repeated, and the remainder will entail

$$\int_0^{\infty} \frac{t^{\mu}e^{-t}}{1+\varepsilon^2 t^2} dt = \frac{1}{2}\{J(\mu, i\varepsilon) + J(\mu, -i\varepsilon)\}.$$

Thus the error function will have an important role in the transition through a Stokes line for the associated asymptotic series.

Another type of series in which the error function comes in has terms originating from logarithms. For example, let $\mathcal{F}(\varepsilon)$ in Theorem 7 be redefined as

$$\mathcal{F}(\varepsilon) = \int_0^{\infty e^{i\beta}} e^{-t} F(\varepsilon t) \ln t \, dt.$$

The terms in the asymptotic series are calculated easily, and the remainder will contain the factor

$$\frac{\partial}{\partial \mu} J(\mu, \varepsilon).$$

It is plain from (2) that the partial derivative will not eliminate the error function nor introduce any new transcendental functions. Clearly, more complicated definitions of $\mathcal{F}$ could be used, but the details are left to the paper, Jones (1993), already referred to.

Appendix A. In this appendix are derived some approximations that are required in the main text.

The first is a simple result, but it is quoted so frequently that it is worth a reference.

LEMMA A.1. *If $f(x) \simeq 0$ for all x of an interval, then*

$$\sup |f(x)| \simeq 0.$$

Proof. For any standard $b > 0$, $|f(x)| < b$ for every x of the interval by hypothesis. Hence $\sup |f(x)| < b$, and, since b is any positive standard number, the lemma follows. □

LEMMA A.2. *Let $f(x) \simeq g(x)$ for all limited x. Suppose there is an $h(x) \in L(0, \infty)$ such that $|f(x)| \leq h(x)$, $|g(x)| \leq h(x)$ for all $x \in \mathbb{R}$. Then*

$$\int_0^\infty f(x)dx \simeq \int_0^\infty g(x)dx.$$

Proof. Let n be a positive limited integer. Then

$$\left| \int_0^n \{f(x) - g(x)\} dx \right| \leq \int_0^n \text{ł} \, dx \leq \text{ł} n$$

by Lemma A.1. The right-hand side is infinitesimal since n is limited. Therefore, by Robinson's lemma there is an unlimited ν such that

$$\int_0^\nu f(x)dx \simeq \int_0^\nu g(x)dx.$$

Furthermore,

$$\left| \int_\nu^\infty \{f(x) - g(x)\} dx \right| \leq 2 \int_\nu^\infty h(x)dx \simeq 0$$

since $h \in L(0, \infty)$. Addition of these formulas completes the proof. □

A slight variant of Lemma A.2 is sometimes useful.

LEMMA A.3. *If $f(x) = \dagger h(x)$ for all limited x, $h(x) \in L(0,\infty)$ and $|f(x)| \leq h(x)$ for all $x \in \mathbb{R}$, then*

$$\int_0^\infty f(x)dx = \dagger \int_0^\infty h(x)dx.$$

Proof. The argument goes along the same lines as that of Lemma A.2, but it starts from

$$\int_0^n f(x)dx = \dagger \int_0^n h(x)dx. \qquad \square$$

Infinite series can be handled in the same way. Since the principle of proof is virtually unchanged, only the lemmas themselves will be stated.

LEMMA A.4. *Let $a_n \simeq b_n$ for all limited integers n. Suppose that $\sum_{n=0}^\infty C_n$ is a standard convergent series such that $|a_n| \leq C_n$, $|b_n| \leq C_n$ for all $n \in \mathbb{N}$. Then*

$$\sum_{n=0}^\infty a_n \simeq \sum_{n=0}^\infty b_n.$$

LEMMA A.5. *Let $a_n = \dagger C_n$ for all limited integers n, $\sum_{n=0}^\infty C_n$ convergent and $|a_n| \leq C_n$ for all $n \in \mathbb{N}$, then*

$$\sum_{n=0}^\infty a_n = \dagger \int_{n=0}^\infty C_n.$$

Appendix B. Here some of the properties of a_n when $a_{n+1}/a_n = 1 + \dagger/n^{1/2}$ for unlimited n are elucidated.

With n unlimited and m a positive integer,

(B.1)

$$\ln \frac{a_{n+m}}{a_n} = \sum_{p=n}^{n+m-1} \ln \frac{a_{p+1}}{a_p} = \sum_{p=n}^{n+m-1} \ln\left(1 + \frac{\dagger}{p^{1/2}}\right) = \sum_{p=n}^{n+m-1} \frac{\dagger}{p^{1/2}} = \dagger \sum_{p=n}^{n+m-1} \frac{1}{p^{1/2}}$$

on account of Lemma A.1. Now

$$\int_n^{n+m} \frac{dt}{t^{1/2}} < \sum_{p=n}^{n+m-1} \frac{1}{p^{1/2}} < \int_{n-1}^{n+m-1} \frac{dt}{t^{1/2}},$$

so that

$$(n+m)^{1/2} - n^{1/2} < \sum_{p=n}^{n+m-1} \frac{1}{2p^{1/2}} < (n+m-1)^{1/2} - (n-1)^{1/2}.$$

Since the two sides of the inequality differ by an infinitesimal multiple, it can be asserted that

$$\sum_{p=n}^{n+m-1} \frac{1}{2p^{1/2}} = \{(n+m)^{1/2} - n^{1/2}\}(1+\dagger).$$

Inserting this result into (B.1), we have the following lemma.

LEMMA B.1. *If n is unlimited and if m is a positive integer,*

$$\frac{a_{n+m}}{a_n} = \exp[\cancel{I}\{(n+m)^{1/2} - n^{1/2}\}]$$

and

$$\left|\frac{a_{n+m}}{a_n}\right| \leq \exp\left(\frac{\cancel{I}m}{n^{1/2}}\right).$$

The second statement is a consequence of the inequality

$$(n+m)^{1/2} - n^{1/2} \leq \frac{m}{2n^{1/2}}.$$

The expression for a_{n+m}/a_n in Lemma B.1 suggests that $a_n = \exp(\cancel{I}n^{1/2})$ for unlimited n. That this is true can be proved as follows.

If n is unlimited and q is a standard integer, $q/n \simeq 0$ and $(\ln a_q)/n^{1/2} \simeq 0$. Therefore, by Robinson's lemma there is an unlimited Q such that $Q/n \simeq 0$ and that $(\ln a_Q)/n^{1/2} \simeq 0$. Therefore,

$$\ln a_n = \ln\left(\frac{a_n}{a_Q}\right) + \ln a_Q = \cancel{I}(n^{1/2} - Q^{1/2}) + \ln a_Q$$

from Lemma B.1. By virtue of the definition of Q, the conjectured expression for a_n is confirmed.

LEMMA B.2. *If n is unlimited*

$$a_n = \exp(\cancel{I}n^{1/2}).$$

Appendix C. This appendix gives a couple of comparisons of the numerical values of the two approximations for J. The parameters chosen are $\mu = 3$, $|\varepsilon| = 0.5$ and $\mu = 6$, $|\varepsilon| = 0.1$, with the phase of ε going from 0 to $-6\pi/5$ in steps of $\pi/10$. These are relatively moderate values of the parameters and offer a severe test for the formulae. In particular, ε is nowhere near the infinitesimal specified in the Link Theorem. Nevertheless, the performance of the formulas can be regarded as very satisfactory.

The first column under each pair of parameters in Table 1 shows the ratio of relation (2) to the function in the Link Theorem. The second column gives the ration of (2) to J, but no effort was made to evaluate J by analytic continuation when $\text{ph}\,\varepsilon = -\pi$. As can be seen from the table, the agreement of the Link Theorem prediction worsens as $\text{ph}\,\varepsilon$ diminishes, whereas that of (2) improves. Yet, at the smaller value of $|\varepsilon|$ the two do not vary by much, the main difference being in the phase predicted. Overall, it can be concluded that (2) is always reliable and that the simpler result of the Link Theorem should be perfectly adequate for ballpark estimates so long as $|\varepsilon|$ is not too large.

TABLE 1

Phase of ε	$\mu = 3$, $\lvert\varepsilon\rvert = 0.5$		$\mu = 6$, $\lvert\varepsilon\rvert = 0.1$	
	(2)/Link	(2)/J	(2)/Link	(2)/J
0	$0.944 - 0.003i$	$1.021 - 0.003i$	$0.972 + 0.007i$	$1.009 + 0.007i$
$-\pi/10$	$0.938 - 0.004i$	$1.016 - 0.002i$	$0.969 + 0.008i$	$1.007 + 0.005i$
$-\pi/5$	$0.929 - 0.007i$	$1.012 - 0.001i$	$0.965 + 0.011i$	$1.005 + 0.004i$
$-3\pi/10$	$0.918 - 0.010i$	$1.009 - 0.001i$	$0.960 + 0.014i$	$1.004 + 0.003i$
$-2\pi/5$	$0.903 - 0.016i$	1.006	$0.953 + 0.021i$	$1.002 + 0.003i$
$-\pi/2$	$0.881 - 0.025i$	1.004	$0.944 + 0.031i$	$1.001 + 0.002i$
$-3\pi/5$	$0.849 - 0.041i$	1.003	$0.930 + 0.049i$	$1.001 + 0.002i$
$-7\pi/10$	$0.800 - 0.071i$	1.002	$0.913 + 0.084i$	$1 + 0.001i$
$-4\pi/5$	$0.723 - 0.131i$	1.001	$0.892 + 0.158i$	$1 + 0.001i$
$-9\pi/10$	$0.605 - 0.264i$	1	$0.890 + 0.329i$	1
$-\pi$	0.440	–	1.039	–
$-11\pi/10$	$0.605 + 0.264i$	1	$0.890 - 0.329i$	$1 - 0.001i$
$-6\pi/5$	$0.723 + 0.131i$	1	$0.892 - 0.158i$	$1 - 0.001i$

REFERENCES

M. V. BERRY (1989), Proc. Roy. Soc. London Ser. A, 422, pp. 7–21.

——— (1990), Proc. Roy. Soc. London Ser. A, 427, pp. 265–280.

——— (1991a), Proc. Roy. Soc. London Ser. A, 434, pp. 465–472.

——— (1991b), Proc. Roy. Soc. London Ser. A, 435, pp. 437–444.

W. G. C. BOYD (1990), Proc. Roy. Soc. London Ser. A, 429, pp. 227–246.

R. B. DINGLE (1973), *Asymptotic Expansions: Their Derivation and Interpretation*, Academic Press, London.

S. IZUMI (1927), Japan J. Math., 4, pp. 29–32.

D. S. JONES (1990), in Proc. International Conference on Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, pp. 241–264.

——— (1993), in Ordinary and Partial Differential Equations, B. D, Sleeman and R. J. Jarvis, eds., Research Notes in Mathematics, Vol. 4, Longman Press, London, pp. 126–145.

E. NELSON (1977), Bull. Amer. Math. Soc., 83, pp. 1165–1198.

F. W. J. OLVER (1974), *Asymptotics and Special Functions*, Academic Press, New York.

——— (1990), in Proc. International Conference on Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, pp. 329–355.

——— (1991a), SIAM J. Math. Anal., 22, pp. 1475–1489.

——— (1991b), SIAM J. Math. Anal., 22, pp. 1460–1474.

R. B. PARIS AND A. D. WOOD (1985), Proc. Roy. Irish Acad. Sect. A, 85, pp. 201–220.

R. B. PARIS (1991), Proc. Roy. Soc. London Ser. A, 432, pp. 391–426.

——— (1992), Proc. Roy. Soc. London Ser. A, 436, pp. 165–186.

E. C. TITCHMARCH (1939), *The Theory of Functions*, Oxford University Press, London.

F. J. URSELL (1990), in Elasticity, Mathematical Methods and Applications, G. Eason and R. W. Ogden, eds., Ellis Horwood, Chichester, England, pp. 391–395.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 491–512, March 1994

017

APPLICATION OF SZEGÖ POLYNOMIALS TO FREQUENCY ANALYSIS*

WILLIAM B. JONES†, OLAV NJÅSTAD‡, AND HAAKON WAADELAND§

Abstract. This paper is concerned with the frequency analysis problem of determining unknown frequencies ω_j from a sample of N observed values of a time signal $x_{N,I}$, that is, the superposition of I sinusoidal waves. In recent work [Jones, Njåstad, Thron, and Waadeland, Journal CAM, 46 (1993), pp. 217–228], [Pan and Saff, J. Approx. Theory, 71 (1992), pp. 239–251] it has been shown that, for $1 \le I < \infty$ and $k \ge 2I(+1)$, $\lim_{N\to\infty} z_j(k,\psi_{N,I}) = e^{i\omega_j}$ for $2I(+1)$ zeros $z_j(k,\psi_{N,I})$ of the kth degree Szegö polynomials $\rho_k(\psi_{N,I};z)$ orthogonal on the unit circle with respect to a distribution function $\psi_{N,I}(\theta)$ determined by the signal $x_{N,I}$. Hence the ω_j can be approximated by computing Arg $z_j(k,\psi_{N,I})$. In the present paper the Szegö polynomial method is extended to apply to signals $x_{N,I}$ with $1 \le I \le \infty$. It is shown that as the amplitudes $|\alpha_j|$ of high-order sinusoidal waves approach zero and as $N \to \infty$, $z_j(k,\psi_{N,I}) \to e^{i\omega_j}$ is obtained. It is proved that the sequence of distribution functions $\{\psi_{N,I}(\theta)/N\}$ converges in the weak star sense to a function $\psi_{\infty,I}(\theta)$ as $N \to \infty$. That result is used to establish the convergence of a sequence of two-point Padé approximants to a Carathéodory function $F_{\infty,I}(z)$. The significance of this for frequency analysis lies in the fact that $\rho_k(\psi_{N,I};z)$ is the polynomial denominator of a two-point Padé approximant, and the critical points $e^{i\omega_j}$ are the poles of $F_{\infty,I}(z)$. Results from numerical examples are given to illustrate the use of the method.

Key words. frequency analysis, orthogonal polynomials, Padé approximants

AMS subject classifications. 33C45, 40A15, 41A21

1. Introduction.

Recently, attention has been focused on the study of Szegö polynomials $\rho_k(\psi_{N,I};z)$ associated with discrete time signals $x_{N,I} = \{x_{N,I}(m)\}_{m=-\infty}^{\infty}$ of the form

$$x_{N,I}(m) = \begin{cases} \displaystyle\sum_{j=-I}^{I} \alpha_j e^{i\omega_j m}, & 0 \le m < N, \quad (x_{N,I}(o) \ne 0), \\ 0 \quad \text{elsewhere}, \end{cases} \tag{1.1a}$$

where

$$1 \le N \le \infty, \tag{1.1b}$$

$$\omega_0 = 0, \quad 0 < \omega_j = -\omega_{-j} < \pi, \quad \omega_j \ne \omega_k \quad \text{for } j \ne k, \tag{1.1c}$$

$$0 \le \alpha_0 \le r_0, \quad 0 < |\alpha_1| \le r_1, \quad 0 \le |\alpha_j| \le r_j, \quad \alpha_j = \overline{\alpha}_{-j}, \quad r_j = r_{-j} \quad \text{for } j \ge 1 \tag{1.1d}$$

(see, e.g., [3], [4], [8], [9], [10], [18], [19]). In those papers "I" was taken to be a natural number. In the present paper we consider all I such that $1 \le I \le \infty$; hence

* Received by the editors April 13, 1992; accepted for publication (in revised form) February 12, 1993.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309-0395. This research was supported in part by the United States Educational Foundation in Norway through a Fulbright Grant, the Norwegian Research Council for Science and the Humanities (NAVF) and the United States National Science Foundation under grant DMS-9103141.

‡ Department of Mathematical Sciences, University of Trondheim, Norwegian Institute of Technology, N-7034 Trondheim, Norway.

§ Department of Mathematics and Statistics, College of Arts and Sciences, University of Trondheim N-7055 Dragvoll, Norway.

this additional requirement is imposed:

$$\sum_{\substack{j=-I \\ j\neq -k}}^{I} \sum_{k=-I}^{I} r_j r_k \left|\csc\left(\tfrac{\omega_j+\omega_k}{2}\right)\right| < \infty, \tag{1.1e}$$

which itself implies

$$\sum_{j=-I}^{I} r_j < \infty. \tag{1.2}$$

We denote by $\rho_n(\psi_{N,I}; z)$ the monic nth degree Szegö polynomial orthogonal on the unit circle with respect to the distribution function $\psi_{N,I}(\theta)$ defined as follows.

(a) If $1 \leq N < \infty$ and $1 \leq I \leq \infty$, then $\psi_{N,I}(\theta)$ is an absolutely continuous function on $[-\pi, \pi]$, such that

$$\psi'_{N,I}(\theta) := \frac{1}{2\pi}|X_{N,I}(e^{i\theta})|^2, \qquad X_{N,I}(z) := \sum_{m=0}^{N-1} x_{N,I}(m) z^{-m}, \tag{1.3}$$

which implies

$$\psi'_{N,I}(-\theta) = \psi'_{N,I}(\theta) \quad \text{for } -\pi \leq \theta \leq \pi. \tag{1.4}$$

(b) For $N = \infty$ and $1 \leq I \leq \infty$,

$$\psi_{\infty,I}(\theta) := \sum_{\substack{j \\ \omega_j \leq \theta}} |\alpha_j|^2, \qquad -\pi \leq \theta \leq \pi. \tag{1.5}$$

Thus, if $1 \leq I < \infty$, $\psi_{\infty,I}(\theta)$ is a step function with jumps $|\alpha_j|^2$ at the points $\theta = \omega_j$.

Determination of the unknown normalized frequencies ω_j from the values of the signal $x_{N,I}(m)$ is called the frequency analysis problem. It arises in the study of physical phenomena described by functions $G(t)$ of time t (seconds) of the form

$$G(t) = \sum_{j=-I}^{I} \alpha_j e^{i2\pi f_j t}, \quad f_j = -f_{-j} > 0, \quad \alpha_j = \overline{\alpha}_{-j} \in \mathbf{C} \quad \text{for } j \geq 1, f_o = 0, \alpha_o \geq 0. \tag{1.6}$$

Using equally spaced instants of time $t_m = m\Delta t$, $m = 0, 1, 2, \ldots$, we obtain from (1.1a) and (1.6) that

$$x_{N,I}(m) = G(t_m) \quad \text{for } 0 \leq m < N \quad \text{and} \quad \omega_j = 2\pi f_j \Delta t \quad \text{for } j \geq 1. \tag{1.7a}$$

Since our study restricts ω_j to $0 < \omega_j < \pi$ for $j \geq 1$, we must have

$$0 < \Delta t < \frac{1}{2f_j} \left(\text{i.e., } 0 < f_j < \frac{1}{2\Delta t}\right) \quad \text{for } j \geq 1. \tag{1.7b}$$

Thus the choice of a sampling interval Δt imposes a limit on the frequencies f_j that can be found. Frequency analysis problems arise from phenomena such as human speech [14], ocean tides [20], and radar.

The following related conjecture on the asymptotics of Szegö polynomial zeros was introduced in [4]: For $1 \le I < \infty$, as n and N tend to infinity (in a manner to be determined), the $n_0(I) := 2I + L$ zeros $z_j(n, \psi_{N,I})$ of $\rho_n(\psi_{N,I}; z)$ of largest moduli approach the critical points $e^{i\omega_j}$, $j = \pm 1, \pm 2, \dots, \pm I$ and also $e^{i0} = 1$ if $L = 1$. Here $L := 1$ if $\alpha_0 > 0$ and $L := 0$ if $\alpha_0 = 0$.

This conjecture has recently been verified by results of two papers [8] and [19] for n fixed, $n \ge n_0(I)$, and $N \to \infty$. In both papers it is shown that for each $n \ge n_0(I)$ there exists an arrangement of the zeros $z_j(n, \psi_{N,I})$ such that

$$\lim_{N\to\infty} z_j(n, \psi_{N,I}) = e^{i\omega_j}, \qquad j = \pm 1, \pm 2, \dots, \pm I, \tag{1.8a}$$

$$\lim_{N\to\infty} z_0(n, \psi_{N,I}) = e^{i0} = 1 \quad \text{if } L = 1 \,(i.e., \alpha_0 > 0). \tag{1.8b}$$

It can be easily seen from the proof in [19] that for each $n \ge n_0(I)$, there exists a number λ_n, with $0 < \lambda_n < 1$, such that for all of the remaining $n - n_0(I)$ zeros $z_j(n, \psi_{N,I})$ of $\rho_n(\psi_{N,I}; z)$, we have

$$|z_j(n, \psi_{N,I})| \le \lambda_n < 1 \quad \text{for all } 1 \le N < \infty. \tag{1.9}$$

Thus, for N large enough, those zeros of $\rho_n(\psi_{N,I}; z)$ which can be used to find the unknown frequencies can be distinguished from the other zeros. The Szegö polynomial method of solving frequency analysis problems (FAP) consists of computing the polynomials $\rho_n(\psi_{N,I}; z)$ and its zeros $z_j(n, \psi_{N,I})$, or at least the ones nearest the unit circle $|z| = 1$. This method has been shown [3], [4] to be a reformulation of the Wiener–Levinson method [13], [24] based on linear prediction and digital filters. Other methods for solving the FAP are discussed in [1], [12], [20], and [21].

In the present paper we assume that $1 \le I \le \infty$ and that $\{\omega_j\}$ and $\{r_j\}$ are sequences of numbers satisfying (1.1c, d, e). The results we obtain apply to all signals $x_{N,I}$ in (1.1a), where $\{\alpha_j\}_{-I}^{I}$ satisfies (1.1d). Although one cannot expect to determine infinitely many frequencies by a polynomial method, it is shown (§4) that under certain conditions one can approximate frequencies ω_j associated with relatively large amplitudes $|\alpha_j|$. More specifically we show (Theorem 4.1) that for each K, with $1 \le K < I \le \infty$, the zeros $z_j(n_0(K), \psi_{N,I})$ of $\rho_{n_0(K)}(\psi_{N,I}; z)$ can be arranged so that

$$\lim_{\substack{N\to\infty \\ \alpha(K,I)\to 0}} z_j(n_0(K), \psi_{N,I}) = e^{i\omega_j}, \qquad j = \pm 1, \pm 2, \dots, \pm K, \quad \alpha(K, I) := \sum_{j=K+1}^{I} |\alpha_j|. \tag{1.10}$$

Therefore, for $\alpha(K, I)$ small enough and N large enough, $\operatorname{Arg} z_j(n_0(K), \psi_{N,I})$ is an approximation of ω_j for each $j = \pm 1, \pm 2, \dots, \pm K$.

Computational results are considered in §5 for an example where $I = 4$ and $|\alpha_1|$ is much larger than other amplitudes. The first four rows of the table — Table 2a — ($k = 2, 4, 6, 8$, corresponding to $K = 1, 2, 3, 4$) illustrate (1.10) for $j = \pm 1$. Further evidence supporting the applicability of the Szegö polynomial method is described in §3. It is shown (Theorem 3.2) that, for $1 \le I \le \infty$,

$$\lim_{\substack{N\to\infty \\ k\to\infty}} \frac{1}{N} \frac{P_{2k+1}(\psi_{N,I}; z)}{Q_{2k+1}(\psi_{N,I}; z)} = F_{\infty,I}(z) = \sum_{j=-I}^{I} |\alpha_j|^2 \frac{e^{i\omega_j} + z}{e^{i\omega_j} - z} \quad \text{for } |z| > 1, \tag{1.11}$$

where $P_{2k+1}(\psi_{N,I};z)/Q_{2k+1}(\psi_{N,I};z)$ is the $(2k+1)$th (two-point Padé) approximant of a positive PC-continued fraction (3.8) associated with the signal $x_{N,I}$. The significance of this result for frequency analysis lies in the fact that

$$\rho_k(\psi_{N,I};z) = Q_{2k+1}(\psi_{N,I};z), \tag{1.12}$$

and hence (1.11) suggests that, under suitable conditions, zeros $z_j(k,\psi_{N,I})$ of $\rho_k(\psi_{N,I};z)$ approach singularities of the Carathéodory function $F_{\infty,I}(z)$. This result is of interest for its own sake since, together with Theorem 4.1, it provides information on the location of the poles of two-point Padé approximants for a large class of Carathéodory functions. Our proof of Theorem 3.2 makes use of a weak star convergence result (Theorem 2.1): $\psi_{N,I}/N \overset{*}{\to} \psi_{\infty,I}$ as $N \to \infty$ for all $1 \le I \le \infty$. This result extends a similar theorem given in [4, Thm. 7] for $1 \le I < \infty$.

In the remainder of this introduction we summarize known properties of Szegö polynomials that are subsequently used. Further results of Szegö polynomials can be found in [5], [7], [15], [17], and [23].

The mth moments $\mu_m^{(N,I)}$ with respect to $\psi_{N,I}$ are defined by

$$\mu_m^{(N,I)} := \int_{-\pi}^{\pi} e^{-im\theta} d\psi_{N,I}(\theta), \qquad m = 0, \pm 1, \pm 2, \ldots, \quad 1 \le N, I \le \infty. \tag{1.13}$$

They can be computed by the following formulas valid for $1 \le I \le \infty$:

$$\mu_m^{(N,I)} = \begin{cases} \displaystyle\sum_{k=0}^{N-m-1} x_{N,I}(k)x_{N,I}(k+m), & m \ge 0 \\ \mu_{-m}^{(N,I)}, & m < 0 \end{cases} \qquad \text{for } 1 \le N < \infty, \tag{1.14a}$$

$$\mu_m^{(\infty,I)} = \sum_{j=-I}^{I} |\alpha_j|^2 e^{i\omega_j m} \qquad m = 0, \pm 1, \pm 2, \ldots \quad \text{for } N = \infty. \tag{1.14b}$$

For $[1 \le N < \infty, 1 \le I \le \infty]$ and for $[N = \infty, I = \infty]$ the bisequence $\{\mu_m^{(N,I)}\}_{m=-\infty}^{\infty}$ is positive definite, that is,

$$\mu_{-n}^{(N,I)} = \mu_n^{(N,I)} \quad \text{and} \quad \Delta_n^{(N,I)} := T_{n+1}^{(0)}(N,I) > 0, \qquad n = 0, 1, 2, \ldots, \tag{1.15}$$

where the Toeplitz determinants $T_k^{(m)}(N,I)$ are defined by

$$T_0^{(m)}(N,I) := 1, \; T_k^{(m)}(N,I) := \det(\mu_{m+j-n}^{(N,I)})_{j,n=0}^{k-1}, \qquad k = 1, 2, 3, \ldots, 1 \le N, I \le \infty.$$

For $[N = \infty, 1 \le I < \infty]$, $\{\mu_m^{(\infty,I)}\}_{m=-\infty}^{m=\infty}$ is positive $n_0(I)$-definite, since

$$\begin{aligned} &\mu_{-n}^{(\infty,I)} = \mu_n^{(\infty,I)}, \qquad \Delta_n^{(\infty,I)} := T_{n+1}^{(0)}(\infty,I) > 0 \quad \text{for } 0 \le n \le n_0(I) - 1, \\ &\Delta_{n_0(I)}^{(\infty,I)} := T_{n_0(I)+1}^{(0)}(\infty,I) = 0. \end{aligned} \tag{1.16}$$

For $1 \le N$, $I \le \infty$ we define

$$\langle f, g \rangle_{\psi_{N,I}} := \int_{-\pi}^{\pi} f(e^{i\theta}) \overline{g(e^{i\theta})} d\psi_{N,I}(\theta).$$

The monic Szegö polynomials $\rho_n(\psi_{N,I};z)$ and reciprocal (reversed) polynomials $\rho_n^*(\psi_{N,I};z) := z^n\overline{\rho_n(\psi_{N,I};1/\bar{z})} = z^n\rho_n(\psi_{N,I};z^{-1})$ are defined by $\rho_0(\psi_{N,I};z) = \rho_0^*(\psi_{N,I};z) = 1$ and, for $1 \le n < n_0(I)+1$ where $n_0(\infty) := \infty$, by

$$\rho_n(\psi_{N,I};z) := \frac{1}{\Delta_{n-1}^{(N,I)}} \begin{vmatrix} \mu_0^{(N,I)} & \mu_{-1}^{(N,I)} & \cdots & \mu_{-n}^{(N,I)} \\ \mu_1^{(N,I)} & \mu_0^{(N,I)} & \cdots & \mu_{-n+1}^{(N,I)} \\ \vdots & \vdots & & \vdots \\ \mu_{n-1}^{(N,I)} & \mu_{n-2}^{(N,I)} & \cdots & \mu_{-1}^{(N,I)} \\ 1 & z & \cdots & z^n \end{vmatrix},$$

(1.17)

$$\rho_n^*(\psi_{N,I};z) := \frac{1}{\Delta_{n-1}^{(N,I)}} \begin{vmatrix} \mu_0^{(N,I)} & \mu_1^{(N,I)} & \cdots & \mu_n^{(N,I)} \\ \mu_{-1}^{(N,I)} & \mu_0^{(N,I)} & \cdots & \mu_{n-1}^{(N,I)} \\ \vdots & \vdots & & \vdots \\ \mu_{-n+1}^{(N,I)} & \mu_{-n+2}^{(N,I)} & \cdots & \mu_1^{(N,I)} \\ z^n & z^{n-1} & \cdots & 1 \end{vmatrix}.$$

They satisfy orthogonality relations, for $1 \le n < n_0(I)+1$,

$$\text{(1.18a)} \quad \langle \rho_n(\psi_{N,I};z), z^m\rangle_{\psi_{N,I}} = \begin{cases} 0, & 0 \le m \le n-1, \\ \Delta_n^{(N,I)}/\Delta_{n-1}^{(N,I)}, & m = n, \end{cases}$$

$$\text{(1.18b)} \quad \langle \rho_n^*(\psi_{N,I};z), z^m\rangle_{\psi_{N,I}} = \begin{cases} \Delta_n^{(N,I)}/\Delta_{n-1}^{(N,I)}, & m = 0, \\ 0, & 1 \le m \le n. \end{cases}$$

These yield the recurrence relations

$$\text{(1.19a)} \quad \rho_n(\psi_{N,I};z) = z\rho_{n-1}(\psi_{N,I};z) + \delta_n^{(N,I)}\rho_{n-1}^*(\psi_{N,I};z), \qquad 1 \le n < n_0(I)+1,$$

$$\text{(1.19b)} \quad \rho_n^*(\psi_{N,I};z) = \delta_n^{(N,I)}z\rho_{n-1}(\psi_{N,I};z) + \rho_{n-1}^*(\psi_{N,I};z), \qquad 1 \le n < n_0(I)+1,$$

where the reflection coefficients $\delta_n^{(N,I)} := \rho_n(\psi_{N,I};0)$ satisfy

$$\text{(1.20)} \quad \delta_n^{(N,I)} = \frac{(-1)^n T_n^{(-1)}(N,I)}{T_n^{(0)}(N,I)} = -\frac{\sum_{j=0}^{n-1} q_j^{(n-1,N,I)}\mu_{-j-1}^{(N,I)}}{\sum_{j=0}^{n-1} q_j^{(n-1,N,I)}\mu_{-j+1-n}^{(N,I)}}, \qquad 1 \le n < n_0(I)+1,$$

where $\sum_{j=0}^{n-1} q_j^{(n-1,N,I)} z^j := \rho_{n-1}(\psi_{N,I};z)$. Levinson's algorithm utilizes (1.18) and (1.19) to compute successively the $\delta_n^{(N,I)}$ and $q_j^{(n-1,N,I)}$.

2. Weak star convergence. Before giving the main results of this section (Theorems 2.1 and 2.2), we describe a family of signals $x_{N,I}$ satisfying conditions (1.1) with $I = \infty$ and with $\omega_0 = 0$ being the limit of an infinite sequence of frequencies ω_j.

Example 2.1. Let $\{r_j\}$ and $\{\omega_j\}$ be defined by

$$r_0 := 0,\ r_j = \frac{1}{|j|^3} \quad \text{and} \quad \omega_j = \frac{1}{j} \quad \text{for} \quad j = \pm 1, \pm 2, \ldots . \tag{2.1}$$

We show that $\{r_j\}$ and $\{\omega_j\}$ satisfy (1.1e). Since $\frac{2}{\pi}x < \sin x$ for $0 < x < \frac{\pi}{2}$, it follows that for all integers m and n such that $0 \neq m \neq -n \neq 0$,

$$\left| \frac{\frac{1}{m} \cdot \frac{1}{n}}{\sin(\frac{1}{2m} + \frac{1}{2n})} \right| < \frac{|\frac{1}{mn}|}{\frac{2}{\pi}|\frac{1}{2m} + \frac{1}{2n}|} = \frac{\pi}{|m+n|} \leq \pi.$$

Therefore,

$$\sum_{0 \neq m \neq -n \neq 0}^{\infty} \sum^{\infty} \left| \frac{\frac{1}{m^3} \cdot \frac{1}{n^3}}{\sin(\frac{\frac{1}{m} + \frac{1}{n}}{2})} \right| < 4\pi \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \frac{1}{m^2} \cdot \frac{1}{n^2} \leq 4\pi \sum_{m=1}^{\infty} \frac{1}{m^2} \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^5}{9}.$$

Additional examples can be verified in a similar manner.

THEOREM 2.1 (weak star convergence). *For each* $1 \leq I \leq \infty$, *as* $N \to \infty$,

$$\frac{1}{N}\psi_{N,I}(\theta) \overset{*}{\to} \psi_{\infty,I}(\theta) = \sum_{j:\omega_j \leq \theta} |\alpha_j|^2, \qquad -\pi \leq \theta \leq \pi. \tag{2.2}$$

To prove (2.2) it suffices to show that for every function $f(\theta)$ continuous on $-\pi \leq \theta \leq \pi$,

$$\lim_{N\to\infty} \frac{1}{N} \int_{-\pi}^{\pi} f(\theta) d\psi_{N,I}(\theta) = \int_{-\pi}^{\pi} f(\theta) d\psi_{\infty,I}(\theta) = \sum_{j=-I}^{I} |\alpha_j|^2 f(\omega_j). \tag{2.3}$$

For use in our proof we recall a few well-known results and prove one basic lemma. From [22], e.g., it is known that, for $\omega \in [-\pi, \pi]$ and $N = 1, 2, 3, \ldots$

$$\int_{-\pi}^{\pi} \left[\frac{\sin N(\frac{\theta-\omega}{2})}{\sin(\frac{\theta-\omega}{2})} \right]^2 d\theta = 2\pi N\ . \tag{2.4}$$

Hence, by Schwarz's inequality,

$$\int_{-\pi}^{\pi} \left| \frac{\sin N(\frac{\theta-\omega}{2})}{\sin(\frac{\theta-\omega}{2})} \right| d\theta \leq \left[\int_{-\pi}^{\pi} \left[\frac{\sin N(\frac{\theta-\omega}{2})}{\sin(\frac{\theta-\omega}{2})} \right]^2 d\theta \right]^{\frac{1}{2}} \left[\int_{-\pi}^{\pi} 1^2 d\theta \right]^{\frac{1}{2}} = 2\pi\sqrt{N}. \tag{2.5}$$

Furthermore (also from [22]),

$$\lim_{N\to\infty} \frac{1}{2\pi N} \int_{\omega-\varepsilon_1}^{\omega+\varepsilon_2} \left[\frac{\sin N(\frac{\theta-\omega}{2})}{\sin(\frac{\theta-\omega}{2})} \right]^2 f(\theta) d\theta = f(\omega), \qquad \varepsilon_1 > 0,\ \varepsilon_2 > 0. \tag{2.6}$$

From $|\sin\varphi| = 2|\sin\varphi/2 \cos\varphi/2| \leq 2|\sin\varphi/2|$, one obtains

$$|\sin\tfrac{\varphi}{2}| \geq \tfrac{1}{2}|\sin\varphi|. \tag{2.7}$$

For convenience we introduce the notation

$$S_N(\theta,\alpha,\beta) := \frac{\sin N(\frac{\alpha-\theta}{2})\sin N(\frac{\beta-\theta}{2})}{\sin(\frac{\alpha-\theta}{2})\sin(\frac{\beta-\theta}{2})}. \tag{2.8}$$

LEMMA 2.2. *There exists a number $K > 0$ such that for all $-\pi < \alpha < \beta < \pi$ and $N = 1, 2, 3, \ldots$*

$$\int_{-\pi}^{\pi} |S_N(\theta,\alpha,\beta)|d\theta \le \frac{K\sqrt{N}}{\sin(\frac{\beta-\alpha}{2})}. \tag{2.9}$$

Proof. Since the integrand is a periodic function of θ with period 2π, we have

$$\int_{-\pi}^{\pi} |S_N(\theta,\alpha,\beta)|d\theta = \int_{-\pi+v}^{\pi+v} |S_N(\theta,\alpha,\beta)|d\theta \quad \text{for all } v \in \mathbf{R}.$$

We set $\gamma := \beta - \alpha$ so that $0 < \gamma < 2\pi$ and consider the change of variables $\theta = \varphi + \alpha$. Then it suffices to prove that there exists $K > 0$ such that for all $0 < \gamma < 2\pi$, $N = 1, 2, 3, \ldots$,

$$\int_{-\pi}^{\pi} |S_N(\varphi,0,\gamma)|d\varphi \le \frac{K\sqrt{N}}{\sin(\frac{\gamma}{2})}. \tag{2.10}$$

From (2.5) we know that one can find a $K_0 > 0$ such that for $N \ge 1$ and $\gamma \in \mathbf{R}$,

$$\int_{-\pi}^{\pi} \left|\frac{\sin N(\frac{\varphi}{2})}{\sin(\frac{\varphi}{2})}\right| d\varphi \le K_0\sqrt{N} \text{ and } \int_{-\pi}^{\pi} \left|\frac{\sin N(\frac{\gamma-\varphi}{2})}{\sin(\frac{\gamma-\varphi}{2})}\right| d\varphi \le K_0\sqrt{N}. \tag{2.11}$$

We divide the interval $[-\pi,\pi]$ into two parts $[-\pi,\pi] = A \cup B$, $A := [-\gamma/2, \gamma/2]$, $B := [-\pi,\pi]\backslash A$. First we consider the part of the integral (2.10) over the set B. Since

$$B = [\varphi : \tfrac{\gamma}{2} < |\varphi| < \pi] = [\varphi : \tfrac{\gamma}{4} < |\tfrac{\varphi}{2}| < \tfrac{\pi}{2}],$$

we obtain by (2.7)

$$|\sin\tfrac{\gamma}{2}| > \sin\tfrac{\varphi}{4} > \tfrac{1}{2}\sin\tfrac{\gamma}{2} \quad \text{for all } \varphi \in B,$$

and hence

$$\int_{\beta} |S_N(\varphi,0,\gamma)|d\varphi \le \frac{2K_0\sqrt{N}}{\sin(\frac{\gamma}{2})}. \tag{2.12}$$

To deal with the part of the integral (2.10) over A we consider two cases.

Case 1. Suppose $0 < \gamma \le \pi$. Then $(3\gamma/4) \le \pi - (\gamma/4)$; hence

$$A := [\varphi : -\tfrac{\gamma}{2} \le \varphi \le \tfrac{\gamma}{2}] = [\varphi : \tfrac{\gamma}{4} \le \tfrac{\gamma-\varphi}{2} \le \tfrac{3\gamma}{4}] \subseteq [\varphi : \tfrac{\gamma}{4} \le \tfrac{\gamma-\varphi}{2} \le \pi - \tfrac{\gamma}{4}]$$

and, for all $\varphi \in A$,

$$|\sin(\tfrac{\gamma-\varphi}{2})| \ge \min[\sin\tfrac{\gamma}{4}, \sin(\pi - \tfrac{\gamma}{4})] = \sin\tfrac{\gamma}{4} > \tfrac{1}{2}\sin\tfrac{\gamma}{2}.$$

Then for all $N \ge 1$ and $0 < \gamma \le \pi$,

$$\int_A |S_N(\varphi,0,\gamma)|d\varphi \le \frac{1}{\sin\frac{\gamma}{4}} \int_A \left|\frac{\sin N(\frac{\varphi}{2})}{\sin(\frac{\varphi}{2})}\right| d\theta \le \frac{2K_0\sqrt{N}}{\sin\frac{\gamma}{2}}. \tag{2.13}$$

Case 2. Suppose $\pi < \gamma < 2\pi$. We set

$$\gamma' := 2\pi - \gamma \quad \text{so that } 0 < \gamma' < \pi.$$

Then by (2.13), for all $N \geq 1$ and $\pi < \gamma < 2\pi$, there exists a $K_1 > 0$ such that (with $\psi := -\varphi$)

$$\begin{aligned} \int_A |S_N(\psi, 0, \gamma)| d\psi &= \int_A |S_N(\varphi, 0, -\gamma)| d\varphi = \int_A |S_N(\varphi, 0, \gamma')| d\varphi \\ &\leq \frac{2K_0\sqrt{N}}{\sin\frac{\gamma'}{2}} = \frac{2K_0\sqrt{N}}{\sin\frac{\gamma}{2}}. \end{aligned} \tag{2.14}$$

Combining (2.12), (2.13), and (2.14) yields (2.10). □

Proof of Theorem 2.1. By (1.1), (1.3), and (2.8), we have for all $1 \leq N < \infty$ and $1 \leq I \leq \infty$,

$$\begin{aligned} \psi'_{N,I}(\theta) &= \frac{1}{2\pi}\left|\sum_{m=0}^{N-1} x_{N,I}(m) e^{-im\theta}\right|^2 = \frac{1}{2\pi}\left|\sum_{m=0}^{N-1}\left(\sum_{j=-I}^{I} \alpha_j e^{i\omega_j m}\right) e^{-im\theta}\right|^2 \\ &= \frac{1}{2\pi}\left|\sum_{j=-I}^{I} \alpha_j \frac{1 - e^{-iN(\omega_j-\theta)}}{1 - e^{-i(\omega_j-\theta)}}\right|^2 = \frac{1}{2\pi}\left|\sum_{j=-I}^{I} \alpha_j \frac{e^{iN(\frac{\omega_j-\theta}{2})}\sin N(\frac{\omega_j-\theta}{2})}{e^{i(\frac{\omega_j-\theta}{2})}\sin(\frac{\omega_j-\theta}{2})}\right|^2 \\ &= \frac{1}{2\pi}\sum_{j=-I}^{I}\sum_{k=-I}^{I} \alpha_j\overline{\alpha}_k \frac{e^{iN(\frac{\omega_j-\omega_k}{2})}}{e^{i(\frac{\omega_j-\omega_k}{2})}} S_N(\theta, \omega_j, \omega_k). \end{aligned} \tag{2.15}$$

Therefore,

$$\frac{1}{N}\int_{-\pi}^{\pi} f(\theta) d\psi_{N,I}(\theta) = \frac{1}{2\pi N}\int_{-\pi}^{\pi} f(\theta) \sum_{j,k=-I}^{I} \alpha_j\overline{\alpha}_k \frac{e^{iN(\frac{\omega_j-\omega_k}{2})}}{e^{i(\frac{\omega_j-\omega_k}{2})}} S_N(\theta, \omega_j, \omega_k) d\theta. \tag{2.16}$$

We first show that the order of integration and summation in (2.16) can be interchanged (even if $I = \infty$). Let j and k be any to integers such that $\omega_j < \omega_k (j, k \in [-I, I])$. Let M be chosen such that $|f(\theta)| \leq M$ for $\theta \in [-\pi, \pi]$. Then by Lemma 2.2 there exists a $K > 0$ such that

$$\int_{-\pi}^{\pi} |f(\theta) S_N(\theta, \omega_j, \omega_k)| d\theta \leq \frac{MK\sqrt{N}}{\sin(\frac{\omega_k-\omega_j}{2})} \quad \text{for } N = 1, 2, 3 \ldots. \tag{2.17}$$

Therefore by condition (1.1e) and the assumption $r_j = r_{-j}$ and $\omega_j = -\omega_{-j}$

$$\begin{aligned} &\frac{1}{2\pi N}\left|\sum_{j=-I_{j\neq k}}^{I}\sum_{k=-I}^{I} \alpha_j\overline{\alpha}_k \int_{-\pi}^{\pi} f(\theta) \frac{e^{iN(\frac{\omega_j-\omega_k}{2})}}{e^{i(\frac{\omega_j-\omega_k}{2})}} S_n(\theta, \omega_j, \omega_k) d\theta\right| \\ &\quad \leq \frac{MK}{\sqrt{2\pi N}} \sum_{j=-I_{j\neq k}}^{I}\sum_{k=-I}^{I} r_j r_k |\csc(\tfrac{\omega_k-\omega_j}{2})| \longrightarrow 0 \text{ as } N \to \infty. \end{aligned} \tag{2.18}$$

(Note that the double sum is equal to the double sum in (1.1e).) By (2.4)

$$\begin{aligned}(2.19)\qquad &\frac{1}{2\pi N}\left|\sum_{j=-I}^{I}|\alpha_j|^2\int_{-\pi}^{\pi}f(\theta)\left[\frac{\sin N(\frac{\omega_j-\theta}{2})}{\sin(\frac{\omega_j-\theta}{2})}\right]^2 d\theta\right| \\ &\quad\le\frac{M}{2\pi N}\sum_{j=-I}^{I}|\alpha_j|^2\int_{-\pi}^{\pi}\left[\frac{\sin N(\frac{\omega_j-\theta}{2})}{\sin(\frac{\omega_j-\theta}{2})}\right]^2 d\theta = M\sum_{j=-I}^{I}|\alpha_j|^2<\infty.\end{aligned}$$

It follows from (2.18) and (2.19) and an application of Fubini's theorem [16] that the order of integration and summation in (2.16) can be interchanged. From this we obtain, for $N\ge 1$,

(2.20)

$$\begin{aligned}\frac{1}{N}\int_{-\pi}^{\pi}f(\theta)d\psi_{N,I}(\theta) &= \frac{1}{2\pi N}\sum_{j=-I}^{I}|\alpha_j|^2\int_{-\pi}^{\pi}f(\theta)\left[\frac{\sin N(\frac{\omega_j-\theta}{2})}{\sin(\frac{\omega_j-\theta}{2})}\right]^2 d\theta \\ &\quad+\frac{1}{2\pi N}\sum_{j=-I_{j\ne k}}^{I}\sum_{k=-I}^{I}\alpha_j\overline{\alpha}_k\int_{-\pi}^{\pi}f(\theta)\frac{e^{iN(\frac{\omega_j-\omega_k}{2})}}{e^{i(\frac{\omega_j-\omega}{2})}}S_N(\theta,\omega_j,\omega_k)d\theta.\end{aligned}$$

By (2.18) the second sum on the right side of (2.20) tends to zero as $N\to\infty$. By (2.6) the first sum on the right side of (2.20) approaches, as $N\to\infty$,

$$\sum_{j=-I}^{I}|\alpha_j|^2 f(\omega_j)=\int_{-\infty}^{\infty}f(\theta)d\psi_{\infty,I}(\theta),$$

from which we can conclude that (2.3) holds. □

THEOREM 2.3 (uniform convergence). *For each* $1\le N\le\infty$,

$$(2.21)\qquad \psi_{N,I}(\theta)\longrightarrow\psi_{N,\infty}(\theta)\quad -\pi\le\theta\le\pi,\quad \textit{as } I\to\infty,\ \textit{uniformly.}$$

Proof. First we consider $N=\infty$. Let $\varepsilon>0$ be given. By (1.2) there exists a number $I_0(\varepsilon)$ such that

$$(2.22)\qquad \sum_{|j|>I_0(\varepsilon)}r_j^2<\varepsilon.$$

Then for all I, $I_0(\varepsilon)\le I<\infty$,

$$(2.23)\qquad |\psi_{\infty,\infty}(\theta)-\psi_{\infty,I}(\theta)|=\sum_{\substack{\omega_j\le\theta\\ j>I}}|\alpha_j|^2\le\sum_{j>I}|\alpha_j|^2\le\sum_{j>I_0(\varepsilon)}|\alpha_j|^2<\varepsilon.$$

Thus (2.21) holds for $N=\infty$.

Next we consider $1\le N<\infty$. Then by (2.15), the triangle inequality, and

$$\left|\frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}}\right|=\left|\sum_{m=0}^{N-1}e^{im(\omega_j-\theta)}\right|\le N,$$

it can be seen that

(2.24)

$$\psi'_{N,\infty}(\theta) - \psi'_{N,I}(\theta) = \left[\sqrt{\psi'_{N,\infty}(\theta)} + \sqrt{\psi'_{N,I}(\theta)}\right] \cdot \left[\sqrt{\psi'_{N,\infty}(\theta)} - \sqrt{\psi'_{N,I}(\theta)}\right]$$

$$= \frac{1}{2\pi}\left\{\left|\sum_{j=-\infty}^{\infty} \alpha_j \frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}}\right| + \left|\sum_{j=-I}^{I} \alpha_j \frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}}\right|\right\}$$

$$\cdot\left\{\left|\sum_{j=-\infty}^{\infty} \alpha_j \frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}}\right| - \left|\sum_{j=-I}^{I} \alpha_j \frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}}\right|\right\}$$

$$\leq \frac{1}{\pi}\left\{\sum_{j=-\infty}^{\infty} |\alpha_j| \cdot \left|\frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}}\right|\right\}$$

$$\cdot\left|\sum_{j=-\infty}^{\infty} \alpha_j \frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}} - \sum_{j=-I}^{I} \alpha_j \frac{1-e^{iN(\omega_j-\theta)}}{1-e^{i(\omega_j-\theta)}}\right|$$

$$\leq \frac{N^2}{\pi}\left(\sum_{j=-\infty}^{\infty} |\alpha_j|\right)\left(\sum_{|j|>I} |\alpha_j|\right) \leq \frac{N^2}{\pi}\left(\sum_{j=-I}^{I} r_j\right)\left(\sum_{|j|>I} r_j\right).$$

Let $\varepsilon > 0$ be given. By (1.2) there exists a number $I_0(\varepsilon)$ such that

$$\sum_{|j|>I} r_j < \frac{\varepsilon\pi}{N^2 \sum_{j=-\infty}^{\infty} r_j} \quad \text{for all } I \geq I_0(\varepsilon).$$

It follows from (2.24) that for all $I \geq I_0(\varepsilon)$,

$$|\psi'_{N,\infty}(\theta) - \psi'_{N,I}(\theta)| < \varepsilon$$

for all $\theta \in [-\pi, \pi]$. From this result the uniform convergence of $\psi_{N,I}(\theta)$ follows for $N < \infty$. □

3. Moment generating functions. In this section we consider the function of a complex variable

$$F_{N,I}(z) := \int_{-\pi}^{\pi} \frac{e^{i\theta}+z}{e^{i\theta}-z} d\psi_{N,I}(\theta), \quad 1 \leq N, I \leq \infty, \quad |z| \neq 1, \tag{3.1}$$

where $\psi_{N,I}(\theta)$ is a distribution function of the form (1.3) or (1.5). Each such function $F_{N,I}(z)$ belongs to the class of normalized Carthéodory functions (see, e.g., [5] and [7]). For completeness some properties of the functions (3.1) are summarized in Theorem 3.1. In view of (3.4) we refer to $F_{N,I}(z)$ as the moment generating function for $\psi_{N,I}(\theta)$. A related convergence result is described by Theorem 3.2.

THEOREM 3.1. (a) *For each N and I such that $1 \leq N$, $I \leq \infty$, $F_{N,I}(z)$ is holomorphic for $|z| \neq 1$ and satisfies the following properties*:

$$F_{N,I}(0) = \mu_0^{(N,I)} > 0, \tag{3.2}$$

(3.3) $\quad \operatorname{Re} F_{N,I}(z) \geq 0 \quad \textit{for} \quad |z| < 1 \quad \textit{and} \quad \operatorname{Re} F_{N,I}(z) \leq 0 \quad \textit{for} \quad |z| > 1,$

$$\text{(3.4)} \qquad F_{N,I}(z) = \begin{cases} L_0^{(N,I)}(z) := \mu_0^{(N,I)} + 2\sum_{k=1}^{\infty} \mu_k^{(N,I)} z^k & \textit{for} \quad |z| < 1, \\ L_\infty^{(N,I)}(z) := -\mu_0^{(N,I)} - 2\sum_{k=1}^{\infty} \mu_k^{(N,I)} z^{-k} & \textit{for} \quad |z| > 1, \end{cases}$$

$$\text{(3.5)} \qquad F_{N,I}\left(\frac{1}{z}\right) = -F_{N,I}(z) \quad \textit{for} \quad |z| \neq 1.$$

(b) *For each I such that $1 \leq I \leq \infty$,*

$$\text{(3.6)} \qquad F_{\infty,I}(z) = \sum_{j=-I}^{I} |\alpha_j|^2 \frac{e^{i\omega_j} + z}{e^{i\omega_j} - z} \quad \textit{for} \quad |z| \neq 1.$$

Proof. It is readily seen that

$$\text{(3.7a)} \qquad \left|\frac{e^{i\theta} + z}{e^{i\theta} - z}\right| \leq \frac{1 + |z|}{|1 - |z||} \quad \text{for} \quad |z| \neq 1, \ \theta \in [-\pi, \pi],$$

$$\text{(3.7b)} \qquad \operatorname{Re}\left(\frac{e^{i\theta} + z}{e^{i\theta} - z}\right) > 0 \quad \text{for} \quad |z| < 1 \quad \text{and} \quad \operatorname{Re}\left(\frac{e^{i\theta} + z}{e^{i\theta} - z}\right) < 0 \quad \text{for} \quad |z| > 1.$$

First suppose that N and I satisfy $1 \leq N < \infty$ and $1 \leq I \leq \infty$. It follows from (1.3), (3.1), and well known properties of functions defined by integrals that $F_{N,I}(z)$ is holomorphic for $|z| \neq 1$. Equation (3.2) is an immediate consequence of (1.3) and (1.13). Assertion (3.3) follows from (1.3), (3.1), and (3.7b). The power series expansions in (3.4) are obtained from (1.13), substitution of

$$\frac{e^{i\theta} + z}{e^{i\theta} - z} = \begin{cases} 1 + 2\sum_{k=1}^{\infty} e^{-ik\theta} z^k, & |z| < 1, \\ -1 - 2\sum_{k=1}^{\infty} e^{ik\theta} z^{-k}, & |z| > 1, \end{cases}$$

into (3.1) and then term-by-term integration. To verify (3.5) we use (1.4), (3.1), and $\varphi := -\theta$ to obtain

$$F_{N,I}\left(\frac{1}{z}\right) = \int_{-\pi}^{\pi} \frac{e^{i\theta} + z^{-1}}{e^{i\theta} - z^{-1}} d\psi_{N,I}(\theta) = \int_{-\pi}^{\pi} \frac{z + e^{-i\theta}}{z - e^{-i\theta}} \psi'_{N,I}(\theta) d\theta$$

$$= -\int_{-\pi}^{\pi} \frac{e^{i\varphi} + z}{e^{i\varphi} - z} \psi'_{N,I}(\varphi) d\varphi = -F_{N,I}(z).$$

Next we consider $N = \infty$ and $1 \leq I \leq \infty$. The series representation (3.6) follows from (1.5) and (3.1). If $1 \leq I \leq \infty$, then (3.2), (3.3), (3.4), and (3.5) can be deduced from (1.14b), (3.6), and (3.7). □

For each N and I such that $[1 \le N < \infty$ and $1 \le I \le \infty]$ or $[N = \infty$ and $I = \infty]$, we consider the positive PC fraction (Perron–Carathéodory continued fraction)

(3.8a)
$$\delta_0^{(N,I)} - \frac{2\delta_0^{(N,I)}}{1} + \frac{1}{\delta_1^{(N,I)} z} + \frac{(1-|\delta_1^{(N,I)}|^2)z}{\delta_1^{(N,I)}} + \frac{1}{\delta_2^{(N,I)} z} + \frac{(1-|\delta_2^{(N,I)}|^2)z}{\delta_2^{(N,I)}} + \cdots,$$

where

(3.8b)
$$\delta_0^{(N,I)} := \mu_0^{(N,I)} = \sum_{m=0}^{N-1} [x_{N,I}(m)]^2 > 0.$$

The nth numerator $P_n(\psi_{N,I}; z)$ and denominator $Q_n(\psi_{N,I}; z)$ of (3.8) are defined by

(3.9a)
$$P_0(\psi_{N,I}; z) := -P_1(\psi_{N,I}; z) := \delta_0^{(N,I)}, \qquad Q_0(\psi_{N,I}; z) := Q_1(\psi_{N,I}; z) := 1,$$

(3.9b)
$$\begin{bmatrix} P_{2n}(\psi_{N,I}; z) \\ Q_{2n}(\psi_{N,I}; z) \end{bmatrix} = \delta_n^{(N,I)} z \begin{bmatrix} P_{2n-1}(\psi_{N,I}; z) \\ Q_{2n-1}(\psi_{N,I}; z) \end{bmatrix} + \begin{bmatrix} P_{2n-2}(\psi_{N,I}; z) \\ Q_{2n-2}(\psi_{N,I}; z) \end{bmatrix}, \qquad n \ge 1,$$

(3.9c)
$$\begin{bmatrix} P_{2n+1}(\psi_{N,I}; z) \\ Q_{2n+1}(\psi_{N,I}; z) \end{bmatrix} = \delta_n^{(N,I)} \begin{bmatrix} P_{2n}(\psi_{N,I}; z) \\ Q_{2n}(\psi_{N,I}; z) \end{bmatrix} + (1-|\delta_1^{(N,I)}|^2) z \begin{bmatrix} P_{2n-1}(\psi_{N,I}; z) \\ Q_{2n-1}(\psi_{N,I}; z) \end{bmatrix}, \qquad n \ge 1.$$

The close connection between Szegö polynomials and PC fractions (3.8) can be seen from

(3.10)
$$\rho_n(\psi_{N,I}; z) = Q_{2n+1}(\psi_{N,I}; z), \qquad \rho_n^*(\psi_{N,I}; z) = Q_{2n}(\psi_{N,I}; z), \quad n \ge 0,$$

which can be deduced by means of (1.19) and (3.9).

If $N = \infty$ and $1 \le I < \infty$, the corresponding PC fraction is the terminating one

(3.11a)
$$\delta_0^{(\infty,I)} - \frac{2\delta_0^{(\infty,I)}}{1} + \frac{1}{\delta_1^{(\infty,I)} z} + \frac{(1-|\delta_1^{(\infty,I)}|^2)z}{\delta_1^{(\infty,I)}} + \cdots + \frac{1}{\delta_{n_0(I)-1}^{(\infty,I)} z} + \frac{(1-|\delta_{n_0(I)-1}^{(\infty,I)}|^2)z}{\delta_{n_0(I)-1}^{(\infty,I)}} + \frac{1}{\delta_{n_0(I)}^{(\infty,I)} z},$$

where

(3.11b)
$$\delta_0^{(\infty,I)} := \mu_0^{(\infty,I)} = \sum_{j=-I}^{I} |\alpha_j|^2, \quad \delta_n^{(\infty,I)} := \rho_n(\psi_{\infty,I}; 0), \quad 1 \le n \le n_0(I), \quad |\delta_{n_0}^{(\infty,I)}| = 1.$$

The nth numerators and denominators of (3.11) are defined by (3.9) with $N = \infty$, $1 \le n \le 2n_0(I) + 1$. We recall [8, eq. (2.11)] that

(3.12)

$$\rho_{n_0(I)}(\psi_{\infty,I};z) = Q_{2n_0(I)+1}(\psi_{\infty,I};z) = (z-1)^L \prod_{j=1}^{I}(z-e^{i\omega_j})(z-e^{-i\omega_j})$$
$$= (-1)^L \rho^*_{n_0(I)}(\psi_{\infty,I};z) = (-1)^L Q_{2n_0(I)}(\psi_{\infty,I};z).$$

THEOREM 3.2. *For each I such that $1 \le I \le \infty$,*

$$\lim_{\substack{N\to\infty \\ k\to\infty}} \frac{1}{N}\frac{P_{2k}(\psi_{N,I};z)}{Q_{2k}(\psi_{N,I};z)} = F_{\infty,I}(z) = \sum_{j=-I}^{I} |\alpha_j|^2 \frac{e^{i\omega_j}+z}{e^{i\omega_j}-z} \quad \text{for } |z|<1, \tag{3.13a}$$

$$\lim_{\substack{N\to\infty \\ k\to\infty}} \frac{1}{N}\frac{P_{2k+1}(\psi_{N,I};z)}{Q_{2k+1}(\psi_{N,I};z)} = F_{\infty,I}(z) = \sum_{j=-I}^{I} |\alpha_j|^2 \frac{e^{i\omega_j}+z}{e^{i\omega_j}-z} \quad \text{for } |z|>1. \tag{3.13b}$$

Proof. From results given in [7, Thm. 3.2] and [11, Thm. 3.1] we know that for each z with $|z|<1$,

$$\left|\frac{P_{2k}(\psi_{N,I};z)}{Q_{2k}(\psi_{N,I};z)} - F_{N,I}(z)\right| \le \frac{4\delta_0^{(N,I)}|z|^{k+1}}{1-|z|^2}, \quad 1 \le N < \infty,\ 1 \le I \le \infty,\ k \ge 1. \tag{3.14}$$

By (1.1) and (3.8b) there exists a constant $A>0$, such that

$$0 < \delta_0^{(N,I)} \le AN \quad \text{for } 1 \le N < \infty \quad \text{and} \quad 1 \le I \le \infty. \tag{3.15}$$

It follows from (3.14) and (3.15) that, for $1 \le N < \infty$ and $1 \le I \le \infty$,

$$\left|\frac{1}{N}\frac{P_{2k}(\psi_{N,I};z)}{Q_{2k}(\psi_{N,I};z)} - F_{\infty,I}(z)\right| \le \frac{4A|z|^{k+1}}{1-|z|^2} + \left|\frac{1}{N}F_{N,I}(z) - F_{\infty,I}(z)\right|, \quad |z|<1. \tag{3.16}$$

By weak star convergence (Theorem 2.1) we have, for $1 \le I \le \infty$,

$$\lim_{N\to\infty}\frac{1}{N}F_{N,I}(z) = \lim_{N\to\infty}\frac{1}{N}\int_{-\pi}^{\pi}\frac{e^{i\theta}+z}{e^{i\theta}-z}d\psi_{(N,I)}(\theta) = \int_{-\pi}^{\pi}\frac{e^{i\theta}+z}{e^{i\theta}-z}d\psi_{(\infty,I)}(\theta) \tag{3.17}$$
$$= F_{\infty,I}(z), \qquad |z| \ne 1.$$

Now the assertion (3.13a) follows from (3.16) and (3.17). We can obtain (3.13b) from (3.13a), (3.5), and

$$\frac{P_{2k}(\psi_{N,I};z)}{Q_{2k}(\psi_{N,I};z)} = -\frac{P_{2k+1}(\psi_{N,I};z^{-1})}{Q_{2k+1}(\psi_{N,I};z^{-1})} \qquad \text{(see, e.g., [6]).} \qquad \square$$

4. Asymptotics of zeros.

THEOREM 4.1. *Let I be such that $2 \le I \le \infty$ and let $\{r_j\}_{-I}^{I}$ and $\{\omega_j\}_{-I}^{I}$ be sequences of real numbers satisfying* (1.1c, d, e). *For every discrete time signal $x_{N,I}$ of the form* (1.1) *let $\rho_n(\psi_{N,I};z)$ denote the monic nth degree Szegö polynomial* (1.17) *associated with $x_{N,I}$. Then for each K with $1 \le K < I$ the zeros $z_j(n_0(K),\psi_{N,I})$ of $\rho_{n_0(K)}(\psi_{N,I};z)$ can be arranged so that*

$$\lim_{\substack{N\to\infty \\ \alpha(K,I)\to 0}} z_j(n_0(K),\psi_{N,I}) = e^{i\omega_j}, \qquad j = \pm1, \pm2, \ldots, \pm K, \tag{4.1a}$$

and

$$\lim_{\substack{N\to\infty \\ \alpha(K,I)\to 0}} z_0(n_0(K),\psi_{N,I}) = e^{i0} = 1 \quad \text{if } L = 1 (\textit{i. e., } \alpha_0 > 0), \tag{4.1b}$$

where

$$\alpha(K,I) := \sum_{j=K+1}^{I} |\alpha_j|. \tag{4.1c}$$

Our proof of Theorem 4.1 makes use of the following lemmas.

LEMMA 4.2. *For each* $1 \le N < \infty$ *and* $1 \le I \le \infty$,

$$\mu_m^{(N,I)} = N\mu_m^{(\infty,I)} + M_m(N,I), \qquad m = 0, \pm 1, \pm 2, \ldots, \tag{4.2a}$$

where

$$M_m(N,I) := \sum_{\substack{j,n=-I \\ j\neq -n}}^{I} c_{j,n}(m,N)\alpha_j\alpha_n - m\mu_m^{(\infty,I)}, \tag{4.2b}$$

$$c_{j,n}(m,N) := e^{i\left[\frac{N-m-1}{2}\omega_j + \frac{N+m-1}{2}\omega_n\right]} \sin[(N-m)(\tfrac{\omega_j+\omega_n}{2})] \csc(\tfrac{\omega_j+\omega_n}{2}), \tag{4.2c}$$

and

$$|M_m(N,I)| \le M_m(I) := |m| \sum_{j=-I}^{I} r_j^2 + \sum_{j=-I}^{I} \sum_{\substack{n=-I \\ j\neq -n}}^{I} r_j r_n |\csc(\tfrac{\omega_j+\omega_n}{2})|, \tag{4.3}$$

$$M_m(I) \le M_m(I+1) \le M_m(\infty) < \infty, \qquad 1 \le I \le \infty, \quad m = 0, \pm 1, \pm 2, \ldots. \tag{4.4}$$

Proof. Substitute (1.1a) into (1.14a) and interchange the order of summation. Rearranging the terms and using (1.14b) gives (4.2). Equation (4.3) is a consequence of

$$c_{j,n}(m,N) = c_{n,j}(m,N), |c_{-j,-n}(m,N)| = |c_{j,n}(m,N)| \le c_{j,n} := |\csc(\tfrac{\omega_j+\omega_n}{2})|, \tag{4.5}$$

and (4.4) follows from (4.3) and (1.1e). □

LEMMA 4.3. *For* $2 \le I < \infty$,

$$\mu_m^{(\infty,I)} = \mu_m^{(\infty,I-1)} + 2|\alpha_I|^2 \cos(\omega_I m), \qquad m = 0, \pm 1, \pm 2, \ldots. \tag{4.6}$$

Proof. Apply (1.14b). □

LEMMA 4.4. *For* $1 \le N < \infty, 2 \le I < \infty$, *and* $m = 0, \pm 1, \pm 2, \ldots$,

$$\mu_m^{(N,I)} = \mu_m^{(N,I-1)} + |\alpha_I| U_m(N,I), \tag{4.7a}$$

where

$$U_m(N,I) = 2(N-m)|\alpha_I| \cos(\alpha_I m) + V_m(N,I), \tag{4.7b}$$

$$V_m(N,I) := 4 \sum_{j=-I+1}^{I} c_{j,I}(m,N)|\alpha_j| \cos(\varphi_I + \varphi_j) \quad (\varphi_0 := 0, \varphi_j := \operatorname{Arg} \alpha_j \text{for } j \neq 0), \tag{4.7c}$$

$$(4.8a)\quad |V_m(N,I)| \le 4 \sum_{j=-I+1}^{I} c_{j,I} r_j =: V(I) \le V(I+1) \le \lim_{I\to\infty} V(I) =: V(\infty) < \infty,$$

$$(4.8b)\quad |U_m(N,I)| \le \hat{U}_m(N,I) := 2(N+|m|)r_I + V(I),$$

$$(4.8c)\quad \hat{U}_m(N,I) \le 2(N+|m|) \max_{1\le j<\infty} r_j + V(\infty) =: \hat{U}_m(N,\infty) < \infty.$$

Proof. Apply (4.5) and Lemmas 4.2 and 4.3 to obtain (4.7). Then use (1.1e) and (4.7) to obtain (4.8). □

LEMMA 4.5. *For* $1 \le N < \infty$, $1 \le I < \infty$, $k \ge 1$, $m = 0, \pm 1, \pm 2, \ldots$, *let* $G_k^{(m)}(N,I)$ *be defined by*

$$(4.9)\qquad \frac{T_k^{(m)}(N,I)}{N^k} = T_k^{(m)}(\infty, I) + \frac{G_k^{(m)}(N,I)}{N}.$$

Then for each $k \ge 1$ *and* $m \in \mathbf{Z}$, *there exist numbers* $G_k^{(m)}(I)$ *and* $G_k^{(m)}(\infty)$ *such that, for all* $1 \le N < \infty$ *and all sequences* $\{\alpha_j\}_{-I}^{I}$ *satisfying* (1.1 c, d, e),

$$(4.10)\qquad |G_k^{(m)}(N,I)| \le G_k^{(m)}(I) \le G_k^{(m)}(\infty) < \infty.$$

Proof. Substituting (4.2a) into $T_k^{(m)}(N,I) := \det(\mu_{m+j-n}^{(N,I)})_{j,n=0}^{k-1}$, we see that $G_k^{(m)}(N,I)$ is a sum of terms, each a product of $(k-j)$ factors of the form $M_m(N,I)/N$ and j factors of the form $\mu_p^{(\infty,I)} = \sum_{j=-I}^{I} |\alpha_j|^2 e^{i\omega_j p}, 0 \le j \le k-1$. Since

$$|\mu_p^{(\infty,I)}| \le \sum_{j=-I}^{I} |\alpha_j|^2 \le \sum_{j=-\infty}^{\infty} |\alpha_j|^2,$$

it follows from Lemma 4.2 that each such product is of the form

$$H_{k,j}^{(m)}(N,I)/N^{k-j}, \quad \text{where } |H_{k,j}^{(m)}(N,I)| \le H_k^{(m)}(I) \le H_k^{(m)}(\infty),$$

$H_k^{(m)}(I)$ (and $H_k^{(m)}(\infty)$) being a number independent of j and N (and I). Summing these products yield (4.10). □

LEMMA 4.6. *For* $1 \le N < \infty$, $2 \le I < \infty$, $k \ge 1, m = 0, \pm 1, \pm 2, \ldots$, *let* $W_k^{(m)}(N,I)$ *be defined by*

$$(4.11)\qquad \frac{T_k^{(m)}(N,I)}{N^k} = \frac{T_k^{(m)}(N,I-1)}{N^k} + |\alpha_I| W_k^{(m)}(N,I).$$

Then for each $k \ge 1$ *and* $m \in \mathbf{Z}$, *there exist numbers* $W_k^{(m)}(I)$ *and* $W_k^{(m)}(\infty)$ *such that for all* $1 \le N < \infty$ *and all sequences* $\{\alpha_j\}_{-I}^{I}$ *satisfying* (1.1c, d, e),

$$(4.12)\qquad |W_k^{(m)}(N,I)| \le W_k^{(m)}(I) \le W_k^{(m)}(\infty) < \infty.$$

Proof. By substituting (4.7a) into $T_k^{(m)}(N,I) := \det(\mu_{m+j-n}^{(N,I)})_{j,n=0}^{k-1}$, one can show that $W_k^{(m)}(N,I)$ is a sum of terms, each a product of $(k-j)$ factors of the form

$|\alpha_I|U_p(N,I)/N$ and j factors of the form $\mu_p^{(N,I-1)}/N = \mu_p^{(\infty,I-1)} + M_p(N,I-1)/N$. By Lemmas 4.2 and 4.4 each such product is of the form

$$|\alpha_I|^{k-j}Y_{k,j}^{(m)}(N,I), \quad \text{where } |Y_{k,j}^{(m)}(N,I)| \le Y_k^{(m)}(I) \le Y_k^{(m)}(\infty),$$

$Y_k^{(m)}(I)$ (and $Y_k^{(m)}(\infty)$) being a number independent of j and N (and I). Summing these products yields (4.12). □

LEMMA 4.7. *For $1 \le N < \infty$ and $1 \le I < \infty$ let $\hat{\delta}_n(N,I)$ be defined by*

$$\delta_n^{(N,I)} = \delta_n^{(\infty,I)} + \frac{\hat{\delta}_n(N,I)}{N}, \qquad 1 \le n \le n_0(I). \tag{4.13}$$

Then there exist numbers $\hat{\delta}_n(I)$ and $\hat{\delta}_n(\infty)$ such that for all $1 \le N < \infty$ and all $\{\alpha_j\}$ satisfying (1.1c, d, e) *with α_1 fixed and $|\alpha_1| > 0$,*

$$|\hat{\delta}_n(N,I)| \le \hat{\delta}_n(I) \le \hat{\delta}_n(\infty) < \infty, \qquad 1 \le n \le n_0(I). \tag{4.14}$$

Proof. By (1.20), (4.13), and (4.9) with $k = n$

$$\begin{aligned}\frac{\hat{\delta}_n(N,I)}{N} &= \delta_n^{(N,I)} - \delta_n^{(\infty,I)} = (-1)^n\left[\frac{T_n^{(-1)}(N,I)}{T_n^{(0)}(N,I)} - \frac{T_n^{(-1)}(\infty,I)}{T_n^{(0)}(\infty,I)}\right] \\ &= \frac{(-1)^n}{N}\frac{T_n^{(0)}(\infty,I)G_n^{(-1)}(N,I) - T_n^{(-1)}(\infty,I)G_n^{(0)}(N,I)}{N^{-n}T_n^{(0)}(N,I)T_n^{(0)}(\infty,I)} \\ &= \frac{(-1)^n}{N}\frac{T_n^{(0)}(\infty,I)G_n^{(-1)}(N,I) - T_n^{(-1)}(\infty,I)G_n^{(0)}(N,I)}{(T_n^{(0)}(\infty,I) + \frac{G_n^{(0)}(N,I)}{N})T_n^{(0)}(\infty,I)}.\end{aligned}$$

The denominator is positive.

Furthermore, $T_n^{(0)}(\infty,I)$ has a positive lower bound for all $0 \le |\alpha_j| \le r_j$, $j = 2,3,\ldots$ and all $1 \le I \le \infty$ (continuity and compactness).

Thus, by Lemma 4.5 we get

$$|\hat{\delta}_n(N,I)| \le \frac{T_n^{(0)}(\infty,I)G_n^{(-1)}(I) + |T_n^{(-1)}(\infty,I)|G_n^{(0)}(I)}{T_n^{(0)}(\infty,I)\inf_{N\ge1}(T_n^{(0)}(\infty,I) + \frac{G_n^{(0)}(N,I)}{N})} =: \hat{\delta}_n(I)$$

and

$$\begin{aligned}\hat{\delta}_n(I) &\le \frac{G_n^{(-1)}(\infty)\sup_{I\ge\frac{n-L}{2}} T_n^{(0)}(\infty,I) + G_n^{(0)}(\infty)\sup_{I\ge\frac{n-L}{2}}|T_n^{(-1)}(\infty,I)|}{\inf_{\substack{I\ge\frac{n-L}{2}\\ N\ge1}}\left[T_n^{(0)}(\infty,I)(T_n^{(0)}(\infty,I) + \frac{G_n^{(0)}(N,I)}{N})\right]} \\ &:= \hat{\delta}_n(\infty) < \infty.\end{aligned}$$

(Remark: The I-condition is equivalent to the condition $n_0(I) \ge n$.)

LEMMA 4.8. *For $1 \le N < \infty$, $2 \le I < \infty$ and $n \ge 1$, let $\tilde{\delta}_n(N,I)$ be defined by*

$$\delta_n(N,I) = \delta_n^{(N,I-1)} + |\alpha_I|\tilde{\delta}_n(N,I). \tag{4.15}$$

Then there exist numbers $\tilde{\delta}_n(I)$ and $\tilde{\delta}_n(\infty)$ such that for all $1 \le N < \infty$ and all $\{\alpha_j\}$ satisfying (1.1c, d, e) *with α_1 fixed and $|\alpha_1| > 0$,*

$$|\tilde{\delta}_n(N,I)| \le \tilde{\delta}_n(I) \le \tilde{\delta}_n(\infty), \qquad n = 1,2,3,\ldots. \tag{4.16}$$

An argument for proving Lemma 4.8 can be given that it is completely analogous to that given for Lemma 4.7. Hence it is omitted.

LEMMA 4.9. *For* $1 \le N < \infty$, $1 \le I < \infty$, $1 \le n \le n_0(I)$, $0 \le j \le n$, *let* $\hat{q}_j^{(n,N,I)}$ *be defined by*

(4.17)
$$q_j^{(n,N,I)} = q_j^{(n,\infty,I)} + \frac{\hat{q}_j^{(n,N,I)}}{N} \qquad \left(\rho_n(\psi_{N,I}; z) := \sum_{j=0}^{n} q_j^{(n,N,I)} z^j,\ 1 \le N \le \infty\right).$$

Then there exist numbers $\hat{q}_j(n,I)$ *and* $\hat{q}_j(n,\infty)$ *such that for all* $1 \le N < \infty$ *and all* $\{\alpha_j\}$ *satisfying* (1.1 c, d, e) *with* α_1 *fixed and* $|\alpha_1| > 0$,

$$|\hat{q}_j^{(n,N,I)}| \le \hat{q}_j(n,I) \le \hat{q}_j(n,\infty), \quad 0 \le j \le n, \quad 1 \le n \le n_0(I), \quad 1 \le I < \infty. \tag{4.18}$$

Proof. From the recurrence relations (1.19) we obtain the recurrence relations

$$q_0^{(n,N,I)} = \delta_n^{(N,I)}, \quad q_j^{(n,N,I)} = q_{j-1}^{(n-1,N,I)} + \delta_n^{(N,I)} q_{n-1-j}^{(n-1,N,I)} \quad \text{for } 1 \le j \le n-1,\ q_n^{(n,N,I)} = 1, \tag{4.19}$$

valid for all $1 \le N < \infty$, $n \ge 1$. The relations (4.19) are also valid for $n = \infty, 1 \le I < \infty$, $1 \le n \le n_0(I)$. Our proof is by induction on n. First we have, since $\delta_n^{(N,I)} = \rho_n(\psi_{N,I}; 0)$,

$$q_0^{(n,N,I)} = \delta_n^{(N,I)} = \delta_n^{(\infty,I)} + \frac{\hat{\delta}_n(N,I)}{N} = q_0^{(n,\infty,I)} + \frac{\hat{q}_0^{(n,N,I)}}{N},$$

where $\hat{q}_0^{(n,N,I)} = \hat{\delta}_n(N,I)$, hence by Lemma 4.7,

$$|\hat{q}_0^{(n,N,I)}| \le |\hat{\delta}(N,I)| \le \hat{\delta}_n(I) =: \hat{q}_0(n,I) \le \hat{\delta}_n(\infty) =: \hat{q}_0(n,\infty).$$

Therefore, for $j = 0$ Lemma 4.9's assertions hold for all $n \ge 1$. For $n = 1$ the assertion in (4.18) also holds for $j = 1$, since $q_1^{(1,N,I)} = q_1^{(1,\infty,I)} = 1$.

We assume (induction hypothesis) that for some integer $n \ge 1$, there exist numbers $\hat{q}_j(m,I)$ and $\hat{q}_j(m,\infty)$ such that for all $1 \le N < \infty$ and $\{\alpha_j\}$ satisfying (1.1c, d, e) and α_1 fixed $|\alpha_1| > 0$,

$$|\hat{q}_j^{(m,N,I)}| \le \hat{q}_j(m,I) \le \hat{q}_j(m,\infty), \qquad 1 \le j \le m, \quad 1 \le m \le n-1. \tag{4.20}$$

Then, by (4.19), (4.20), (4.17), and (4.13),

$$q_j^{(n,N,I)} = \left(q_{j-1}^{(n-1,\infty,I)} + \frac{\hat{q}_j^{(n-1,N,I)}}{N}\right) + \left(\delta_n^{(\infty,I)} + \frac{\hat{\delta}_n(N,I)}{N}\right)\left(q_{n-1-j}^{(n-1,\infty,I)} + \hat{q}_{n-1-j}^{(n-1,N,I)}\right)$$

$$= q_j^{(n,\infty,I)} + \frac{\hat{q}_j^{(n,N,I)}}{N},$$

where

$$|\hat{q}_j^{(n,N,I)}| = \left|\hat{q}_j^{(n-1,N,I)} + \delta_n^{(\infty,I)}\hat{q}_{n-1-j}^{(n-1,N,I)} + \hat{\delta}_n(N,I)q_{n-1-j}^{(n-1,\infty,I)} + \frac{1}{N}\hat{\delta}_n(N,I)\hat{q}_{n-1-j}^{(n-1,N,I)}\right|$$

$$\leq |\hat{q}_j^{(n-1,N,I)}| + |\delta_n^{(\infty,I)}\hat{q}_{n-1-j}^{(n-1,N,I)}| + |\hat{\delta}_n(N,I)|\left[|q_{n-1-j}^{(n-1,\infty,I)}| + \frac{|\hat{q}_{n-1-j}^{(n-1,N,I)}|}{N}\right]$$

$$\leq \hat{q}_j(n-1,I) + |\delta_n^{(\infty,I)}|\hat{q}_{n-1-j}(n-1,I)$$

$$+\hat{\delta}_n(I)\left[|q_{n-1-j}^{(n-1,\infty,I)}| + \hat{q}_{n-1-j}(n-1,I)\right] =: \hat{q}_j(n,I)$$

$$\leq \hat{q}_j(n-1,\infty) + \hat{q}_{n-1-j}(n-1,\infty)$$

$$+\hat{\delta}_n(\infty)\left[\sup_{1\leq I<\infty} q_{n-1-j}^{(n-1,\infty,I)}| + \hat{q}_{n-1-j}(n-1,\infty)\right] =: \hat{q}_j(n,\infty).$$

Here we have used the fact that $|\delta_n^{(n,\infty)}| \leq 1$ and $\lim_{I\to\infty}\hat{q}_{n-1-j}^{(n-1,\infty,I)} = \hat{q}_{n-1-j}^{(n-1,\infty,\infty)}$ exists and hence $\sup_{1\leq I<\infty}|q_{n-1-j}^{(n-1,\infty,I)}| < \infty$. It follows that (4.20) also holds for $m = n$. □

LEMMA 4.10. *For $1 \leq N < \infty$, $2 \leq I < \infty$ and $n \geq 1$, let $\tilde{q}_j^{(n,N,I)}$ be defined by*

$$q_j^{(n,N,I)} = q_j^{(n,N,I-1)} + |\alpha_I|\,\tilde{q}_j^{(n,N,I)}, \qquad 0 \leq j \leq n. \tag{4.21}$$

Then there exist numbers $\tilde{q}_j(n,I)$ and $\tilde{q}_j(n,\infty)$ such that for all $1 \leq N < \infty$ and all $\{\alpha_j\}$ satisfying (1.1 c, d, e) *with α_1 fixed, $|\alpha_1| > 0$,*

$$|\tilde{q}_j^{(n,N,I)}| \leq \tilde{q}_j(n,I) \leq \tilde{q}_j(n,\infty), \qquad 0 \leq j \leq n, \quad 1 \leq I < \infty. \tag{4.22}$$

A proof for this lemma can be given that it is completely analogous to that given for Lemma 4.9. Hence it is omitted.

LEMMA 4.11. *Let $x_{N,I}$ denote a discrete time signal of the form* (1.1), *where $1 \leq N < \infty$ and $1 \leq I \leq \infty$. Let $\rho_n(\psi_{N,I};z)$ denote the monic nth degree Szegö polynomial* (1.11) *associated with $x_{N,I}$. Let $\hat{\rho}_n(\psi_{N,I};z)$ and $\tilde{\rho}_n(\psi_{N,I};z)$ be defined, for $1 \leq N < \infty, 2 \leq I < \infty, 1 \leq n \leq n_0(I)$, by*

$$\begin{aligned}\rho_n(\psi_{N,I};z) &= \rho_n(\psi_{\infty,I};z) + \tfrac{\hat{\rho}_n(\psi_{N,I};z)}{N},\\ \rho_n(\psi_{N,I};z) &= \rho_n(\psi_{N,I-1};z) + |\alpha_I|\,\tilde{\rho}_n(\psi_{N,I};z).\end{aligned} \tag{4.23}$$

(a) *Then for all $0 < R < \infty$ there exist numbers $\hat{\rho}_n(I,R), \hat{\rho}_n(\infty,R), \tilde{\rho}_n(I,R)$, and $\tilde{\rho}_n(\infty,R)$ such that for all $1 \leq N < \infty$ and all $\{\alpha_j\}$ satisfying* (1.1 c, d, e) *with α_1 fixed, $|\alpha_1| > 0$,*

$$|\hat{\rho}_n(\psi_{N,I};z)| \leq \hat{\rho}_n(I,R) \leq \hat{\rho}_n(\infty,R), \qquad |\tilde{\rho}_n(\psi_{N,I};z)| \leq \tilde{\rho}_n(I,R) \leq \tilde{\rho}_n(\infty,R) \tag{4.24}$$

for $2 \leq I < \infty$, $1 \leq n \leq n_0(I)$, $|z| \leq R$.

(b) *For all I, K, and R such that $2 \leq I \leq \infty, 1 \leq K < I$, there exist positive numbers $\hat{\rho}_{n_0(K)}(R)$ and $\tilde{\rho}_{n_0(K)}(R)$ such that for $1 \leq N < \infty$ and $|z| \leq R$,*

$$\left|\rho_{n_0(K)}(\psi_{N,I};z)-(z-1)^L\prod_{j=1}^{K}(z-e^{i\omega_j})(z-e^{-i\omega_j})\right| \tag{4.25}$$
$$\le \tilde{\rho}_{n_0(K)}(\infty,R)\sum_{j=K+1}^{I}|\alpha_j|+\frac{1}{N}\hat{\rho}_{n_0(K)}(R).$$

Remark. The assertion involving $\hat{\rho}_n(\psi_{N,I};z)$ holds also for $I=\infty$. This can be seen by letting $I\to\infty$ in the first part of (4.23) and (4.24) and using the fact that $\lim_{I\to\infty}\rho_n(\psi_{N,I};z)=\rho_n(\psi_{N,\infty};z)$.

Proof. (a) By Lemma 4.9 we have for $|z|\le R$ and $1\le N<\infty$, $1\le n\le n_0(I)$,

$$|\rho_n(\psi_{N,I};z)-\rho_n(\psi_{\infty,I};z)|=\left|\sum_{j=0}^{n}(q_j^{(n,N,I)}-q_j^{(n,\infty,I)})z^j\right|\le\frac{1}{N}\sum_{j=0}^{n}|\hat{q}_j^{(n,N,I)}|R^j$$
$$\le\frac{1}{N}\sum_{j=0}^{n}\hat{q}_j(n,I)R^j=:\frac{\hat{\rho}_n(I,R)}{N}$$
$$\le\frac{1}{N}\sum_{j=0}^{n}\hat{q}_j(n,\infty)R^j=:\frac{\hat{\rho}_n(\infty,R)}{N}.$$

A similar proof for the second part of (a) can be given and hence is omitted.

(b) First we assume that $2\le I<\infty$. By Lemma 4.11(a) we obtain for $|z|\le R$ and $1\le N<\infty$,

$$\left|\rho_{n_0(K)}(\psi_{N,I};z)-(z-1)^L\prod_{j=1}^{K}(z-e^{i\omega_j})(z-e^{-i\omega_j})\right|$$
$$=\left|\rho_{n_0(K)}(\psi_{N,I};z)-\rho_{n_0(K)}(\psi_{\infty,K};z)\right|$$
$$\le\sum_{j=K+1}^{I}\left|\rho_{n_0(K)}(\psi_{N,j};z)-\rho_{n_0(K)}(\psi_{N,j-1};z)\right|$$
$$+\left|\rho_{n_0(K)}(\psi_{N,K};z)-\rho_{n_0(K)}(\psi_{\infty,K};z)\right|$$
$$=\sum_{j=K+1}^{I}|\alpha_j|\,|\tilde{\rho}_{n_0(K)}(\psi_{N,j};z)|+\frac{1}{N}|\hat{\rho}_{n_0(K)}(\psi_{N,K};z)|$$
$$\le\sum_{j=K+1}^{I}|\alpha_j|\tilde{\rho}_{n_0(K)}(j,R)+\frac{1}{N}\hat{\rho}_{n_0(K)}(K,R)$$
$$\le\tilde{\rho}_{n_0(K)}(\infty,R)\sum_{j=K+1}^{I}|\alpha_j|+\frac{1}{N}\hat{\rho}_{n_0(K)}(\infty,R).$$

To see that this result also holds for $I=\infty$ we let $I\to\infty$ in (4.25) and use the fact that $\lim_{I\to\infty}\rho_{n_0(K)}(\psi_{N,I};z)=\rho_{n_0(K)}(\psi_{N,\infty};z)$. □

Proof of Theorem 4.1. By Lemma 4.11(b) the functions $\rho_{n_0(K)}(\psi_{N,I}; z)$ converge to $(z-1)^L \prod_{j=1}^{K}(z - e^{i\omega_j})(z - e^{-i\omega_j})$ uniformly on each compact set $|z| \leq R$ as $N \to \infty$ and $\alpha(K, I) \to 0$. The assertions (4.1) follow from this and an application of Hurwitz's theorem [2, Thm. 14.3.4]. □

5. Computional results. To illustrate the approximations given by the Szegö polynomial method, we consider signals $x_{N,4}$ of the form (1.4), where $I = 4$ and the α_j and ω_j are as follows in Table 1.

TABLE 1

j	0	1	2	3	4
ω_j	0	$\frac{\pi}{4}$	$\frac{\pi}{6}$	$\frac{\pi}{2}$	$\frac{5\pi}{6}$
α_j	0	100	1	1	1

We let $z_1(k, \psi_{N,4})$ denote the zero of $\rho_k(\psi_{N,4}; z)$ nearest to $e^{i\omega_1}$, where $\rho_k(\psi_{N,4}; z)$ is the Szegö polynomial (1.17) associated with $x_{N,4}$. Values of $z_1(k, \psi_{N,4})$ were computed for the degrees $k = 2, 4, 6, 8, 10, 20, 30, 40, 50$ and for various sample sizes N ranging from $N = 100$ to $N = 3000$. Values of Arg $z_1(k, \psi_{N,4})$ used to approximate $\omega_1 = \pi/4 \doteq 0.785398164$ (for $k \leq 8$ *not* meaning convergence) are given in Table 2a and values of the number of significant digits of these approximations are given in Table 2b. Values of $|z_1(k, \psi_{N,4}) - e^{i\omega_1}|$ are given in Table 3. For these examples $n_0(I) = n_0(4) = 8$, since $L = 0$ (i.e., $\alpha_0 = 0$). By the convergence results (1.8) stated in §1, for all $k \geq n_0(4) = 8$, $z_1(k, \psi_{N,4}) \to e^{i\omega_1}$ and hence Arg $z_1(k, \psi_{N,4}) \to \pi/4$ as $N \to \infty$. The numerical results in Tables 2 and 3 are consistent with this. For degree $k < n_0(4) = 8$ the method gives approximations of $\omega_1 = \pi/4$ with significant digits ranging from 1 to 4 depending on the choices of k and N. For each fixed N, as k takes on the values 2, 4, 6, the significant digits (Table 2b) are nondecreasing. Moreover, for each fixed $k(= 2, 4, 6)$ the significant digits in Table 2b are nondecreasing as N increases from 100 to 1000, with one exception at $k = 6$ and $N = 401$. These results are all consistent with the assertions of Theorem 4.1.

TABLE 2a

Values of Arg $z_1(k, \psi_{N,4})$, *where* $z_1(k, \psi_{N,4})$ *is the zero of* $\rho_k(\psi_{N,4}; z)$ *nearest to the critical point* $e^{i\omega_1}, \omega_1 = \frac{\pi}{4} \doteq 0.785398164$. *Significant digits are underlined.*

$k \backslash N$	100	202	401	601	1000
2	0.8054 40	0.7905 98	0.7906 49	0.7891 07	0.7876 92
4	0.7905 81	0.7860 77	0.7866 15	0.7862 13	0.7858 77
6	0.7855 29	0.7852 78	0.7853 83	0.7853 43	0.7853 29
8	0.7841 59	0.7855 21	0.7851 59	0.7852 68	0.7853 00
10	0.7859 11	0.7854 33	0.7854 27	0.7854 49	0.7854 10
20	0.7849 56	0.7852 83	0.7853 22	0.7853 21	0.7853 70
30	0.7860 54	0.7853 17	0.7853 81	0.7853 48	0.7853 39
40	0.7847 45	0.7854 89	0.7853 95	0.7854 33	0.7854 26
50	0.7713 08	0.7894 09	0.7853 87	0.7853 59	0.7853 73

TABLE 2b

Number of significant digits in the approximation of $\omega_1 = \pi/4 \doteq 0.785398164$ *by* Arg $z_1(k, \psi_{N,4})$ *where* $z_1(k, \psi_{N,4})$ *is the zero of* $\rho_k(\psi_{N,4}; z)$ *nearest to the critical point* $e^{i\omega_1}$.

$k\backslash N$	100	202	401	601	1000
2	1	2	2	2	2
4	2	2	2	2	2
6	2	3	4	3	3
8	1	2	3	3	3
10	2	4	4	4	4
20	1	3	3	3	4
30	2	3	4	3	3
40	1	3	5	4	4
50	1	2	4	4	4

TABLE 3

Values of $|z_1(k, \psi_{N,4}) - e^{i\omega_1}|$, *where* $\omega_1 \doteq \frac{\pi}{4} = 0.785398164$ *and* $z_1(k, \psi_{N,4})$ *is the zero of* $\rho_k(\psi_{N,4}; z)$ *nearest to* $e^{i\omega_1}$. N = *sample size of observed signal* $x_{N,4}$; k = *degree of the Szegö polynomial* $\rho_k(\psi_{N,4}; z)$.

$k\backslash N$	100	202	401	601	1000	2000	3000
2	0.03591	0.01167	0.00957	0.00684	0.00426		
4	0.01034	0.00477	0.00247	0.00168	0.00099		
6	0.00829	0.00521	0.00196	0.00134	0.00080		
8	0.01104	0.00521	0.00258	0.00175	0.00106	0.00055	0.00037
10	0.01175	0.00522	0.00271	0.00184	0.00112	0.00058	0.00040
20	0.01265	0.00540	0.00259	0.00166	0.00098	0.00047	0.00031
30	0.01523	0.00575	0.00268	0.00181	0.00106	0.00055	0.00037
40	0.01835	0.00617	0.00275	0.00173	0.00102	0.00048	0.00032
50	0.02699	0.00683	0.00293	0.00188	0.00107	0.00052	0.00035

Acknowledgments. The authors wish to thank Anne C. Jones for able assistance in computing the numerical examples.

REFERENCES

[1] F. B. HILDEBRAND, *Introduction to Numerical Analysis*, McGraw-Hill Inc., New York, 1956.

[2] E. HILLE, *Analytic Function Theory, Vol.* II, Ginn and Company, Boston, 1962.

[3] W. B. JONES AND O. NJÅSTAD, *Applications of Szegö polynomials to digital signal processing*, Rocky Mountain Math., 21 (1991), pp. 387–436.

[4] W. B. JONES, O. NJÅSTAD, AND E. B. SAFF, *Szegö polynomials associated with Wiener-Levinson filters*, J. Comp. Appl. Math., 32 (1990), pp. 387–406.

[5] W. B. JONES, O. NJÅSTAD, AND W. J. THRON, *Schur fractions, Perron-Carathéodory fractions and Szegö polynomials, a survey*, Analytic Theory of Continued Fraction II, W. J. Thron, ed., Lecture Notes in Math. 1199, Springer-Verlag, New York, 1986, pp. 127–158.

[6] ———, *Moment theory, orthogonal polynomials, quadrature, and continued fractions associated with the unit circle*, Bull. London Math. Soc., 21 (1989), pp. 113–152.

[7] ———, *Continued fractions associated with trigonometric and other strong moment problems*, Constr. Approx., 2 (1986), pp. 197–211.

[8] W. B. JONES, O. NJÅSTAD, W. J. THRON, AND H. WAADELAND, *Szegö polynomials applied to frequency analysis*, Journal CAM J. Comp. Appl. Math., 46 (1993), pp. 217–228.

[9] W. B. JONES, O. NJÅSTAD, AND H. WAADELAND, *Asymptotics for Szegö polynomial zeros*, Numerical algorithms, 3 (1992), pp. 255–264.

[10] W. B. JONES AND E. B. SAFF, *Szegö polynomials and frequency analysis*, Approximation Theory, G. Anastassiou, ed., Marcel Dekker, Inc., New York, 1992, pp. 341–352.

[11] W. B. JONES AND W. J. THRON, *A constructive proof of convergence of the even approximants of positive PC-fractions*, Rocky Mountain Math., 19 (1989), pp. 199–210.

[12] R. KUMARESAM, L. L. SCHARF, AND A. K. SHAW, *An algorithm for pole-zero modeling and spectral analysis*, IEEE Trans ASSP, 34 (1986), pp. 637–640.

[13] N. LEVINSON, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[14] J. D. MARKEL AND A. N. GRAY, JR., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[15] H. N. MHASKAR AND E. B. SAFF, *On the distribution of zeros of polynomials orthogonal on the unit circle*, J. Approx. Theory, 63 (1990), pp. 30–38.

[16] M. E. MUNROE, *Introduction to Measure and Integration*, Addison-Wesley, Cambridge, MA, 1953.

[17] P. NEVAI AND V. TOTIK, *Orthogonal polynomials and their zeros*, Acta Sci. Math., 53 (1989), pp. 99–104.

[18] K. PAN, *Asymptotics for Szegö polynomials associated with Wiener-Levinson filters*, J. Comp. Appl. Math., 46 (1993), pp. 387–394.

[19] K. PAN AND E. B. SAFF, *Asymptotics for zeros of the Szegö polynomials associated with trigonometric polynomial signals*, J. Approx. Theory, 71 (1992), pp. 239–251.

[20] A. K. PAUL, *Anharmonic frequency analysis*, Math. Comp., 26 (1972), pp. 437–447.

[21] L. L. SCHARF, *Statistical Signal Processing*, Addison-Wesley, Reading, MA, 1991.

[22] G. YE. SHILOV, *Mathematical Analysis*, J. D. Davis, trans., Pergamon Press, New York, 1965.

[23] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Coll. Pub. 23, fourth ed., Providence, RI, 1975.

[24] N. WIENER, *Extrapolation, interpolation and smoothing of stationary time series*, published jointly by the Technology Press of Massachusetts Institute of Technology, Cambridge, MA and John Wiley and Sons Inc., New York, 1949.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 513–527, March 1994

018

MODELS OF Q-ALGEBRA REPRESENTATIONS: THE GROUP OF PLANE MOTIONS*

E. G. KALNINS†, W. MILLER, JR.‡, AND S. MUKHERJEE§

Abstract. This paper continues a study of one- and two-variable function space models of irreducible representations of q-analogs of Lie enveloping algebras, motivated by recurrence relations satisfied by q-hypergeometric functions. The algebra considered is the Lie algebra $m(2)$ of the group of plane motions. It is shown that various q-analogs of the exponential function can be used to mimic the exponential mapping from a Lie algebra to its Lie group, and the corresponding matrix elements of the group operators are computed on these representation spaces. This local approach applies to more general families of special functions, e.g., those with complex arguments and parameters, than does the quantum group approach. A simple one-variable model of the infinite-dimensional irreducible representations is used to compute the Clebsch–Gordan coefficients for $m(2)$ considered as a true quantum algebra. The authors derive a generalization of Koelink's addition formula for Hahn–Exton q-Bessel functions. It is interpreted here as the expansion of the matrix elements of a group operator in a tensor product basis in terms of the matrix elements in a reduced basis.

Key words. basic hypergeometric functions, q-algebras, quantum groups, motion group, Clebsch–Gordan series

AMS subject classifications. 33A75, 33A65, 20N99

1. Introduction.

This paper continues the study of function space models of irreducible representations of q-algebras [8]. These algebras and models are motivated by recurrence relations satisfied by q-hypergeometric functions [2], [9]. Here, we consider the irreducible representations of the Lie algebra $m(2)$ of the group $M(2)$ of plane motions. We replace the usual exponential-function mapping from the Lie algebra to the Lie group by the q-exponential mappings E_q and e_q. In place of the usual matrix elements on the group (arising from an irreducible representation), which are expressible in terms of Bessel functions of integer order, we find four types of matrix elements expressible in terms of the Jackson and the Hahn–Exton q-Bessel functions. These q-matrix elements do not satisfy group homomorphism properties, and so they do not lead to addition theorems in the usual sense. However, they do satisfy orthogonality relations. (This was shown earlier by Koornwinder and Swarttouw [15].) Furthermore, in analogy with true group representation theory, we can show that each of the four families of matrix elements determines a two-variable model for irreducible representations of $m(2)$. By q-exponentiating these models we get q-analogs of addition theorems for Bessel functions.

In §3 we use the definition of $m(2)$ as a true quantum algebra and take the tensor product of two infinite-dimensional unitary irreducible representations of this quantum algebra. The tensor product decomposes into a direct sum of irreducible

*Received by the editors January 14, 1992; accepted for publication (in revised form) December 3, 1992.

†Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand.

‡School of Mathematics and Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported in part by National Science Foundation grant DMS 91-100324.

§Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported by the Study Abroad fellowship of the Government of India.

representations (rather than a direct integral, as in the $q = 1$ case), and the decomposition is nonunique. Focusing our attention on the two simplest decompositions of the infinite-parameter family of choices, we compute the corresponding Clebsch–Gordon coefficients. Special cases of the unitarity relations for the Clebsch–Gordon coefficients are the q-Hankel orthogonality relations. Moreover, expressing the matrix elements of the group operators in the tensor-product basis as a linear combination of the matrix elements in the basis adapted to the direct sum decomposition, by means of the Clebsch–Gordon coefficients, we obtain a generalization of Koelink's addition theorem for Hahn–Exton q-Bessel functions.

Our approach to the derivation and understanding of q-series identities is based on the study of q-algebras as q-analogs of Lie algebras [7]. Essentially, we are attempting to find q-analogs of the theory relating Lie algebra and local Lie transformation groups. A similar approach has been adopted by Floreanini and Vinet [3]–[5]. This is an alternative to the elegant papers [10]–[17] that are based primarily on the theory of quantum groups. The main justification of the local approach is that it is more general; it applies to more general families of special functions than does the quantum group approach.

The notation used for q-series in this paper follows that of Gasper and Rahman [6]. We wish to thank the referee for suggestions that considerably improved the exposition and accuracy of §2.

2. Matrix elements of $m(2)$ representations. The three-dimensional Lie algebra $m(2)$ is determined by its generators H, E_+, E_-, which obey the commutation relations

$$[H, E_+] = E_+, \qquad [H, E_-] = -E_-, \qquad [E_+, E_-] = 0. \tag{2.1}$$

The irreducible representations $Q(\omega, m_0)$ are characterized by the complex numbers ω and m_0, with $\omega \neq 0$ and $0 \leq \Re\, m_0 < 1$. The spectrum of H corresponding to $Q(\omega, m_0)$ is the set $S = \{m_0 + n : n \in Z\}$, and the complex representation space has basis vectors f_m, $m \in S$, such that

$$E_\pm f_m = \omega f_{m\pm 1}, \quad Hf_m = mf_m, \quad E_+E_-f_m = \omega^2 f_m, \tag{2.2}$$

where $C \equiv E_+E_-$ is an invariant operator. Note that the representations $Q(\omega, m_0)$ and $Q(-\omega, m_0)$ are equivalent.

A simple realization of $Q(\omega, m_0)$ is given by the operators

$$H = m_0 + z\frac{d}{dz}, \quad E_+ = \omega z, \quad E_- = \frac{\omega}{z} \tag{2.3}$$

acting on the space of all linear combinations of the functions z^n, z a complex variable, $n \in Z$, with basis vectors $f_m(z) = z^n$, where $m = m_0 + n$.

The representations $(\omega) \cong Q(\omega, 0)$ with $\omega > 0$ are of special interest. In this case we can introduce an inner product such that $\langle f_n, f_{n'} \rangle = \delta_{nn'}$, $n, n' \in Z$. On the dense subspace $\mathcal{K}$ of all finite linear combinations of the basis vectors we have

$$\langle E_+f, f' \rangle = \langle f, E_-f' \rangle, \quad \langle Hf, f' \rangle = \langle f, Hf' \rangle \tag{2.4}$$

for all $f, f' \in \mathcal{K}$, so that $H = H^*$ and $E_+^* = E_-$. In terms of the operators (2.3) we can obtain a realization of (ω) and its Hilbert-space structure by setting $z = e^{i\theta}$:

$$
\begin{aligned}
&H = -i\frac{d}{d\theta}, \quad E_+ = \omega e^{i\theta}, \quad E_- = \omega e^{-i\theta}, \\
&f_n(z) = e^{in\theta}, \qquad \langle f, f' \rangle = \frac{1}{2\pi}\int_0^{2\pi} f(e^{i\theta})\overline{f'(e^{i\theta})}\, d\theta.
\end{aligned} \tag{2.5}
$$

Matrix elements $T_{m'm}$ of the complex motion group in the representation $Q(\omega, m_o)$ are typically defined by the expansions

$$
e^{\beta E_+} e^{\alpha E_-} e^{\tau H} f_m = \sum_{m'=-\infty}^{\infty} T_{m'm}(\alpha, \beta, \tau) f_{m'}, \tag{2.6}
$$

with $m, m' \in S$ [3], [5], [19]. The group multiplication property of the operators on the left-hand side of (2.6) leads to addition theorems for the matrix elements. For convenience in the computations to follow we shall limit ourselves to the case in which $\tau = 0$ and ω is real. For this case we have $T_{m'm}(\alpha, \beta) \equiv T_{m'm}(\alpha, \beta, 0) \equiv T_{n'n}(\alpha, \beta)$, where $m = m_0 + n$, $m' = m_0 + n'$, i.e., the representations $Q(\omega, m_0)$ and (ω) have exactly the same matrix elements. (The extension of the matrix elements to complex ω will usually be obvious.)

With the q-analogs of the exponential function

$$
\begin{aligned}
e_q(x) &= \sum_{k=0}^{\infty} \frac{x^k}{(q;q)_k} = \frac{1}{(x;q)_\infty}, \qquad |x| < 1, \\
E_q(x) &= \sum_{k=0}^{\infty} \frac{q^{k(k-1)/2} x^k}{(q;q)_k} = (-x;q)_\infty
\end{aligned} \tag{2.7}
$$

we use the model (2.3) to define the following q-analogs of matrix elements of (ω):

$$
e_q(\beta E_+) e_q(\alpha E_-) f_n = \sum_{n'=-\infty}^{\infty} T_{n'n}^{(e,e)}(\alpha, \beta) f_{n'}, \qquad |\omega\alpha|, |\omega\beta| < 1, \tag{2.8a}
$$

$$
e_q(\beta E_+) E_q(\alpha E_-) f_n = \sum_{n'=-\infty}^{\infty} T_{n'n}^{(e,E)}(\alpha, \beta) f_{n'}, \qquad |\omega\beta| < 1, \tag{2.8b}
$$

$$
E_q(\beta E_+) e_q(\alpha E_-) f_n = \sum_{n'=-\infty}^{\infty} T_{n'n}^{(E,e)}(\alpha, \beta) f_{n'}, \qquad |\omega\alpha| < 1, \tag{2.8c}
$$

$$
E_q(\beta E_+) E_q(\alpha E_-) f_n = \sum_{n'=-\infty}^{\infty} T_{n'n}^{(E,E)}(\alpha, \beta) f_{n'}. \tag{2.8d}
$$

Here, $0 < q < 1$ and $\alpha, \beta \in \mathbb{C}$. Since $E_+^* = E_-$, we have

$$
\begin{aligned}
T_{n'n}^{(e,e)}(\alpha, \beta) &= \langle e_q(\beta E_+) e_q(\alpha E_-) f_n,\ f_{n'} \rangle \\
&= \langle f_n,\ e_q(\overline{\alpha} E_+) e_q(\overline{\beta} E_-) f_{n'} \rangle \\
&= \overline{T_{nn'}^{(e,e)}(\overline{\beta}, \overline{\alpha})},
\end{aligned} \tag{2.9}
$$

$$T^{(e,E)}_{n'n}(\alpha,\beta) = \overline{T^{(E,e)}_{nn'}(\overline{\beta},\overline{\alpha})}, \tag{2.10}$$

$$T^{(E,E)}_{n'n}(\alpha,\beta) = \overline{T^{(E,E)}_{nn'}(\overline{\beta},\overline{\alpha})}. \tag{2.11}$$

Furthermore, since $e_q(x)E_q(-x) = 1$, we have the identities

$$\sum_{\ell=-\infty}^{\infty} T^{(e,e)}_{n\ell}(\alpha,\beta)T^{(E,E)}_{\ell n'}(-\alpha,-\beta) = \delta_{nn'}, \qquad |\omega\alpha|, |\omega\beta| < 1, \tag{2.12}$$

$$\sum_{\ell=-\infty}^{\infty} T^{(e,E)}_{n\ell}(\alpha,\beta)T^{(E,e)}_{\ell n'}(-\alpha,-\beta) = \delta_{nn'}, \qquad |\omega\alpha|, |\omega\beta| < 1. \tag{2.13}$$

(Note that our operator derivations of these formulas and of many formulas to follow lead automatically to formal power series identities in the group parameters. These identities must then be examined case by case to determine when the series are convergent as analytic functions of the group parameters.) Using the model (2.3) to treat (2.8) as generating functions for the matrix elements and computing the coefficients of $z^{n'}$ in the resulting expressions, we obtain the explicit results:

$$\begin{aligned} T^{(e,e)}_{n'n}(\alpha,\beta) &= \frac{(q^{n-n'+1};q)_\infty(\alpha\omega)^{n-n'}}{(q;q)_\infty} {}_2\phi_1\left(\begin{matrix}0, & 0\\ q^{n-n'+1}\end{matrix}; q, \alpha\beta\omega^2\right) \\ &= \frac{(q^{n-n'+1};q)_\infty(\alpha\omega)^{n-n'}}{(q;q)_\infty(\alpha\beta\omega^2;q)_\infty} {}_0\phi_1\left(\begin{matrix}-\\ q^{n-n'+1}\end{matrix}; q, \alpha\beta\omega^2 q^{n-n'+1}\right), \end{aligned} \tag{2.14}$$

$$\begin{aligned} T^{(e,E)}_{n'n}(\alpha,\beta) &= \frac{(q^{n-n'+1};q)_\infty(\alpha\omega)^{n-n'}}{(q;q)_\infty} q^{(n-n')(n-n'-1)/2} \\ &\quad \times {}_1\phi_1\left(\begin{matrix}0\\ q^{n-n'+1}\end{matrix}; q, -\alpha\beta\omega^2 q^{n-n'}\right) \\ &= \frac{(q^{n'-n+1};q)_\infty(\beta\omega)^{n'-n}}{(q;q)_\infty} {}_1\phi_1\left(\begin{matrix}0\\ q^{n'-n+1}\end{matrix}; q, -\alpha\beta\omega^2\right), \end{aligned} \tag{2.15}$$

$$T^{(E,e)}_{n'n}(\alpha,\beta) = T^{(e,E)}_{nn'}(\beta,\alpha), \tag{2.16}$$

$$\begin{aligned} T^{(E,E)}_{n'n}(\alpha,\beta) &= \frac{(q^{n-n'+1};q)_\infty(\alpha\omega)^{n-n'}}{(q;q)_\infty} q^{(n-n')(n-n'-1)/2} \\ &\quad \times {}_0\phi_1\left(\begin{matrix}-\\ q^{n-n'+1}\end{matrix}; q, \alpha\beta\omega^2 q^{n-n'}\right). \end{aligned} \tag{2.17}$$

If $\alpha\beta \neq 0$, we can express these results in terms of the Jackson q-Bessel functions [6, p. 25],

$$J^{(1)}_\nu(z;q) = \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty}\left(\frac{z}{2}\right)^\nu {}_2\phi_1\left(\begin{matrix}0, & 0\\ q^{\nu+1}\end{matrix}; q, -\frac{z^2}{4}\right),$$

$$J_\nu^{(2)}(z;q) = \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty}\left(\frac{z}{2}\right)^\nu {}_0\phi_1\left(\begin{matrix}-\\ q^{\nu+1}\end{matrix};q,-\frac{z^2q^{\nu+1}}{4}\right), \tag{2.18}$$

$$J_\nu^{(2)}(z;q) = (-z^2/4;q)_\infty J_\nu^{(1)}(z;q),$$

and the Hahn–Exton q-Bessel function [15],

$$J_\nu(z;q) = \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty} z^\nu {}_1\phi_1\left(\begin{matrix}0\\ q^{\nu+1}\end{matrix};q,qz^2\right). \tag{2.19}$$

Indeed, setting $\alpha = ire^{i\psi}$, $\beta = ire^{-i\psi}$, we see that in terms of the new complex coordinates $[r, e^{i\psi}]$ we have

$$\begin{aligned}
T_{n'n}^{(e,e)}(\alpha,\beta) \equiv T_{n'n}^{(e,e)}[r,e^{i\psi}] &= \frac{e^{i(\frac{\pi}{2}+\psi)(n-n')}}{(-r^2\omega^2;q)_\infty} J_{n-n'}^{(2)}(2r\omega;q),\\
T_{n'n}^{(e,E)}[r,e^{i\psi}] &= e^{i(\frac{\pi}{2}-\psi)(n'-n)} q^{(n'-n)/2} J_{n'-n}(r\omega q^{-\frac{1}{2}};q),\\
T_{n'n}^{(E,e)}[r,e^{i\psi}] &= e^{i(\frac{\pi}{2}+\psi)(n-n')} q^{(n-n')/2} J_{n-n'}(r\omega q^{-\frac{1}{2}};q),\\
T_{n'n}^{(E,E)}[r,e^{i\psi}] &= e^{i(\frac{\pi}{2}+\psi)(n-n')} q^{(n-n')^2/2} J_{n-n'}^{(2)}(2r\omega q^{-\frac{1}{2}};q).
\end{aligned} \tag{2.20}$$

(Note that $J_{-n}(z;q) = (-1)^n q^{n/2} J_n(zq^{n/2};q)$, $J_{-n}^{(2)}(z;q) = (-1)^n J_n^{(2)}(z;q)$ for integer n.)

The matrix elements $T_{n'n}(\alpha,\beta)$ themselves define models of the representations (ω). We can see this directly from the commutation relations (2.1). It is a simple consequence of these relations and $e_q(x) = (x;q)_\infty^{-1}$, $E_q(x) = (-x;q)_\infty$ that

$$\begin{aligned}
e_q(\beta E_+)e_q(\alpha E_-)E_+ &= \frac{1}{\beta}(I - T_\beta)e_q(\beta E_+)e_q(\alpha E_-),\\
e_q(\beta E_+)e_q(\alpha E_-)E_- &= \frac{1}{\alpha}(I - T_\alpha)e_q(\beta E_+)e_q(\alpha E_-),
\end{aligned} \tag{2.21}$$

where I is the identity operator and $T_\beta g(\alpha,\beta) = g(\alpha,\beta q)$ for a function $g(\alpha,\beta)$. Thus

$$\begin{aligned}
\omega T_{n',n+1}^{(e,e)}(\alpha,\beta) &= \langle e_q(\beta E_+)e_q(\alpha E_-)E_+ f_n, f_{n'}\rangle\\
&= \frac{1}{\beta}(I - T_\beta)T_{n'n}^{(e,e)}(\alpha,\beta),\\
\omega T_{n',n-1}^{(e,e)}(\alpha,\beta) &= \frac{1}{\alpha}(I - T_\alpha)T_{n'n}^{(e,e)}(\alpha,\beta).
\end{aligned} \tag{2.22}$$

Similarly, the relations

$$\begin{aligned}
e_q(\beta E_+)E_q(\alpha E_-)E_+ &= \frac{1}{\beta}(I - T_\beta)e_q(\beta E_+)E_q(\alpha E_-),\\
e_q(\beta E_+)E_q(\alpha E_-)E_- &= \frac{q}{\alpha}(T_\alpha^{-1} - I)e_q(\beta E_+)E_q(\alpha E_-)
\end{aligned} \tag{2.23}$$

yield

$$\begin{aligned}
\omega T_{n',n+1}^{(e,E)}(\alpha,\beta) &= \frac{1}{\beta}(I - T_\beta)T_{n'n}^{(e,E)}(\alpha,\beta),\\
\omega T_{n',n-1}^{(e,E)}(\alpha,\beta) &= \frac{q}{\alpha}(T_\alpha^{-1} - I)T_{n'n}^{(e,E)}(\alpha,\beta).
\end{aligned} \tag{2.24}$$

In the same way we find

$$\begin{aligned}\omega T^{(E,e)}_{n',n+1}(\alpha,\beta) &= \frac{q}{\beta}(T_\beta^{-1} - I)T^{(E,e)}_{n'n}(\alpha,\beta),\\ \omega T^{(E,e)}_{n',n-1}(\alpha,\beta) &= \frac{1}{\alpha}(I - T_\alpha)T^{(E,e)}_{n'n}(\alpha,\beta)\end{aligned} \tag{2.25}$$

and

$$\begin{aligned}\omega T^{(E,E)}_{n',n+1}(\alpha,\beta) &= \frac{q}{\beta}(T_\beta^{-1} - I)T^{(E,E)}_{n'n}(\alpha,\beta),\\ \omega T^{(E,E)}_{n',n-1}(\alpha,\beta) &= \frac{q}{\alpha}(T_\alpha^{-1} - I)T^{(E,E)}_{n'n}(\alpha,\beta).\end{aligned} \tag{2.26}$$

Furthermore, induction with respect to $k+\ell$ yields $[H, \alpha^\ell\beta^k E_+^k E_-^\ell] = (k-\ell)\alpha^\ell\beta^k E_+^k E_-^\ell$, and this implies

$$[H, e_q(\beta E_+)e_q(\alpha E_-)] = (\beta\partial_\beta - \alpha\partial_\alpha)e_q(\beta E_+)e_q(\alpha E_-),$$

so that

$$\begin{aligned}(n'-n)T^{(e,e)}_{n'n}(\alpha,\beta) &= \langle e_q(\beta E_+)e_q(\alpha E_-)f_n, Hf_{n'}\rangle - \langle e_q(\beta E_+)e_q(\alpha E_-)Hf_n, f_{n'}\rangle\\ &= \langle [H, e_q(\beta E_+)e_q(\alpha E_-)]f_n, f_{n'}\rangle = (\beta\partial_\beta - \alpha\partial_\alpha)T^{(e,e)}_{n'n}(\alpha,\beta).\end{aligned}$$

Similarly,

$$(n'-n)T_{n'n}(\alpha,\beta) = (\beta\partial_\beta - \alpha\partial_\alpha)T_{n'n}(\alpha,\beta) \tag{2.27}$$

for all the remaining cases.

Thus, denoting operators

$$\tilde{E}^x = \frac{1}{x}(I - T_x), \quad \hat{E}^x = \frac{q}{x}(T_x^{-1} - I), \quad \tilde{H} = \alpha\partial_\alpha - \beta\partial_\beta, \tag{2.28}$$

we see that the following sets of operators and basis functions each define a two-variable realization of relations (2.2) and hence a realization of the representation (ω):

$$\tilde{E}_+^\beta, \tilde{E}_-^\alpha, \tilde{H}, \qquad f_{-n'+n} = T^{(e,e)}_{n'n}(\alpha,\beta), \tag{2.29a}$$

$$\tilde{E}_+^\beta, \hat{E}_-^\alpha, \tilde{H}, \qquad f_{-n'+n} = T^{(e,E)}_{n'n}(\alpha,\beta), \tag{2.29b}$$

$$\hat{E}_+^\beta, \tilde{E}_-^\alpha, \tilde{H}, \qquad f_{-n'+n} = T^{(E,e)}_{n'n}(\alpha,\beta), \tag{2.29c}$$

$$\hat{E}_+^\beta, \hat{E}_-^\alpha, \tilde{H}, \qquad f_{-n'+n} = T^{(E,E)}_{n'n}(\alpha,\beta). \tag{2.29d}$$

Moreover, from the explicit expressions (2.14) for the matrix elements $T_{n'n}(\alpha,\beta)$ it is easy to verify that relations (2.2) remain valid for $-n' = m_0$, a complex number with $0 \le \Re\, m_0 < 1$, ω a nonzero complex number, and n an integer. Thus the four families (2.29) provide realizations of all the representations $Q(\omega, m_0)$, although for $m_0 \neq 0$ the basis functions are no longer matrix elements. (Indeed, each of the families (2.29a)–(2.29d) defines two realizations of $Q(\omega, m_0)$ for $m_0 \neq 0$, one where

the q-Bessel functions are expressed in terms of the subscript $n - n'$ and one where they are expressed in terms of $n' - n$. For $m_0 = 0$ these expressions coincide.)

Before passing on to a deeper consideration of these models we note that the identities (2.12) and (2.13) for $\alpha\beta \neq 0$ are essentially the following identities for q-Bessel functions [15, eqs. (2.14), (2.10)]:

$$\sum_{\ell=-\infty}^{\infty} J^{(2)}_{\ell-n}(r;q) q^{\ell^2/2} J^{(2)}_{\ell}(rq^{-\frac{1}{2}};q) = \delta_{n0}(-r^2/4;q)_\infty, \tag{2.30a}$$

$$\sum_{\ell=-\infty}^{\infty} q^{\ell} J_{\ell+n}(r;q) J_{\ell+n'}(r;q) = \delta_{nn'} q^{-n}, \qquad |r| < q^{-\frac{1}{2}}. \tag{2.30b}$$

The identities (2.30a) can be further generalized by considering matrix element relations of the form

$$S_{n'n}(\beta,\gamma) = \sum_{\ell=-\infty}^{\infty} T^{(e,e)}_{n'\ell}(\alpha,\beta) T^{(E,E)}_{\ell n}(-\alpha,\gamma), \tag{2.12'}$$

where

$$\begin{aligned} S_{n'n}(\beta,\gamma) &= \langle E_q(\gamma E_+) e_q(\beta E_+) f_n, f_{n'} \rangle \\ &= \begin{cases} 0 & \text{if } n' - n < 0, \\ \frac{(\omega\beta)^{n'-n}}{(q;q)_{n'-n}} (-\frac{\gamma}{\beta};q)_{n'-n} & \text{otherwise.} \end{cases} \end{aligned}$$

(There are relations of a similar type for the E_- operator.) These identities yield a q-analog of the Hansen–Lommel orthogonality relations [10], [11]. Indeed, if we set $\alpha = \beta = ir, \gamma = -irq^{n'-n-1}$ in (2.12'), we see that the left-hand side of this equation vanishes unless $n = n'$. Furthermore, with the choice $2r\omega = zq^{n'/2}, m = -\ell$ the expression becomes Koelink's formula:

$$\begin{aligned} &\sum_{m=-\infty}^{\infty} J^{(2)}_{m+n'}(zq^{n'/2};q) q^{(m+n')(m+n'-1)/4} J^{(2)}_{m+n}(zq^{n/2};q) q^{(m+n)(m+n-1)/4} \\ &\qquad = \delta_{nn'}(-z^2 q^n/4;q)_\infty. \end{aligned}$$

As additional examples the identities

$$T^{(E,E)}_{n'n}(\gamma,\alpha) = \sum_{\ell=-\infty}^{n'} S_{n'\ell}(-\beta,\alpha) T^{(E,E)}_{\ell n}(\gamma,\beta),$$

$$T^{(e,E)}_{n'n}(\alpha,\beta) = \sum_{\ell=-\infty}^{n'} S_{n'\ell}(\beta,\gamma) T^{(e,E)}_{\ell n}(\alpha,-\gamma)$$

yield q-analogs of the Lommel relations:

$$x^{-m} J^{(2)}_m(xz;q) = \sum_{j=0}^{\infty} \left(\frac{q^m z}{2}\right)^j q^{j(j+1)/2} \frac{(x^2;q)_j}{(q;q)_j} J^{(2)}_{m+j}(z;q),$$

$$x^m J_m\left(\frac{z}{x} q^{m/2};q\right) = \sum_{j=0}^{\infty} \left(-\frac{zq^{1/2}}{x^2}\right)^j \frac{(x^2;q)_j}{(q;q)_j} J_{m+j}(zq^{(m+j)/2};q),$$

where m is an integer.

We can use the relations (2.22), (2.24)–(2.26) to derive addition theorems for the basis functions $T_{n'n}(\alpha,\beta)$, including the cases in which $n' = -m_0$ is a complex number. First of all, note that (for m, s complex numbers)

$$(2.31)\qquad e_q(\gamma\hat{E}^x)x^m = x^m\frac{(-q^{-m+1}\gamma/x;q)_\infty}{(-q\gamma/x;q)_\infty}, \qquad E_q(\xi\tilde{E}^x)x^s = x^s\frac{(-\xi/x;q)_\infty}{(-q^s\xi/x;q)_\infty}.$$

Now consider the operator $e_q(\gamma\hat{E}_+^\beta)E_q(\xi\tilde{E}_-^\alpha)$ applied to the basis function $f_{m_0} = T^{(E,e)}_{-m_0,0}(\alpha,\beta) = T^{(e,E)}_{0,-m_0}(\beta,\alpha)$:

(2.32)

$$\begin{aligned} &e_q(\gamma\hat{E}_+^\beta)E_q(\xi\tilde{E}_-^\alpha)T^{(E,e)}_{-m_0,0}(\alpha,\beta)\\ &\quad= (\alpha\omega)^{m_0}\frac{(q^{m_0+1};q)_\infty}{(q;q)_\infty}\sum_{k=0}^{\infty}\frac{q^{k(k-1)/2}\omega^{2k}\alpha^k\beta^k(-q^{-k+1}\gamma/\beta;q)_\infty(-\xi/\alpha;q)_\infty}{(q^{m_0+1};q)_k(q;q)_k(-q\gamma/\beta;q)_\infty(-q^{m_0+k}\xi/\alpha;q)_\infty}\\ &\quad= (\alpha\omega)^{m_0}\frac{(q^{m_0+1};q)_\infty(-\xi/\alpha;q)_\infty}{(q;q)_\infty(-q^{m_0}\xi/\alpha;q)_\infty}{}_2\phi_1\left(\begin{matrix}-q^{m_0}\xi/\alpha, & -\beta/\gamma\\ q^{m_0+1}\end{matrix};q,\alpha\gamma\omega^2\right), \end{aligned}$$

convergent for $|q^{m_0}\xi/\alpha| < 1$, $|\alpha\gamma\omega^2| < 1$. Since

$$(2.33)\qquad e_q(\gamma\hat{E}_+^\beta)E_q(\xi\tilde{E}_-^\alpha)T^{(E,e)}_{-m_0,0}(\alpha,\beta) = \sum_{n=-\infty}^{\infty} T^{(e,E)}_{n0}(\xi,\gamma)T^{(E,e)}_{-m_0,n}(\alpha,\beta)$$

from (2.8b), or

$$(2.33')\qquad \begin{aligned} &\frac{x^{m_0}(q^{m_0+1};q)_\infty(qz/xy;q)_\infty}{(q;q)_\infty(q^{m_0+1}z/xy;q)_\infty}{}_2\phi_1\left(\begin{matrix}q^{m_0+1}z/xy, & qx/yz\\ q^{m_0+1}\end{matrix};q,xyz\right)\\ &\quad= \sum_{n=-\infty}^{\infty} y^nJ_n(z;q)J_{m_0+n}(x;q), \qquad |q^{m_0+1}z/xy| < 1, \qquad |xyz| < 1, \end{aligned}$$

this can be considered as an addition theorem for the $T^{(E,e)}_{-m_0,n}$ basis functions that generalizes the Lommel and Graf addition theorems for ordinary Bessel functions, [15, eq. (4.5)], [18], [20], [22]. In [15, eq. (4.10)] (for integer m_0) and in [5] (for complex m_0) the addition theorem corresponding to

$$(2.34)\qquad E_q(\gamma\tilde{E}_+^\beta)E_q(\xi\tilde{E}_-^\alpha)T^{(e,e)}_{-m_0,0}(\alpha,\beta) = \sum_{n=-\infty}^{\infty} T^{(E,E)}_{n0}(\xi,\gamma)T^{(e,e)}_{-m_0,n}(\alpha,\beta)$$

is worked out. The result is

(2.34′)

$$\begin{aligned} &\left(\frac{x}{2}\right)^{m_0}\frac{(qz/xy;q)_\infty(q^{m_0+1};q)_\infty}{(q^{m_0+1}z/xy;q)_\infty(q;q)_\infty}{}_2\phi_1\left(\begin{matrix}q^{m_0+1}z/xy, & yz/x\\ q^{m_0+1}\end{matrix};q,-\frac{x^2}{4}\right)\\ &\quad= \sum_{n=-\infty}^{\infty} q^{n(n-1)/2}y^nJ_n^{(2)}(z;q)J^{(1)}_{m_0+n}(x;q), \qquad |q^{m_0+1}z/xy| < 1, \qquad |x^2| < 4. \end{aligned}$$

To better understand and to extend identities (2.33) we note that, formally at least, the operators $e_q(\gamma\hat{E}_+^\beta)$, $E_q(\xi\tilde{E}_-^\alpha)$, and H are symmetry operators for the q-difference equation

$$\hat{E}_+^\beta \tilde{E}_-^\alpha \Psi(\alpha,\beta) = \ell^2 \Psi(\alpha,\beta). \tag{2.35}$$

That is, these operators map solutions of the equations to new solutions. Thus if $\Psi(\alpha,\beta)$ is a solution, then so are $e_q(\gamma\hat{E}_+^\beta)E_q(\xi\tilde{E}_-^\alpha)\Psi(\alpha,\beta) = \Psi'(\alpha,\beta)$ and $H\Psi(\alpha,\beta)$. It is easy to check that the only solution Ψ_λ of (2.35) and $H\Psi_\lambda = \lambda\Psi_\lambda$ for complex (noninteger) λ, such that $r^{-\lambda}\Psi_\lambda$ is analytic in r at $r = 0$, is $\Psi_\lambda(\alpha,\beta) \equiv \Psi_\lambda[r,t] = t^\lambda J_\lambda(r\omega q^{-1/2};q)$, unique up to multiplication by a constant. Here, $\alpha = irt$, $\beta = ir/t$. (For $\lambda = n$, an integer, the only solution Ψ_λ analytic in r at $r = 0$ is $\Psi_n = t^n J_n(r\omega q^{-1/2};q)$.) Thus if $\Psi'_\lambda(\alpha,\beta)$ is analytic in the variables $R = r$, $S = rt$ at $R = 0$, then it must have an expansion of the form [2]

$$\Psi' = \sum_\lambda c_\lambda t^\lambda J_\lambda(r\omega q^{-1/2};q), \tag{2.36}$$

where the c_λ are complex constants. Expression (2.33) is a special case of this expansion.

It is easy to see from (2.31) that the operators $e_q(\gamma\hat{E}_+^\beta)$, $E_q(\xi\tilde{E}_-^\alpha)$ do commute with $\hat{E}_+^\beta$ and $\tilde{E}_-^\alpha$ when acting on monomials $\beta^m\alpha^s$, and so modulo convergence problems are symmetry operators mapping analytic solutions of $\hat{E}_+^\beta\tilde{E}_-^\alpha\Psi = \ell^2\Psi$ to analytic solutions. Now, instead of the definition $E_q(\xi\tilde{E}^x)x^s = x^s(-\xi/x;q)_\infty/(-q^s\xi/x;q)_\infty$, let us adopt the definition ($\xi \neq 0$)

$$E_q(\xi\tilde{E}_-^x)'x^s = \xi^s q^{s(s-1)/2}\frac{(-q^{-s+1}x/\xi;q)_\infty}{(-qx/\xi;q)_\infty}. \tag{2.37}$$

For s an integer these definitions coincide, but they are distinct for s complex. It is again easy to verify that

$$E_q(\xi\tilde{E}_-^\alpha)'\tilde{E}_-^\alpha\alpha^s = \tilde{E}_-^\alpha E_q(\xi\tilde{E}_-^\alpha)'\alpha^s,$$

so that $E_q(\xi\tilde{E}_-^\alpha)'$ is also a symmetry operator. As an example of the use of this observation consider

(2.38)

$$\begin{aligned}\Psi' &= E_q(\xi\tilde{E}_-^\alpha)'T_{-m_0,0}^{(E,e)}(\alpha,\beta)\\ &= \frac{(\xi\omega)^{m_0}(q^{m_0+1};q)_\infty q^{m_0(m_0-1)/2}(-q^{-m_0+1}\alpha/\xi;q)_\infty}{(q;q)_\infty(-q\alpha/\xi;q)_\infty}\,{}_1\phi_1\left(\begin{matrix}-q^{m_0}\xi/\alpha\\ q^{m_0+1}\end{matrix};q,-\alpha\beta\omega^2\right),\\ &\qquad |q\alpha/\xi| < 1.\end{aligned}$$

Thus we have

$$\frac{(-q^{-m_0+1}irt/\xi;q)_\infty}{(-qirt/\xi;q)_\infty}\,{}_1\phi_1\left(\begin{matrix}-q^{m_0}\xi/irt\\ q^{m_0+1}\end{matrix};q,r^2\omega^2\right) = \sum_{n=-\infty}^{\infty} c_n t^n J_n(r\omega q^{-1/2};q). \tag{2.39}$$

Now set $t = s/r$ and let $r \to 0$ in (2.39). We obtain $c_n = (-iq^{3/2}/\xi\omega)^n (q^{-m_0};q)_\infty/(q^{-m_0+n};q)_\infty$ for $n = 0, 1, \dots$. Similarly, if we set $t = sr$ and let $r \to 0$, we find the same expression for c_n with $-n = 0, 1, \dots$. Thus

$$\frac{(q^{-m_0}x;q)_\infty}{(x;q)_\infty} {}_1\phi_1\left(\begin{matrix} q^{m_0+1}/x \\ q^{m_0+1}\end{matrix}; q, qz^2\right) = \sum_{n=-\infty}^{\infty} \left(\frac{qx}{z}\right)^n \frac{(q^{-m_0};q)_\infty}{(q^{-m_0+n};q)_\infty} J_n(z;q), \tag{2.40}$$
$$0 < |x| < 1.$$

Similarly, applying the symmetry operator $E_q(\xi \tilde{E}_-^\alpha)'$ to the basis function $T^{(e,e)}_{-m_0,0}$, we obtain the identity

(2.41)
$$\frac{(q^{-m_0}ry;q)_\infty}{(ry;q)_\infty} {}_2\phi_1\left(\begin{matrix} q^{m_0+1}/ry, & 0 \\ q^{m_0+1}\end{matrix}; q, -r^2\right) = \sum_{n=-\infty}^{\infty} y^n \frac{(q^{-m_0};q)_\infty}{(q^{-m_0+n};q)_\infty} J_n^{(1)}(2r;q),$$
$$0 < |y|, \quad |r^2| < 1, \quad |ry| < 1.$$

Also, we can apply the symmetry operator

$$e_q(\gamma \hat{E}_+^\beta)' \beta^m = \gamma^m q^{-m(m-1)/2} \frac{(-\beta/\gamma;q)_\infty}{(-q^m\beta/\gamma;q)_\infty},$$

rather than $e_q(\gamma \hat{E}_+^\beta)$, to obtain new identities.

3. Tensor products. In addition to the standard definition of the tensor product of two irreducible representations (ω_1) and (ω_2) of $m(2)$, there is a definition given by the nontrivial Hopf algebra structure in which the coproduct is [1], [4], [11]

$$\begin{aligned} F_+ &= \Delta(E_+) = E_+ \otimes q^{-\frac{1}{2}H} + q^{\frac{1}{2}H} \otimes E_+, \\ F_- &= \Delta(E_-) = E_- \otimes q^{-\frac{1}{2}H} + q^{\frac{1}{2}H} \otimes E_-, \\ L &= \Delta(H) = H \otimes I + I \otimes H. \end{aligned} \tag{3.1}$$

The operators $F_\pm$, L satisfy the same commutation relations as the operators $E_\pm$, H:

$$[L, F_\pm] = \pm F_\pm, \quad [F_+, F_-] = 0. \tag{3.2}$$

The standard definition is obtained by letting $q \to 1$ in (3.1). Each irreducible representation (ω_1), (ω_2) is defined on $L_2[0, 2\pi]$ by the prescription (2.5). To make sense of the operators (3.1) on a dense subspace of the tensor-product space $L[0, 2\pi] \otimes L[0, 2\pi]$ we proceed as follows. The Hilbert-space inner product is

$$\langle f, g \rangle = \frac{1}{(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} f(\theta_1, \theta_2) \overline{g(\theta_1, \theta_2)}\, d\theta_1 d\theta_2.$$

An orthonormal basis is $\{f_{n_1 n_2} = e^{i(n_1\theta_1 + n_2\theta_2)}, n_j = 0, \pm 1, \pm 2, \dots\}$. In terms of this basis a general element f of the Hilbert space can be expanded as

$$f(\theta_1, \theta_2) \sim \sum_{n_1, n_2 = -\infty}^{\infty} a_{n_1 n_2} e^{i(n_1\theta_1 + n_2\theta_2)} = \sum_{n,k=-\infty}^{\infty} A_{nk} e^{i(n\Theta + k\Phi)}, \tag{3.3}$$

where $\sum |a_{n_1n_2}|^2 < \infty$, $k = n_1 + n_2$, $n = n_1$, $\Theta = \theta_1 - \theta_2$, $\Phi = \theta_2$, and $A_{nk} = a_{n_1n_2}$. Now set $z = e^{i\Theta}$, $t = e^{i\Phi}$. Since $L = \Delta(H) = -i\partial_\Phi = t^{-1}\partial_t$, it is clear that the eigenvalues of L are $k_0 = 0, \pm 1, \pm 2, \ldots$ and that the corresponding eigenspace of $L[0, 2\pi] \otimes L[0, 2\pi]$ consists of those square integrable functions f for which $A_{nk} = 0$ if $k \neq k_0$. The operators (3.2) make sense on the dense subspace $\mathcal{K}$ of the Hilbert space consisting of finite sums of functions $f[z,t] = \sum f_k[z]t^k$, $(z,t) \in \mathcal{C}_2$, ($f[e^{i\Theta}, e^{i\Phi}] = f(\theta_1, \theta_2)$ belongs to the Hilbert space) such that each $f_k[z]$ is analytic except possibly for an essential singularity at $z = 0$ and a countable number of poles in the complex plane (not on the unit circle $|z| = 1$). Then

$$
\begin{aligned}
F_+f[z,t] &= \omega_1 ztf[zq^{1/2}, tq^{-1/2}] + \omega_2 tf[zq^{1/2}, t], \\
F_-f[z,t] &= \frac{\omega_1}{zt} f[zq^{1/2}, tq^{-1/2}] + \frac{\omega_2}{t} f[zq^{1/2}, t], \\
Lf[z,t] &= \frac{1}{t}\partial_t f[z,t].
\end{aligned} \tag{3.4}
$$

We further require that $F_\pm f[e^{i\Theta}, e^{i\Phi}]$ and $F_+F_-f[e^{i\Theta}, e^{i\Phi}]$ are square integrable. Formally, the invariant operator $\Delta(E_+E_-) = F_+F_-$ takes the form

$$
F_+F_-f[z,t] = \left(\omega_1^2 T_t^{-1} + \omega_1\omega_2\left(zq^{1/2} + \frac{1}{zq^{1/2}}\right)T_t^{-1/2} + \omega_2^2 I\right) T_z f[z,t].
$$

Thus if f satisfies

$$
F_+F_-f = \lambda f, \quad Lf = kf, \tag{3.5}
$$

it should have the form $f[z,t] = g_k[z]t^k$, where

$$
(\omega_1 q^{-k/2} + \omega_2 zq^{1/2})\left(\omega_1 q^{-k/2} + \frac{\omega_2}{zq^{1/2}}\right) g_k[zq] = \lambda g_k[z]. \tag{3.6}
$$

To decompose the tensor-product representation into irreducible representations it is sensible to require that the domain of definition of the preceding operators be adjusted so that on this domain F_+ is the adjoint of F_- and L, $K = F_+F_-$ are symmetric. Then we extend the domain to make K self-adjoint. Let us consider the restriction K_k of K to the subspace of functions of the form $f_k[z,t] = g_k[z]t^k$, so that $Lf_k = kf_k$. Then

$$
K_k = (\omega_1 q^{-k/2} + \omega_2 zq^{1/2})\left(\omega_1 q^{-k/2} + \frac{\omega_2}{zq^{1/2}}\right) T_z \tag{3.7}
$$

and the induced Hilbert-space inner product is

$$
(g, g') = \frac{1}{2\pi i}\oint_{|z|=1} g[z]\overline{g}'[z^{-1}]\,\frac{dz}{z}, \tag{3.8}
$$

where $\overline{g[z]} = \overline{g}[\overline{z}]$. It follows that

$$
(K_k g, g') = (g, K_k g') + R(g, g'), \tag{3.9}
$$

where $R(g, g')$ is the sum of residues of the function $z^{-1}(\omega_1 q^{-k/2} + \omega_2 zq^{-1/2})(\omega_1 q^{-k/2} + \omega_2 q^{1/2}/z)g[z]\overline{g}'[q/z]$ in the annulus $q < |z| < 1$. Thus K_k will be symmetric on

a domain $\mathcal{D}$ such that $R(g,g')=0$ for all $g,g'\in\mathcal{D}$. This shows that the role of boundary conditions for symmetric operators is played here by the placement of zeros and poles for the elements of $\mathcal{D}$.

There is an infinite parameter family of self-adjoint operators associated with the symbol (3.7). We shall focus on those two of the self-adjoint operators (and their spectral resolutions) that are the simplest in structure. The first corresponds to the orthonormal basis of eigenvectors

$$h_p^{(1)}[z]=\frac{(-q^{(1+k)/2}\omega_2 z/\omega_1;q)_\infty}{(-q^{(1+k)/2}\omega_2/\omega_1 z;q)_\infty}z^{p+k},\qquad p=0,\pm1,\dots, \tag{3.10}$$

with eigenvalues $\lambda_p^{(1)}=\omega_1^2q^p$. The second has an orthonormal basis of eigenvectors

$$h_p^{(2)}[z]=\frac{(-q^{(1-k)/2}\omega_1 z/\omega_2;q)_\infty}{(-q^{(1-k)/2}\omega_1/\omega_2 z;q)_\infty}z^{p},\qquad p=0,\pm1,\dots, \tag{3.11}$$

with eigenvalues $\lambda_p^{(2)}=\omega_2^2q^p$. Setting $f_{pk}^{(j)}[z,t]=h_p^{(j)}[z]t^k$, $j=1,2$, we have

$$\begin{aligned}F_\pm f_{pk}^{(j)}&=\omega_j q^{p/2}f_{p,k\pm1}^{(j)},\\ Lf_{pk}^{(j)}&=kf_{pk}^{(j)},\qquad Kf_{pk}^{(j)}=\omega_j^2q^pf_{pk}^{(j)}.\end{aligned} \tag{3.12}$$

Thus

$$(\omega_1)\otimes_q^j(\omega_2)\equiv\sum_{p=-\infty}^{\infty}\oplus(\omega_jq^{p/2}), \tag{3.13}$$

i.e., in each case the tensor product decomposes into a direct sum of irreducible representations. (This is dramatically different from the case $q=1$ in which the tensor-product representation is unique and the spectrum of K is continuous [22, p. 224].)

As an indication of the multiplicity of possible self-adjoint operators and resulting spectral decompositions, consider the function

$$S_\lambda[z]=\frac{(-q^{1/2}\lambda/z;q)_\infty(-q^{1/2}z/\lambda;q)_\infty}{(-q^{1/2}\lambda z;q)_\infty(-q^{1/2}/\lambda z;q)_\infty} \tag{3.14}$$

for λ a nonzero real constant. Note that $|S_\lambda[z]|=1$ for $|z|=1$. Now $S_\lambda[zq]=\lambda^2S_\lambda[z]$ so the sets $\{S_\lambda[z]f_{pk}^{(j)}\}$, $j=1,2$, are each orthonormal bases of $L_2[0,2\pi]\otimes L_2[0,2\pi]$ consisting of eigenvectors of K with eigenvalues $\lambda^2\omega_j^2q^p$. Furthermore, the sets $\{f_{pk}^{(j)\prime}=S_\lambda[z^2]f_{pk}^{(j)}\}$ satisfy relations (3.12) with ω_j replaced by $\omega_j\lambda^2$.

The Clebsch–Gordan coefficients for the tensor product corresponding to the spectral resolution (3.10) are defined by

$$f_{pk}^{(1)}[z,t]=f_{pk}^{(1)}(\theta_1,\theta_2)=\sum_{n_1,n_2=-\infty}^{\infty}\begin{bmatrix}\omega_1&\omega_2&p\\ n_1&n_2&k\end{bmatrix}^{(1)}f_{n_1}^{\omega_1}(e^{i\theta_1})\otimes_q^1 f_{n_2}^{\omega_2}(e^{i\theta_2}). \tag{3.15}$$

Clearly, these coefficients vanish unless $k=n_1+n_2$. The orthogonality of the two bases implies the identities

$$\sum_{n_1,n_2}\begin{bmatrix}\omega_1&\omega_2&p\\ n_1&n_2&k\end{bmatrix}^{(1)}\begin{bmatrix}\omega_1&\omega_2&p'\\ n_1&n_2&k\end{bmatrix}^{(1)}=\delta_{pp'}, \tag{3.16}$$

$$\sum_p \begin{bmatrix} \omega_1 & \omega_2 & p \\ n_1 & n_2 & k \end{bmatrix}^{(1)} \begin{bmatrix} \omega_1 & \omega_2 & p \\ n_1' & n_2' & k \end{bmatrix}^{(1)} = \delta_{n_1 n_1'},$$

where $n_1 + n_2 = n_1' + n_2' = k$. Explicitly,

$$\begin{bmatrix} \omega_1 & \omega_2 & p \\ n_1 & n_2 & k \end{bmatrix}^{(1)} = (-1)^{p+n_2} q^{(p+n_2)/2} J_{p+n_2}(q^{-1/2} b_k; q) - \sum_{\ell=0}^{s_k-1} \frac{(q^\ell b_k^2; q)_\infty q^{\ell(\ell+1)/2} (-q^\ell b_k)^{p+n_2} (-1)^\ell}{(q;q)_\infty (q;q)_\ell}, \tag{3.17}$$

where $k = n_1 + n_2$, $b_k = q^{(1+k)/2}\omega_2/\omega_1$, and $s_k \geq 0$ is the smallest integer such that $q^{s_k} b_k < 1$. (We assume that $b_k \neq q^n$ for any integer n.) Indeed, the case $s_k = 0$ of (3.17) follows from (3.15), and this result can be written as a complex contour integral. The case $s_k > 0$ can be obtained from this result by shifting the contour.

Similarly, the Clebsch–Gordan coefficients for the tensor product corresponding to the spectral resolution (3.11) are defined by

$$f_{pk}^{(2)}[z,t] = f_{pk}^{(2)}(\theta_1, \theta_2) = \sum_{n_1, n_2 = -\infty}^{\infty} \begin{bmatrix} \omega_1 & \omega_2 & p \\ n_1 & n_2 & k \end{bmatrix}^{(2)} f_{n_1}^{\omega_1}(e^{i\theta_1}) \otimes_q^2 f_{n_2}^{\omega_2}(e^{i\theta_2}).$$

They satisfy orthogonality relations analogous to (3.16) and are given explicitly by (3.17), where now $b_k = q^{(1-k)/2}\omega_1/\omega_2$ and n_2 is replaced by $-n_1$.

With respect to the tensor-product basis $\{f_{n_1}^{\omega_1} \otimes f_{n_2}^{\omega_2}\}$ the operator $e_q(\beta F_+)E_q(\alpha F_-)$ has matrix elements

$$\begin{aligned} T_{m_1 m_2; n_1 n_2}(\alpha, \beta) &= \langle e_q(\beta\Delta(E_+))E_q(\alpha\Delta(E_-)) f_{n_1}^{\omega_1} \otimes f_{n_2}^{\omega_2}, f_{m_1}^{\omega_1} \otimes f_{m_2}^{\omega_2} \rangle \\ &= (\beta\omega_1)^{m_1 - n_1} (\beta\omega_2)^{m_2 - n_2} \frac{(q^{m_1 - n_1 + 1}; q)_\infty (q^{m_2 - n_2 + 1}; q)_\infty}{(q;q)_\infty^2 q^{(m_1 m_2 + n_1 n_2 - 2n_1 m_2)/2}} \\ &\quad \times {}_1\phi_1\left(\begin{matrix} 0 \\ q^{m_1 - n_1 + 1} \end{matrix}; q, -\alpha\beta\omega_1^2 q^{-m_2} \right) {}_1\phi_1\left(\begin{matrix} 0 \\ q^{m_2 - n_2 + 1} \end{matrix}; q, -\alpha\beta\omega_2^2 q^{n_1} \right), \end{aligned}$$

or, in coordinates $[r,t]$,

$$T_{m_1 m_2; n_1 n_2}[r,t] = (-q^{-1/2} it)^{n_1 + n_2 - m_1 - m_2} J_{m_1 - n_1}(r\omega_1 q^{-(m_2+1)/2}; q) J_{m_2 - n_2}(r\omega_2 q^{(n_1 - 1)/2}; q). \tag{3.18}$$

To see this most simply, recall that if X and Y are linear operators such that $YX = qXY$, then [6, p. 28],

$$(Y + X)^k = \sum_{\ell=0}^{k} \frac{(q;q)_k}{(q;q)_\ell (q;q)_{k-\ell}} X^\ell Y^{k-\ell},$$

$$e_q(X + Y) = e_q(X) e_q(Y), \qquad E_q(X + Y) = E_q(Y) E_q(X).$$

Thus, using the facts that $E_+ q^{\frac{1}{2}H} = q^{-\frac{1}{2}} q^{\frac{1}{2}H} E_+$, $E_- q^{\frac{1}{2}H} = q^{\frac{1}{2}} q^{\frac{1}{2}H} E_-$ for these representations, we have

$$\begin{aligned} &e_q(\beta F_+)E_q(\alpha F_-) \\ &\quad = e_q(\beta E_+ \otimes q^{-\frac{1}{2}H}) E_q(\alpha E_- \otimes q^{-\frac{1}{2}H}) e_q(\beta q^{\frac{1}{2}H} \otimes E_+) E_q(\alpha q^{\frac{1}{2}H} \otimes E_-), \end{aligned}$$

so that the matrix elements with respect to the tensor-product basis are given by

$$T_{m_1m_2;n_1n_2}(\alpha,\beta) = \langle e_q(\beta q^{-m_2/2}E_+)E_q(\alpha q^{-m_2/2}E_-)f_{n_1}^{\omega_1}, f_{m_1}^{\omega_1}\rangle \times \langle e_q(\beta q^{n_1/2}E_+)E_q(\alpha q^{n_1/2}E_-)f_{n_2}^{\omega_2}, f_{m_2}^{\omega_2}\rangle.$$

From the definition of the Clebsch–Gordan coefficients we see immediately that the identities

(3.19)
$$T_{m_1m_2;n_1n_2}(\alpha,\beta) = \sum_{p=-\infty}^{\infty} \begin{bmatrix} \omega_1 & \omega_2 & p \\ n_1 & n_2 & n_1+n_2 \end{bmatrix}^{(j)} T^{(e,E),\omega_j q^{p/2}}_{m_1+m_2,n_1+n_2}(\alpha,\beta) \times \begin{bmatrix} \omega_1 & \omega_2 & p \\ m_1 & m_2 & m_1+m_2 \end{bmatrix}^{(j)}, \qquad j=1,2$$

must hold. Choosing $j=1$ and taking the special case for which $q^{(1+m_1+m_2)/2}\omega_2/\omega_1 < 1$ and $q^{(1+n_1+n_2)/2}\omega_2/\omega_1 < 1$, we obtain the identity

(3.20)
$$(-q)^n J_{x-n}(Sq^{-y};q^2)J_n(RSq^n;q^2) = \sum_{k=-\infty}^{\infty} q^{2k} J_{k-n}(Rq^y;q^2)J_x(Sq^{k-y};q^2)J_k(Rq^{y+x};q^2),$$

which is valid for $0 < Rq^{y+x+1} < 1$, $0 < Rq^{y+1} < 1$, $0 \le S$, and n, x, y integers. The case $S = q^z$ of this formula for z an integer is the addition theorem for Hahn–Exton q-Bessel functions derived by Koelink by using the theory of quantum groups, [10], [21]. For an analytic proof of Koelink's formula, see [20].

Acknowledgment. Willard Miller, Jr. has worked with both Dick Askey and Frank Olver, in an editorial capacity and as a friend and colleague, for some 25 years and respectfully dedicates this paper to them.

REFERENCES

[1] E. Abe (1980), *Hopf Algebras*, Cambridge University Press, Cambridge, England.
[2] A. K. Agarwal, E. G. Kalnins, and W. Miller (1987), *Canonical equations and symmetry techniques for q-series*, SIAM J. Math. Anal., 18, pp. 1519–1538.
[3] R. Floreanini and L. Vinet (1990), *q-Orthogonal Polynomials and the Oscillator Quantum Group*, INFN preprint AE-90/23, Trieste, Italy.
[4] ——— (1991), *Quantum algebras and q-special functions*, Lett. Math. Phys., 22, pp. 45–54.
[5] ——— (1991), *Addition Formulas for q-Bessel Functions*, Université de Montréal preprint UdeM-LPN-TH60, Montreal.
[6] G. Gasper and M. Rahman (1990), *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, England.
[7] M. Jimbo (1985), *A q-difference analogue of U(g) and the Yang–Baxter equation*, Lett. Math. Phys., 10, pp. 63–69.
[8] E. G. Kalnins, H. L. Manocha, and W. Miller (1992), *Models of q-algebra representations: Tensor products of special unitary and oscillator algebras*, J. Math. Phys., 33, pp. 2365–2383.
[9] E. G. Kalnins and W. Miller (1989), *Symmetry techniques for q-series: Askey–Wilson polynomials*, Rocky Mountain J. Math., 19, pp. 223–230.
[10] H. T. Koelink (1991), *Hansen–Lommel Orthogonality Relations for Jackson's q-Bessel Functions*, Report W-91-11, University of Leiden, Leiden, the Netherlands.

[11] ——— (1991), *On Quantum Groups and q-Special Functions*, Ph.D. thesis, University of Leiden, Leiden, the Netherlands.

[12] H. T. KOELINK AND T. H. KOORNWINDER (1989), *The Clebsch–Gordan coefficients for the quantum group $S_\mu U(2)$ and q-Hahn polynomials*, Proc. Kon. Nederl. Akad. Wetensch. Ser. A, 92, pp. 443–456.

[13] T. H. KOORNWINDER (1989), *Representations of the twisted $SU(2)$ quantum group and some q-hypergeometric orthogonal polynomials*, Nederl. Akad. Wetensch. Proc. Ser. A, 92, pp. 97–117.

[14] ——— (1991), *The addition formula for little q-Legendre polynomials and the $SU(2)$ quantum group*, SIAM J. Math. Anal., 22, pp. 295–301.

[15] T. H. KOORNWINDER AND R. F. SWARTTOUW (1991), *On q-analogues of the Fourier and Hankel transforms group*, Trans. Amer. Math. Soc., to appear.

[16] T. MASUDA, K. MIMACHI, Y. NAKAGAMI, M. NOUMI, Y. SABURI, AND K. UENO (1990), *Unitary representations of the quantum groups $SU_q(1,1)$: Structure of the dual space of $\mathcal{U}_q(sl(2))$*, Lett. Math. Phys., 19, pp. 187–194.

[17] ——— (1990), *Unitary representations of the quantum groups $SU_q(1,1)$*: II. *Matrix elements of unitary representations and the basic hypergeometric functions*, Lett. Math. Phys., 19, pp. 194–204.

[18] W. MILLER (1968), *Lie Theory and Special Functions*, Academic Press, New York.

[19] ——— (1970), *Lie theory and q-difference equations*, SIAM J. Math. Anal., 1, pp. 171–188.

[20] R. F. SWARTTOUW (1992), *The Hahn–Exton q-Bessel function*, Ph.D. thesis, Delft University of Technology, Delft, the Netherlands.

[21] L. L. VAKSMAN AND L. I. KOROGODSKIĬ (1989), *An algebra of bounded functions on the quantum group of the motions of the plane, and q-analogues of Bessel functions*, Soviet Math. Dokl., 39, pp. 173–177.

[22] N. JA. VILENKIN (1968), *Special Functions and the Theory of Group Representations*, American Mathematical Society, Providence, RI.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 528–548, March 1994

019

SOME RESULTS ON CO-RECURSIVE ASSOCIATED LAGUERRE AND JACOBI POLYNOMIALS*

JEAN LETESSIER†

Abstract. The author presents results on co-recursive associated Laguerre and Jacobi polynomials which are of interest for the solution of the Chapman–Kolmogorov equations of some birth and death processes with or without absorption. Explicit forms, generating functions, and absolutely continuous parts of the spectral measures are given. Fourth-order differential equations satisfied by the polynomials with a special attention to some simple limiting cases are derived.

Key words. orthogonal polynomials, birth and death processes, hypergeometric functions

AMS subject classifications. 33A65, 60J80, 33A30

1. Introduction. Starting from a sequence of orthogonal polynomials $\{P_n\}_{n\geq 0}$ defined by the recurrence relation

$$P_{n+2}(x) = (x-\beta_{n+1})P_{n+1}(x) - \gamma_{n+1}P_n(x), \qquad n \geq 0, \tag{1.1}$$

and the initial conditions

$$P_0(x) = 1, \qquad P_1(x) = x - \beta_0, \tag{1.2}$$

with β_n, $\gamma_n \in \mathbb{C}$, and $\gamma_n \neq 0$, several modifications were considered:

(i) Associated polynomials arise when we replace n by $n+c$ in the coefficients β_n and γ_n (keeping $\gamma_n \neq 0$). If c is an integer k these polynomials are called associated of order k. The associated polynomials of order one are the numerator polynomials;

(ii) Co-recursive polynomials arise when we replace β_0 by $\beta_0 + \mu$;

(iii) Perturbed polynomials arise when we replace γ_1 by $\lambda\gamma_1$, $(\lambda > 0)$;

(iv) Co-recursive polynomials of these perturbed polynomials arise when the two previous modifications are made together;

(v) Generalized co-recursive and perturbed polynomials arise when we change β_n and/or γ_n at any level n.

In the study of birth and death processes, orthogonal polynomials, in particular all the hypergeometric families of the Askey scheme [1], [19] and their corresponding associated families, play a primordial role in the Karlin–McGregor solution of the Chapman–Kolmogorov equation [11], [16]

$$p_{mn}(t) \sim \int_0^\infty e^{-xt}P_m(x)P_n(x)d\phi(x). \tag{1.3}$$

In certain birth and death processes, zero-related polynomials [12], [13], [14] arise in a natural way. Zero-related polynomials are special co-recursive associated polynomials.

More generally, co-recursive and generalized co-recursive polynomials are involved in the solution of the Chapman–Kolmogorov equation of birth and death processes with absorption or killing [11], [17], [18].

The purpose of this paper is to present some results on co-recursive, associated Laguerre and Jacobi polynomials which are of special interest in the study of birth and

* Received by the editors March 31, 1992; accepted for publication (in revised form) April 13, 1993.

† Laboratoire de Physique Théorique et Hautes Energies, Unité associée au Centre National de la Recherche Scientifique, Université Paris 7, Tour 24, 5è ét., 2 Place Jussieu, F-75251 Cedex 05.

death processes with linear and rational rates, respectively. The Laguerre polynomial families are involved in processes for which the birth and death rates are of the form [12]

$$\lambda_n = n+\alpha+c+1, \quad \mu_{n+1} = n+c+1, \quad n \geq 0, \quad \mu_0 = c-\mu. \tag{1.4}$$

The Jacobi polynomial families are involved when the rates are of the form

$$\lambda_n = \frac{2(n+c+\alpha+\beta+1)(n+c+\beta+1)}{(2n+2c+\alpha+\beta+1)(2n+2c+\alpha+\beta+2)}, \qquad n \geq 0, \tag{1.5a}$$

$$\mu_n = \frac{2(n+c)(n+c+\alpha)}{(2n+2c+\alpha+\beta)(2n+2c+\alpha+\beta+1)}, \qquad n > 0, \tag{1.5b}$$

$$\mu_0 = \frac{2c(c+\alpha)}{(2c+\alpha+\beta)(2c+\alpha+\beta+1)} - \mu. \tag{1.5c}$$

In both cases $\mu_0 = 0$ correspond to the "honest" [24] linear processes (i.e., processes for which the sum of the probabilities $p_{mn}(t)$ is equal to 1). Cases $\mu_0 = \text{Const.} \neq 0$ correspond to processes with absorption and are not "honest." However, if $\mu_0 = c$ in the Laguerre case or $\mu_0 = 2c(c+\alpha)/(2c+\alpha+\beta)(2c+\alpha+\beta+1)$ in the Jacobi case, the corresponding processes are simply solved using associated polynomials.

In §2 we explain the method used by applying it to the Laguerre case. In §2.1 we give an explicit expression for the co-recursive associated Laguerre (CAL) polynomials. In §2.2 we derive a generating function of them and we find the absolutely continuous part of the spectral measure. Section 2.3 is devoted to the derivation, using the Orr's method, of a fourth-order differential equation satisfied by the CAL polynomials. In §2.4 we present results obtained in some limiting cases, among which is a new simple case of associated Laguerre polynomials.

In §3 we briefly present some results corresponding to the Jacobi case. In §3.1 we give an explicit expression for the co-recursive associated Jacobi (CAJ) polynomials. In §3.2 we present a generating function and in §3.3 we give the absolutely continuous component of the spectral measure. Section 3.5 is devoted to some limiting cases of CAJ polynomials for which we give fourth-order differential equations that they satisfy. In §3.6 we give some concluding remarks.

We use the notation of [7] for the special functions used in this work. We do not give validity conditions on the parameters of the used hypergeometric functions, analytic continuations, or limiting processes giving, in general, valid formulas. We use Slater's notation [29] for the product of Γ functions:

$$\Gamma\left(\begin{matrix}\alpha_1,\ldots,\alpha_p\\ \beta_1,\ldots,\beta_q\end{matrix}\right) = \prod_{i=1}^{p}\Gamma(\alpha_i)\Big/\prod_{i=1}^{q}\Gamma(\beta_i). \tag{1.6}$$

2. The case of Laguerre polynomials. Replacing n by $n+c$ in the recurrence relation of the Laguerre polynomials we obtain the recurrence relation satisfied by the associated Laguerre polynomials $L_n^{\alpha}(x;c)$:

$$(2n+2c+\alpha+1-x)p_n = (n+c+1)p_{n+1} + (n+c+\alpha)p_{n-1}. \tag{2.1}$$

To complete the definition of the polynomials $L_n^{\alpha}(x;c)$ the initial condition

$$L_{-1}^{\alpha}(x;c) = 0, \qquad L_0^{\alpha}(x;c) = 1 \tag{2.2}$$

has to be imposed. These polynomials are orthogonal with respect to a positive measure when $(n+c)(n+\alpha+c) > 0$, for all $n > 0$. (See [2] for details.)

Note that if we consider the monic polynomials $\mathbf{L}_n^\alpha(x;c)$ defined by

$$\mathbf{L}_n^\alpha(x;c) = (-1)^n(c+1)_n L_n^\alpha(x;c), \tag{2.3}$$

they satisfy the recurrence relation

$$(x-2n-2c-\alpha-1)\mathbf{L}_n^\alpha(x;c) = \mathbf{L}_{n+1}^\alpha(x;c) + (n+c)(n+c+\alpha)\mathbf{L}_{n-1}^\alpha(x;c), \tag{2.4}$$

We can see that this recurrence is invariant in the transformation $\mathcal{T}$ defined by

$$\mathcal{T}(c,\alpha) = (c+\alpha, -\alpha). \tag{2.5}$$

2.1. Explicit representation of the CAL polynomials. Equations (2.1) and (2.2) give for $L_1^\alpha(x;c)$,

$$L_1^\alpha(x;c) = -\frac{1}{c+1}(x-2c-\alpha-1). \tag{2.6}$$

The CAL polynomials $L_n^\alpha(x;c,\mu)$ satisfy the same recurrence relation (2.1) with a shift μ on the monic polynomial of first degree, i.e.,

$$L_1^\alpha(x;c,\mu) = -\frac{1}{c+1}(x+\mu-2c-\alpha-1). \tag{2.7}$$

To obtain $L_1^\alpha(x;c,\mu)$ with (2.1) we have to impose the initial condition for the CAL polynomials

$$L_{-1}^\alpha(x;c,\mu) = \frac{\mu}{c+\alpha}, \qquad L_0^\alpha(x;c,\mu) = 0. \tag{2.8}$$

Even for $c+\alpha \to 0$ this initial condition in the recurrence relations (2.1) leads to (2.7).

We know two linearly independent solutions of (2.1) in terms of confluent hypergeometric functions:

$$u_n = \frac{(c+\alpha+1)_n}{(c+1)_n} {}_1F_1\left(\begin{matrix}-n-c\\1+\alpha\end{matrix};x\right) \quad \text{and} \quad v_n = {}_1F_1\left(\begin{matrix}-n-c-\alpha\\1-\alpha\end{matrix};x\right). \tag{2.9}$$

Writing the polynomials $L_n^\alpha(x;c,\mu)$ as a linear combination

$$L_n^\alpha(x;c,\mu) = Au_n + Bv_n \tag{2.10}$$

and using the initial condition (2.8) we obtain

$$A = \frac{1}{\Delta}\left[\frac{\mu}{c+\alpha}v_0 - v_{-1}\right] \quad \text{and} \quad B = -\frac{1}{\Delta}\left[\frac{\mu}{c+\alpha}u_0 - u_{-1}\right], \tag{2.11}$$

where Δ may be calculated using contiguous relations of confluent hypergeometric functions [7, pp. 253–254]

$$\Delta = u_{-1}v_0 - u_0v_{-1} = -\frac{\alpha}{c+\alpha}e^x. \tag{2.12}$$

With the help of the relation

$$\gamma_1 F_1\left(\begin{matrix}1-\gamma\\ \alpha\end{matrix};x\right)-\beta_1 F_1\left(\begin{matrix}-\gamma\\ \alpha\end{matrix};x\right)=(\gamma-\beta)_2F_2\left(\begin{matrix}-\gamma,\beta-\gamma+1\\ \alpha,\beta-\gamma\end{matrix};x\right), \tag{2.13}$$

we can write the CAL polynomials as

(2.14)

$$L_n^\alpha(x;c,\mu)=\frac{e^{-x}}{(c+1)_n}(1+\mathcal{T})\times\frac{\mu-c}{\alpha}(c+1)_n{}_2F_2\left(\begin{matrix}-c,\mu-c+1\\ 1+\alpha,\mu-c\end{matrix};x\right){}_1F_1\left(\begin{matrix}-n-c-\alpha\\ 1-\alpha\end{matrix};x\right).$$

This representation is the one we use to derive the fourth-order differential equation in §2.3. It is valid only for $\alpha\neq 0,\pm1,\pm2,\ldots$ but these restrictions can be removed by limiting processes.

Following the method in [2] we can find an explicit representation. We first transform the ${}_1F_1$ in (2.11) using Kummer's transformation [7, p. 253]

$${}_1F_1\left(\begin{matrix}a\\ c\end{matrix};x\right)=e^x{}_1F_1\left(\begin{matrix}c-a\\ c\end{matrix};-x\right). \tag{2.15}$$

Then we use the formula

$${}_1F_1\left(\begin{matrix}a\\ b\end{matrix};x\right){}_1F_1\left(\begin{matrix}c\\ d\end{matrix};-x\right)=\sum_{k=0}^{\infty}\frac{(a)_kx^k}{k!(b)_k}{}_3F_2\left(\begin{matrix}-k,1-k-b,c\\ 1-k-a,d\end{matrix};1\right) \tag{2.16}$$

for the four products in (2.10).

Applying the three-term relation [4, eq. (2), p. 15] to the four resulting ${}_3F_2(1)$ gives a sum of two terms that we can group to obtain the explicit form

$$L_n^\alpha(x;c,\mu)=\frac{(c+\alpha+1)_n}{n!}\sum_{k=0}^{n}\frac{(-n)_k}{(c+1)_k(c+\alpha+1)_k}\times{}_4F_3\left(\begin{matrix}k-n,c,c+\alpha,G+1\\ c+k+1,c+\alpha+k+1,G\end{matrix};1\right)x^k, \tag{2.17}$$

where

$$G=\frac{c(c+\alpha)}{\mu}. \tag{2.18}$$

Take care of the limiting processes when $n=k$ and when c or $c+\alpha=0$. The representation (2.17) can also be proved using the generating function (2.22).

2.2. Generating function, spectral measure. Let $F(x,w)$ be a generating function of the CAL polynomials $L_n^\alpha(x;c,\mu)$,

$$F(x,w)=\sum_{n=0}^{\infty}w^nL_n^\alpha(x;c,\mu). \tag{2.19}$$

The recurrence relation (2.1) and the initial condition (2.8) lead to the following differential equation for $F(x,w)$,

$$w(1-w)^2\frac{\partial}{\partial w}F(x,w)+[(1-w)(c-(c+\alpha+1)w)+xw]F(x,w)=c-\mu w. \tag{2.20}$$

The function $F(x,w)$ is normalized by the condition $F(x,0) = 1$ and due to the orthogonality of the $L_n^{\alpha}(x;c,\mu)$ we have the boundary condition

$$\int_0^{\infty} F(x,w)d\phi(x) = 1, \tag{2.21}$$

where $d\phi(x)$ is the spectral measure.

The solution of the differential equation (2.20) which is bounded at $w = 0$ is easily obtained, for $c > 0$, following the same method as in [12]:

$$\begin{aligned} F(x,w) &= w^{-c}(1-w)^{-\alpha-1}\exp\left[-\frac{x}{1-w}\right] \\ &\quad \times \int_0^w u^{c-1}(1-u)^{\alpha-1}(c-\mu u)\exp\left[\frac{x}{1-u}\right]du. \end{aligned} \tag{2.22}$$

Changing variables according to

$$u = \frac{\tau}{1+\tau}, \qquad w = \frac{z}{1+z}, \tag{2.23}$$

and integrating both sides of (2.22) with respect to $d\phi(x)$, taking (2.21) into account, leads to

$$\begin{aligned} z^c(1+z)^{-1-\alpha-c} &= \int_0^{\infty}\int_0^z \tau^{c-1}(1+\tau)^{-1-\alpha-c} \\ &\quad \times [c+\tau(c-\mu)]\exp[-x(z-\tau)]d\tau d\phi(x). \end{aligned} \tag{2.24}$$

Taking the Laplace transform of the above identity we obtain, for the Stieltjes transform of the measure $d\phi(x)$, the relationship

$$s(p) = \int_0^{\infty} \frac{d\phi(x)}{x+p} = \frac{\Psi(c+1,1-\alpha;p)}{\Psi(c,-\alpha;p) + (c-\mu)\Psi(c+1,1-\alpha;p)}, \tag{2.25}$$

which we can rewrite formaly, using the expression of the Tricomi function Ψ in terms of generalized hypergeometric functions [7, p. 257], on the form

$$s(p) = p^c\Psi(c+1,1-\alpha;p)\,{}_3F_1\left(\begin{matrix} c, c+\alpha, \frac{c(c+\alpha)}{\mu}+1 \\ \frac{c(c+\alpha)}{\mu} \end{matrix}; -\frac{1}{p}\right)^{-1}, \tag{2.26}$$

where the principal branch of the ${}_3F_1$ is considered one of the Ψ functions. The function $\Psi(a,b;p)$ having no zeros for $|\arg p| \le \pi$ the denominator of (2.25) has no zeros in this region at least for $\mu \le c$.

The CAL polynomials belong to the Laguerre–Hahn family of orthogonal polynomials and are of class zero [22]. It is easily verified that the Stieltjes transform $s(p)$ of the measure, calculated in (2.25), is a solution of the Riccati equation

$$ps'(p) = [\mu p - (c-\mu)(\alpha+c-\mu)]\,s^2(p) + [p+\alpha+2(c-\mu)]\,s(p) - 1. \tag{2.27}$$

The absolutely continuous part of the measure $d\phi(x)$ can be computed using the Perron–Stieltjes inversion formula. Details of the method are in [10] and we obtain

$$\phi'(x) = \frac{1}{\Gamma(c+1)\Gamma(c+\alpha+1)} \frac{x^{\alpha}e^{-x}}{|\Psi(c,-\alpha;xe^{i\pi}) + (c-\mu)\Psi(c+1,1-\alpha;xe^{i\pi})|^2}. \tag{2.28}$$

We can formally write

$$\phi'(x) = \frac{x^{\alpha+2c}e^{-x}}{\Gamma(c+1, c+\alpha+1)} \left| {}_3F_1\left(\begin{matrix} c, c+\alpha, \frac{c(c+\alpha)}{\mu}+1 \\ \frac{c(c+\alpha)}{\mu} \end{matrix}; -\frac{e^{i\pi}}{x}\right)\right|^{-2}. \tag{2.29}$$

The CAL polynomials $L_n^\alpha(x; c, \mu)$ satisfy the orthogonality relation, valid at least for $\mu \leq c$, $c \geq 0$, $\alpha + c > -1$,

$$\int_0^\infty L_n^\alpha(x; c, \mu) L_m^\alpha(x; c, \mu) d\phi(x) = \frac{(c+\alpha+1)_n}{(c+1)_n}\delta_{mn}. \tag{2.30}$$

2.3. Fourth-order differential equation. The CAL polynomials verify a differential equation of fourth order [3], [22]. One way to obtain this fourth-order differential equation is to start from their explicit form (2.14). The right-hand side of (2.14) is a sum of two products of a ${}_1F_1$ times a ${}_2F_2$. The ${}_1F_1$ are solutions of a second-order differential equation, but for the ${}_2F_2$ a third-order equation is expected. In fact, the ${}_2F_2$ involved in (2.14) are of the form

$$y(x) = {}_2F_2\left(\begin{matrix} b, e+1 \\ d, e \end{matrix}; x\right), \tag{2.31}$$

which can be shown with little effort to be a solution of the second-order differential equation

$$\begin{aligned} x[(b-e)x + e(d-e-1)]y''(x) - \{(b-e)x^2 + [e(2d-e-2) + b(1-d)]x \\ - ed(d-e-1)\}y'(x) - b[(b-e)x \\ + (e+1)(d-e-1)]y(x) = 0. \end{aligned} \tag{2.32}$$

So we can use the Orr method to obtain the differential equation satisfied by the products involved in (2.14) [23]. Let us notice that because of the second product being obtained by the transformation $\mathcal{T}$ of the first product, the fourth-order differential equation will have to be invariant under this transformation.

The function

$$y = e^{-x}{}_1F_1\left(\begin{matrix} -n-c-\alpha \\ 1-\alpha \end{matrix}; x\right)$$

is a solution of

$$xy'' + (1-\alpha+x)y' + (1+n+c)y = 0, \tag{2.33}$$

and the function

$$z = {}_2F_2\left(\begin{matrix} -c, \mu-c+1 \\ 1+\alpha, \mu-c \end{matrix}; x\right)$$

of

$$\begin{aligned} x[\mu x + (c-\mu)(c+\alpha-\mu)]z''(x) - \{\mu x^2 + [(c-\mu)(c+\alpha-\mu) - \alpha\mu]x \\ - (\alpha+1)(c-\mu)(c+\alpha-\mu)\}z'(x) \\ + c[\mu x + (c-\mu+1)(c+\alpha-\mu)]z(x) = 0. \end{aligned} \tag{2.34}$$

Changing the functions y and z to $y = fv$ and $z = gw$ with

$$f = x^{\frac{\alpha-1}{2}} e^{\frac{x}{2}}, \tag{2.35a}$$

$$g = [(c-\mu)(c+\alpha-\mu)+\mu x]^{\frac{1}{2}} x^{-\frac{\alpha+1}{2}} e^{\frac{x}{2}}, \tag{2.35b}$$

we obtain the normal form of the differential equations (2.33) and (2.34):

$$v'' + Iv = 0, \qquad w'' + Jw = 0. \tag{2.36}$$

The product $u = vw$ is a solution of the fourth-order differential equation (see [30, p. 146])

$$\frac{d}{dx}\left[\frac{u''' + 2(I+J)u' + (I'+J')u}{I-J}\right] = -(I-J)u. \tag{2.37}$$

Finally, we obtain the needed equation for the CAL polynomials setting $y(x) = fgu$.

Details of these calculations are very difficult to write explicitly and were achieved with the help of the MAPLE computer algebra [6]. Although with this help the fourth-order differential equation for the CAL polynomials is not easy to find. We give it as a *curiosity*:

$$c_4 y^{(4)}(x) + c_3 y^{(3)}(x) + c_2 y^{(2)}(x) + c_1 y^{(1)}(x) + c_0 y(x) = 0, \tag{2.38}$$

with

$$c_4 = x^2(2Ax^2 + Bx + 2C), \tag{2.39a}$$

$$c_3 = 2x(3Ax^2 + 2Bx + 5C), \tag{2.39b}$$

$$c_2 = -2Ax^4 + Dx^3 + Ex^2 + Fx + G, \tag{2.39c}$$

$$c_1 = -4Ax^3 + Hx^2 - 4Ax^3 + Ix + J, \tag{2.39d}$$

$$c_0 = n(n+1)(2Ax^2 + Kx + 2L), \tag{2.39e}$$

where

$$\begin{aligned}
A &= \mu^2(1+2n),\\
B &= \mu(2(1+4n)\mu^2 - (4(1+2n)(2c+\alpha)-1)\mu + 2c(3+4n)(c+\alpha)),\\
C &= (c-\mu)(c+\alpha-\mu)(2n\mu^2 - ((1+2n)(2c+\alpha)+1)\mu + 2c(n+1)(c+\alpha)),\\
D &= -\mu(2(1+4n)\mu^2 - (2(1+2n)(8c+4\alpha+1+2n)-1)\mu + 2c(3+4n)(c+\alpha)),\\
E &= -4n\mu^4 + ((1+4n)(4n+12c+6\alpha+3)+3)\mu^3\\
&\quad -2(1+2n)((2c+\alpha)(4n+12c+6\alpha+2) - 2c(c+\alpha)-1)\mu^2\\
&\quad +c(c+\alpha)((3+4n)(4n+12c+6\alpha+1)+3)\mu - 4c^2(c+\alpha)^2(n+1),\\
F &= 8n(2c+\alpha+n+1)\mu^4\\
&\quad -((1+4n)((2c+\alpha)(4n+12c+6\alpha+3) - 8c(c+\alpha) - 1/2) + 10c + 5\alpha + 5/2)\mu^3\\
&\quad +((1+4n)((\alpha^2 + 6c(c+\alpha)(2n+8c+4\alpha+3/2) - (2c+\alpha)(12c(c+\alpha)+1/2))\\
&\qquad +(2\alpha^2 + 6c(c+\alpha) - 1/4)(4c+2\alpha+23/6) - 25/6\alpha^2 + 71/24)\mu^2\\
&\quad -2c(c+\alpha)((1+4n)((2c+\alpha)(2n+4c+2\alpha+3/2) + 2\alpha^2 - 1/4)\\
&\quad +(2c+\alpha)(8c+4\alpha+5/2) + 2\alpha^2 - 7/4)\mu\\
&\quad +8c^2(c+\alpha)^2(n+1)(n+2c+\alpha),
\end{aligned}$$

$$
\begin{aligned}
G &= -2(\alpha-2)(\alpha+2)(c-\mu)(c+\alpha-\mu)(2(c-\mu)(c+\alpha-\mu)n \\
&\quad -(2c+\alpha+1)\mu+2c(c+\alpha)), \\
H &= 2\mu(-2(5n+1)\mu^2+((1+2n)(n+12c+6\alpha+1/2)-3/2)\mu-2c(c+\alpha)(5n+4)), \\
I &= -12n\mu^4+2((n+1/5)(8n+40c+20\alpha+7/5)+93/25)\mu^3 \\
&\quad +2(-(1+2n)((2c+\alpha)(4n+12c+6\alpha+2)+10c(c+\alpha)-1)-6c-3\alpha)\mu^2 \\
&\quad +2c(c+\alpha)((4/5+n)(8n+40c+20\alpha+33/5)+93/25)\mu-12c^2(c+\alpha)^2(n+1), \\
J &= 4(\mu-c)(\mu-c-\alpha)(3n(n+2c+\alpha+1)\mu^2+(-1/4(1+2n)((2c+\alpha)(6n+8c \\
&\quad +4\alpha+3)+8c(c+\alpha)+8)-9/2c-9/4\alpha)\mu+3c(c+\alpha)(n+1)(n+2c+\alpha)), \\
K &= \mu(2(4n-1)\mu^2-(4(1+2n)(2c+\alpha)-3)\mu+2c(4n+5)(c+\alpha)), \\
L &= (c-\mu)(c+\alpha-\mu)(2(n-1)\mu^2-((1+2n)(2c+\alpha)+6)\mu+2c(n+2)(c+\alpha)).
\end{aligned}
$$

Note the invariance of the differential equation (2.38) by the transformation $\mathcal{T}$ defined in (2.5).

2.4. Particular cases. We now give the different results corresponding to limiting cases of special interest.

2.4.1. Limit $c = 0$. In this limit we obtain from (2.14) the co-recursive Laguerre polynomials.

(2.40)
$$
L_n^{\alpha}(x;0,\mu) = \frac{e^{-x}}{\alpha}\left\{\frac{(\alpha-\mu)(\alpha+1)_n}{n!}\,{}_2F_2\left(\begin{matrix}-\alpha,\mu-\alpha+1\\ 1-\alpha,\mu-\alpha\end{matrix};x\right)\,{}_1F_1\left(\begin{matrix}-n\\ 1+\alpha\end{matrix};x\right)\right.
$$
$$
\left.+\mu\,{}_1F_1\left(\begin{matrix}-n-\alpha\\ 1-\alpha\end{matrix};x\right)\right\}.
$$

The limit $\mu = 0$ in (2.40) leads back to the classical Laguerre polynomials. An explicit form is

(2.41)
$$
L_n^{\alpha}(x;0,\mu) = \frac{(\alpha+1)_n}{n!}\sum_{k=0}^{n}\frac{(-n)_k}{k!(1+\alpha)_k}
$$
$$
\times\left[1+\frac{\mu(k-n)}{(1+k)(1+\alpha+k)}\,{}_3F_2\left(\begin{matrix}1+k-n,1+\alpha,1\\ k+\alpha+2,k+2\end{matrix};1\right)\right]x^k,
$$

and the corresponding absolutely continuous part of the measure is given for $\mu \le 0$, $\alpha > -1$, by

(2.42)
$$
\phi'(x) = \frac{x^{\alpha}e^{-x}}{\Gamma(1+\alpha)}\left|1-\mu\Psi(1,1-\alpha;xe^{-i\pi})\right|^{-2},
$$

where the limit $\mu = 0$ is also straightforward.

It is easy to see that the differential equation (2.38) satisfied by the co-recursive Laguerre polynomials can be factorized in the limit $c = 0$ to obtain the fourth-order factorized (2+2) differential equation

(2.43)
$$
\begin{aligned}
&[xA(x)\mathrm{D}^2+\{(2+\alpha-x)A(x)-xB(x)\}\mathrm{D} \\
&\quad +(n-1)A(x)+(x-\alpha-1)B(x)+C(x)] \\
&\times\left[x\mathrm{D}^2+(x+1-\alpha)\mathrm{D}+n+1\right]L_n^{\alpha}(x;0,\mu)=0,
\end{aligned}
$$

where $\mathrm{D} \equiv d/dx$ and

$$A(x) = 3x + 2(x - \alpha + \mu)\{2n(x - \alpha + \mu) + x - \alpha - 1\}, \tag{2.44a}$$

$$B(x) = 1 + 2x - 2\alpha + 2(1 + 4n)(x - \alpha + \mu), \tag{2.44b}$$

$$C(x) = (1 + 2x - 2\alpha)\{1 + \alpha - x - 2n(x - \alpha + \mu)\} + 3x(1 + 4n). \tag{2.44c}$$

The comparison with the differential equation given in [27, eq. 34–35] requires some attention because of a few misprints.

2.4.2. Limit $c = -\alpha$. In this limit we obtain a special class of CAL polynomials corresponding to the associated Laguerre polynomials for which

$$L_n^{\alpha}(x; -\alpha) = \frac{n!}{(1-\alpha)_n} L_n^{-\alpha}(x). \tag{2.45}$$

We can write the $L_n^{\alpha}(x; -\alpha, \mu)$ as

(2.46)

$$L_n^{\alpha}(x; -\alpha, \mu) = \frac{n!}{(1-\alpha)_n} \frac{e^{-x}}{-\alpha} \left\{ \frac{(-\alpha - \mu)(1-\alpha)_n}{n!} {}_2F_2\left(\begin{matrix}\alpha, \mu + \alpha + 1 \\ 1 + \alpha, \mu + \alpha\end{matrix}; x\right) \right.$$
$$\left. \times {}_1F_1\left(\begin{matrix}-n \\ 1 - \alpha\end{matrix}; x\right) + \mu {}_1F_1\left(\begin{matrix}-n + \alpha \\ 1 + \alpha\end{matrix}; x\right) \right\},$$

which gives (2.45) in the limit $\mu = 0$. Except for the global factor $n!/(1-\alpha)_n$, (2.46) is obtained from (2.40) changing α to $-\alpha$. The corresponding measure and differential equation are obtained in the same way as §2.4.1. An explicit form is

(2.47)

$$L_n^{\alpha}(x; -\alpha, \mu) = \sum_{k=0}^{n} \frac{(-n)_k}{k!(1-\alpha)_k}$$
$$\times \left[1 + \frac{\mu(k-n)}{(1+k)(1-\alpha+k)} {}_3F_2\left(\begin{matrix}1 + k - n, 1 - \alpha, 1 \\ k - \alpha + 2, k + 2\end{matrix}; 1\right) \right] x^k.$$

2.4.3. Limit $\mu = 0$. In this limit we obtain the associated Laguerre polynomials studied in [2] and [12]:

$$L_n^{\alpha}(x, c) = \frac{e^{-x}}{(c+1)_n} (1 + \mathcal{T}) \frac{(c+\alpha)_{n+1}}{\alpha} {}_1F_1\left(\begin{matrix}1 - c - \alpha \\ 1 - \alpha\end{matrix}; x\right) {}_1F_1\left(\begin{matrix}-n - c \\ 1 + \alpha\end{matrix}; x\right), \tag{2.48}$$

with the measure

$$\phi'(x) = \frac{x^{\alpha} e^{-x}}{\Gamma(1 + c, 1 + c + \alpha)} \left| \Psi(c, 1 - \alpha; x e^{-i\pi}) \right|^{-2}. \tag{2.49}$$

An explicit form is

$$L_n^{\alpha}(x; c) = \frac{(c + \alpha + 1)_n}{n!} \sum_{k=0}^{n} \frac{(-n)_k}{(c+1)_k (c + \alpha + 1)_k}$$
$$\times {}_3F_2\left(\begin{matrix}k - n, c, c + \alpha \\ c + k + 1, c + \alpha + k + 1\end{matrix}; 1\right) x^k, \tag{2.50}$$

The limit $c = 0$ leads back to the Laguerre polynomial case. The coefficients of the differential equation (2.38) satisfied by the associated Laguerre polynomials are now very simple:

$$c_4 = x^2,\quad c_3 = 5x,\quad c_2 = -x(x-2F) - \alpha^2 + 4,\quad c_1 = 3(F-x),\quad c_0 = n(n+2), \tag{2.51}$$

with $F = n + 2c + \alpha$. This differential equation was first given by Hahn [9, eq. 22]. See [25] and [27] for the special factorizable case $c = 1$ and [5] and [26] when c is an integer.

2.4.4. Limit $\mu = c$. In this limit we obtain the so-called *zero-related* Laguerre polynomials studied in [12]. Note the symmetry $\mathcal{T}$ of the monic polynomials is now broken:

$$\mathcal{L}_n^{\alpha}(x,c) = e^{-x}\left\{\frac{(c+\alpha+1)_n}{(c+1)_n}\,{}_1F_1\left(\begin{matrix}-c-\alpha\\ -\alpha\end{matrix};x\right){}_1F_1\left(\begin{matrix}-n-c\\ 1+\alpha\end{matrix};x\right)\right.$$
$$\left. -\frac{c}{\alpha(\alpha+1)}\,x\,{}_1F_1\left(\begin{matrix}1-c\\ 2+\alpha\end{matrix};x\right){}_1F_1\left(\begin{matrix}-n-c-\alpha\\ 1-\alpha\end{matrix};x\right)\right\}, \tag{2.52}$$

and the measure

$$\phi'(x) = \frac{x^{\alpha}e^{-x}}{\Gamma(1+c,1+c+\alpha)}\left|\Psi(c,-\alpha;xe^{-i\pi})\right|^{-2}. \tag{2.53}$$

Again the limit $c = 0$ leads back to the Laguerre polynomial case.

The explicit form (2.17) simplifies into

$$\mathcal{L}_n^{\alpha}(x,c) = \frac{(c+\alpha+1)_n}{n!}\sum_{k=0}^{n}\frac{(-n)_k}{(c+1)_k(c+\alpha+1)_k}$$
$$\times{}_3F_2\left(\begin{matrix}k-n,c,c+\alpha+1\\ c+k+1,c+\alpha+k+1\end{matrix};1\right)x^k. \tag{2.54}$$

The coefficients of the differential equation satisfied by the zero-related Laguerre polynomials are

$$c_4 = x^2(2(2n+1)x + D),\qquad c_3 = 2x(3(2n+1)x + 2D), \tag{2.55a}$$
$$c_2 = -2(2n+1)x^3 + (8n(F+1) + 8c + D)x^2$$
$$- (4(\alpha^2-1)n - 2\alpha^2 - D(4c+1) - 1)x - 1/4D(D^2-9), \tag{2.55b}$$
$$c_1 = 8(2n+1)x^2 - 4(2n(F+1) + 2c - D)x - 2D(n(D+3) + 6c + 2D), \tag{2.55c}$$
$$c_0 = n(n+1)(2(2n+1)x + 3D), \tag{2.55d}$$

with

$$F = n + 2c + \alpha,\qquad D = 1 + 2\alpha, \tag{2.56}$$

and are no longer invariant under the transformation $\mathcal{T}$.

2.4.5. Limit $\mu = c + \alpha$. This is a new simple case of CAL polynomials lacking in [12]:

$$\mathcal{L}_n^{\alpha}(x,c) = e^{-x}\left\{{}_1F_1\left(\begin{matrix}-c\\ \alpha\end{matrix};x\right){}_1F_1\left(\begin{matrix}-n-c-\alpha\\ 1-\alpha\end{matrix};x\right)\right.$$
$$\left. + \frac{(c+\alpha)(c+\alpha+1)_n}{\alpha(1-\alpha)(c+1)_n}x\,{}_1F_1\left(\begin{matrix}1-c-\alpha\\ 2-\alpha\end{matrix};x\right){}_1F_1\left(\begin{matrix}-n-c\\ 1+\alpha\end{matrix};x\right)\right\}, \tag{2.57}$$

and using [7, (10), p. 258] the measure is obtained:

$$\phi'(x) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(1+c,1+c+\alpha)} \left|\Psi(c+1,2-\alpha;xe^{-i\pi})\right|^{-2}. \tag{2.58}$$

In this case the limit $c = 0$ does not lead back to the Laguerre polynomial case but to the co-recursive Laguerre one with $\mu = \alpha$.

The explicit form is

$$\mathfrak{L}_n^{\alpha}(x,c) = \frac{(c+\alpha+1)_n}{n!} \sum_{k=0}^{n} \frac{(-n)_k}{(c+1)_k(c+\alpha+1)_k} \times {}_3F_2\left(\begin{matrix} k-n, c+1, c+\alpha \\ c+k+1, c+\alpha+k+1 \end{matrix}; 1\right) x^k. \tag{2.59}$$

The coefficients of the differential equation (2.38) satisfied by the polynomials $\mathfrak{L}_n^{\alpha}(x,c)$ are obtained from (2.55), (2.56), changing only $D = 1+2\alpha$ by $D = 1-2\alpha$.

3. The case of Jacobi polynomials. We now present some results on the CAJ polynomials. The recurrence relation of the associated Jacobi polynomials $P_n^{\alpha,\beta}(x;c)$ is [31]

(3.1)
$$\begin{aligned}(2n+2c+\alpha+\beta+1)&[(2n+2c+\alpha+\beta+2)(2n+2c+\alpha+\beta)x+\alpha^2-\beta^2]p_n \\ &= 2(n+c+1)(n+c+\alpha+\beta+1)(2n+2c+\alpha+\beta)p_{n+1} \\ &\quad +2(n+c+\alpha)(n+c+\beta)(2n+2c+\alpha+\beta+2)p_{n-1}.\end{aligned}$$

We can note the invariance of the recurrence relation (3.2) under the transformation $\mathcal{T}'$ defined by

$$\mathcal{T}'(c,\alpha,\beta) = (c+\alpha+\beta,-\alpha,-\beta). \tag{3.2}$$

All polynomials satisfying (3.2), for which the initial conditions are symmetric in α and β and invariant under $\mathcal{T}'$ have the property

$$p_n^{-\alpha,-\beta}(x,c+\alpha+\beta) = p_n^{\alpha,\beta}(x,c). \tag{3.3}$$

As in [31] we use the more convenient shifted polynomials defined as

$$R_n^{\alpha,\beta}(x;c) = P_n^{\alpha,\beta}(2x-1;c). \tag{3.4}$$

Due also to the properties of the recurrence relation (3.2) we have

$$R_n^{\alpha,\beta}(x;c) = (-1)^n R_n^{\beta,\alpha}(1-x;c). \tag{3.5}$$

3.1. Explicit representation for the CAJ polynomials. A solution of the recurrence relation satisfied by the $R_n^{\alpha,\beta}(x;c)$ in terms of the hypergeometric function is [21, p. 280]

$$u_n = \frac{(c+\alpha+1)_n}{(c+1)_n} {}_2F_1\left(\begin{matrix} -n-c, n+c+\alpha+\beta+1 \\ 1+\alpha \end{matrix}; 1-x\right), \tag{3.6}$$

and another linearly independent solution is given by

$$v_n = \mathcal{T}' u_n = \frac{(c+\beta+1)_n}{(c+\alpha+\beta+1)_n} {}_2F_1\left(\begin{matrix}-n-c-\alpha-\beta, n+c+1 \\ 1-\alpha\end{matrix}; 1-x\right). \tag{3.7}$$

The functions u_n and $x^{-\beta}(1-x)^{-\alpha}v_n$ are two independent solutions of the second-order differential equation

$$\begin{aligned} x(1-x)y''(x) + [1+\beta-(\alpha+\beta+2)x]y'(x) \\ + (n+c)(n+c+\alpha+\beta+1)y(x) = 0. \end{aligned} \tag{3.8}$$

The associated Jacobi polynomials are defined by (3.2) and the initial condition

$$P_{-1}^{\alpha,\beta}(x;c) = 0, \qquad P_0^{\alpha,\beta}(x;c) = 1. \tag{3.9}$$

For $P_1^{\alpha,\beta}(x;c)$, this gives

$$\begin{aligned} P_1^{\alpha,\beta}(x;c) = \frac{(2c+\alpha+\beta+1)(2c+\alpha+\beta+2)}{2(c+1)(c+\alpha+\beta+1)} \\ \times \left[x + \frac{\alpha^2-\beta^2}{(2c+\alpha+\beta)(2c+\alpha+\beta+2)}\right]. \end{aligned} \tag{3.10}$$

The CAJ polynomials $P_n^{\alpha,\beta}(x;c,\mu)$ satisfy the recurrence relation (3.2) with a shift μ on the first monic polynomial. This corresponds to the initial condition on the shifted CAJ polynomials $R_n^{\alpha,\beta}(x;c,\mu)$:

$$R_{-1}^{\alpha,\beta}(x;c,\mu) = D = -\frac{(2c+\alpha+\beta)(2c+\alpha+\beta+1)}{2(c+\alpha)(c+\beta)}\mu, \qquad R_0^{\alpha,\beta}(x;c,\mu) = 1. \tag{3.11}$$

If $c+\alpha \to 0$ or $c+\beta \to 0$ this initial condition in (3.2) nevertheless leads to a shift μ on the value of x in $P_1^{\alpha,\beta}(x;c)$.

As in §2.1 writing

$$R_n^{\alpha,\beta}(x;c,\mu) = Au_n + Bv_n \tag{3.12}$$

and using (3.11) we obtain

$$A = \frac{1}{\Delta}[Dv_0 - v_{-1}] \quad \text{and} \quad B = -\frac{1}{\Delta}[Du_0 - u_{-1}], \tag{3.13}$$

where Δ is easily calculated using the fact that u_n and $x^{-\beta}(1-x)^{-\alpha}v_n$ are two independent solutions of (3.8),

$$\Delta = u_{-1}v_0 - u_0 v_{-1} = -\frac{\alpha(2c+\alpha+\beta)}{(c+\alpha)(c+\beta)}. \tag{3.14}$$

The condition $\Delta \neq 0$ leads to $\alpha \neq 0$ and $2c+\alpha+\beta \neq 0$. We can note the invariance of D under $\mathcal{T}'$ and can note that $B = \mathcal{T}'A$.

Grouping the two ${}_2F_1$ involved in the expression (3.13) of A gives

$$A = \frac{c+\alpha+\beta-D(c+\beta)}{\alpha(2c+\alpha+\beta)} {}_3F_2\left(\begin{matrix}-c-\alpha-\beta, c, F+1 \\ 1-\alpha, F\end{matrix}; 1-x\right), \tag{3.15}$$

with

$$F = \frac{c[D(c+\beta) - c - \alpha - \beta]}{D(c+\beta) + c}, \tag{3.16}$$

and the CAJ polynomials could be written as

(3.17)

$$R_n^{\alpha,\beta}(x;c,\mu) = (1+\mathcal{T}')\frac{c+\alpha+\beta-D(c+\beta)}{\alpha(2c+\alpha+\beta)}(c+\alpha)\frac{(c+\alpha+1)_n}{(c+1)_n}$$
$$\times {}_3F_2\left(\begin{matrix}-c-\alpha-\beta, c, F+1\\ 1-\alpha, F\end{matrix}; 1-x\right) {}_2F_1\left(\begin{matrix}-n-c, n+c+\alpha+\beta+1\\ 1+\alpha\end{matrix}; 1-x\right).$$

We will use this expression of the CAJ polynomials in §3.4 to obtain a fourth-order differential equation of them.

Transforming the ${}_2F_1(1-x)$ in (3.12) by [7, eq. 1, p. 108] one obtains, with little algebra,

(3.18)

$$R_n^{\alpha,\beta}(x;c,\mu) = (1+\mathcal{T}')\frac{(-1)^n c(c+\alpha)}{\beta(2c+\alpha+\beta)}$$
$$\times \frac{(c+\alpha+1)_n}{(c+\alpha+\beta+1)_n} {}_2F_1\left(\begin{matrix}-n-c-\alpha-\beta, n+c+1\\ 1-\beta\end{matrix}; x\right)$$
$$\times \left[\frac{c+\beta}{c} D\,{}_2F_1\left(\begin{matrix}-c, c+\alpha+\beta+1\\ 1+\beta\end{matrix}; x\right) - {}_2F_1\left(\begin{matrix}1-c, c+\alpha+\beta\\ 1+\beta\end{matrix}; x\right)\right].$$

This formula generalizes the formula of [31, eq. 28] to the case of the CAJ polynomials. As the explicit form of the CAL polynomials (2.14) the representation (3.17) and (3.18) are valid only for $\alpha \neq 0, \pm 1, \pm 2\ldots$ and $\beta \neq 0, \pm 1, \pm 2\ldots$ but can be extended by limiting processes.

We obtain an explicit formula following the method in [31]. We first use [7, eq. 14, p. 87] for each product of ${}_2F_1$ in (3.18) to obtain four series involving gamma functions and a ${}_4F_3$. For two of them we use [4, eq. 1, p. 56]. The next step is to use [4, eq. 3, p. 62] twice for each ${}_4F_3$. After numerous cancellations only two series of ${}_4F_3$ remain. We can group to obtain the following explicit form:

(3.19)

$$R_n^{\alpha,\beta}(x;c,\mu) = 7(-1)^n \frac{(2c+\alpha+\beta+1)_n(\beta+c+1)_n}{n!(c+\alpha+\beta+1)_n} \sum_{k=0}^{n} \frac{(-n)_k(n+2c+\alpha+\beta+1)_k}{(c+1)_k(c+\beta+1)_k}$$
$$\times {}_5F_4\left(\begin{matrix}k-n, n+k+2c+\alpha+\beta+1, c, c+\beta, G+1\\ c+k+1, c+\beta+k+1, 2c+\alpha+\beta+1, G\end{matrix}; 1\right) x^k,$$

where

$$G = \frac{2c(c+\beta)(2c+\alpha+\beta)}{2c(c+\beta)+\mu(2c+\alpha+\beta)(2c+\alpha+\beta+1)}. \tag{3.20}$$

3.2. Generating function. One can obtain a generating function of the CAJ polynomials following the same strategy as in [31] for the associated one. Let $\mathcal{G}(x,w)$ be a generating function of $R_n^{\alpha,\beta}(x;c,\mu)$:

$$\mathcal{G}(x,w) = \sum_{n=0}^{\infty} \frac{(c+1)_n(c+\alpha+\beta+1)_n}{n!(2c+\alpha+\beta+2)_n} w^n R_n^{\alpha,\beta}(x;c,\mu). \tag{3.21}$$

Starting from the form (3.17) for the $R_n^{\alpha,\beta}(x;c,\mu)$ it follows that

$$\begin{aligned}\mathcal{G}(x,w)&=(1+\mathcal{T}')\frac{c+\alpha+\beta-D(c+\beta)}{\alpha(2c+\alpha+\beta)}(c+\alpha)_3F_2\left(\begin{matrix}-c-\alpha-\beta,c,F+1\\1-\alpha,F\end{matrix};1-x\right)\\&\quad\times\sum_{n=0}^{\infty}\frac{(c+\alpha+1)_n(c+\alpha+\beta+1)_n}{n!(2c+\alpha+\beta+2)_n}w^n{}_2F_1\left(\begin{matrix}-n-c,n+c+\alpha+\beta+1\\1+\alpha\end{matrix};1-x\right),\end{aligned}$$

and using [31, Thm. 4] we obtain

(3.22)

$$\begin{aligned}\mathcal{G}(x,w)&=(1+\mathcal{T}')\frac{c+\alpha+\beta-D(c+\beta)}{\alpha(2c+\alpha+\beta)}\\&\quad\times(c+\alpha)\left[\frac{2}{w(Z_2+1)}\right]^{c+\alpha+\beta+1}{}_3F_2\left(\begin{matrix}-c-\alpha-\beta,c,F+1\\1-\alpha,F\end{matrix};1-x\right)\\&\quad\times{}_2F_1\left(\begin{matrix}-c,c+\alpha+\beta+1\\1+\alpha\end{matrix};\frac{1-Z_1}{2}\right){}_2F_1\left(\begin{matrix}c+\alpha+1,c+\alpha+\beta+1\\2c+\alpha+\beta+2\end{matrix};\frac{2}{1+Z_2}\right),\end{aligned}$$

where

$$Z_1=\frac{1-\sqrt{(1+w)^2-4wx}}{w},\qquad Z_2=\frac{1+\sqrt{(1+w)^2-4wx}}{w},\tag{3.23}$$

which generalize the already exotic generating function [31, eq. 75].

3.3. Spectral measure. The Stieltjes transform of the measure of the shifted associated Jacobi polynomials $R_n^{\alpha,\beta}(x;c)$ is [31, eqs. 63–64]

$$s(p)=\frac{1}{p}{}_2F_1\left(\begin{matrix}c+1,c+\beta+1\\2c+\alpha+\beta+2\end{matrix};\frac{1}{p}\right){}_2F_1\left(\begin{matrix}c,c+\beta\\2c+\alpha+\beta\end{matrix};\frac{1}{p}\right)^{-1}.\tag{3.24}$$

The CAJ polynomials $R_n^{\alpha,\beta}(x;c,\mu)$ satisfy the same recurrence relations as the $R_n^{\alpha,\beta}(x;c)$ with a shift μ on the first monic polynomials

$$R_1^{\alpha,\beta}(x;c,\mu)-R_1^{\alpha,\beta}(x;c)=\frac{(2c+\alpha+\beta+1)(2c+\alpha+\beta+2)}{2(c+1)(c+\alpha+\beta+1)}\mu.\tag{3.25}$$

Using continued J-fractions [14], [8], [28] whose denominators are $R_n^{\alpha,\beta}(x;c,\mu)$ and $R_n^{\alpha,\beta}(x;c)$, we can derive the following for the Stieltjes transform of the measure of the CAJ polynomials:

$$\begin{aligned}s(p;\mu)&=s(p)\left(1+\frac{\mu}{2}s(p)\right)^{-1}\\&=\frac{{}_2F_1\left(\begin{matrix}c+1,c+\beta+1\\2c+\alpha+\beta+2\end{matrix};\frac{1}{p}\right)}{p\,{}_2F_1\left(\begin{matrix}c,c+\beta\\2c+\alpha+\beta\end{matrix};\frac{1}{p}\right)+\frac{\mu}{2}{}_2F_1\left(\begin{matrix}c+1,c+\beta+1\\2c+\alpha+\beta+2\end{matrix};\frac{1}{p}\right)},\end{aligned}\tag{3.26}$$

which using contiguous relations we can also write as

(3.27)

$$s(p;\mu) = {}_2F_1\left(\begin{matrix}c+1, c+\beta+1\\ 2c+\alpha+\beta+2\end{matrix}; \frac{1}{p}\right) \times \left[\left(\frac{c+\alpha}{2c+\alpha+\beta} - \frac{2c+\alpha+\beta+1}{2c}\mu\right) {}_2F_1\left(\begin{matrix}c, c+\beta\\ 2c+\alpha+\beta+1\end{matrix}; \frac{1}{p}\right) + \left(\frac{c+\beta}{2c+\alpha+\beta} + \frac{2c+\alpha+\beta+1}{2c}\mu\right) {}_2F_1\left(\begin{matrix}c, c+\beta+1\\ 2c+\alpha+\beta+1\end{matrix}; \frac{1}{p}\right)\right]^{-1}.$$

Grouping the ${}_2F_1$ in (3.28) gives the compact formula

$$s(p;\mu) = {}_2F_1\left(\begin{matrix}c+1, c+\beta+1\\ 2c+\alpha+\beta+2\end{matrix}; \frac{1}{p}\right) \left[{}_3F_2\left(\begin{matrix}c, c+\beta, G+1\\ 2c+\alpha+\beta+1, G\end{matrix}; \frac{1}{p}\right)\right]^{-1}, \tag{3.28}$$

where G is given by (3.20). A sufficient condition for the positivity of the denominator in (3.28) on $(1,\infty)$ is

$$c \geq 0, \quad c > -\beta, \quad \alpha > -1, \quad \mu \geq -\frac{2c(c+\beta)}{(2c+\alpha+\beta)(2c+\alpha+\beta+1)}, \tag{3.29}$$

but other conditions are possible.

To obtain the absolutely continuous part of the spectral measure we need to evaluate $s^+(p;\mu) - s^-(p;\mu)$, where $s^\pm$ are the values of s above and below the cut [0,1]. Using the analytic continuation [7, eq. 2, p. 108] for each ${}_2F_1$ in (3.26) we find, for the spectral measure of the CAJ polynomials,

(3.30)

$$\begin{aligned}\phi'(x) &= (1-x)^\alpha x^{\beta+2c} \left| {}_2F_1\left(\begin{matrix}c, c+\beta\\ 2c+\alpha+\beta\end{matrix}; \frac{e^{i\pi}}{x}\right) + \frac{\mu}{2x} {}_2F_1\left(\begin{matrix}c+1, c+\beta+1\\ 2c+\alpha+\beta+2\end{matrix}; \frac{e^{i\pi}}{x}\right)\right|^{-2} \\ &= (1-x)^\alpha x^{\beta+2c} \left| {}_3F_2\left(\begin{matrix}c, c+\beta, G+1\\ 2c+\alpha+\beta+1, G\end{matrix}; \frac{e^{i\pi}}{x}\right)\right|^{-2},\end{aligned}$$

valid at least under the conditions (3.29).

3.4. Fourth-order differential equation. The method used to obtain the differential equation satisfied by the $R_n^{\alpha,\beta}(x;c,\mu)$ is the same as in §2.3. In (3.17) the hypergeometric function ${}_2F_1$ is solution of (3.8) and the ${}_3F_2$ is of the form

$${}_3F_2\left(\begin{matrix}a, b, e+1\\ d, e\end{matrix}; x\right),$$

which is also a solution of the second-order differential equation

(3.31)

$$\begin{aligned}&x(x-1)\left[(a-e)(b-e)x + e(d-e-1)\right] y''(x) \\ &+ \left\{(a-e)(b-e)(a+b+1)x^2 + \left[e(a+b+1)(2d-e-2) - d(ab+e^2) + ab\right] x \right. \\ &\left. + de(e-d+1)\right\} y'(x) + ab\left[(a-e)(b-e)x + (e+1)(d-e-1)\right] y(x) = 0.\end{aligned}$$

We don't write the fourth-order differential equation hardly obtained by symbolic MAPLE computation. The coefficients are at most of degree eight in x, and it would take several pages to write them. We give the results only in the following limiting cases.

3.5. Particular cases.

3.5.1. Laguerre case limit. The limit giving the CAL polynomial case is obtained by the replacement

$$\left.\begin{array}{rcl} x & \to & 1-\frac{2x}{\beta} \\ \mu & \to & +\frac{2\mu}{\beta} \end{array}\right\} \qquad \beta \to \infty, \tag{3.32}$$

in $P_n^{\alpha,\beta}(x;c,\mu)$. The representation (3.17) is the more suitable to obtain the form of the CAL polynomials (2.14) using Kummer's transformation (2.15) for one of the confluent hypergeometric functions and his generalization

$$ {}_2F_2\left(\begin{matrix}a, e+1\\ c, e\end{matrix};x\right) = e^x\, {}_2F_2\left(\begin{matrix}c-a-1, \frac{e(c-a-1)}{e-a}+1\\ c, \frac{e(c-a-1)}{e-a}\end{matrix};-x\right) \tag{3.33}$$

for one of the ${}_2F_2$. Note that (3.33) gives (2.15) in the limit $e \to \infty$.

3.5.2. Limit $c = 0$. In this limit we obtain the co-recursive Jacobi polynomials. An explicit form is

(3.34)

$$R_n^{\alpha,\beta}(x;\mu) = (-1)^n \frac{(\beta+1)_n}{n!} \sum_{k=0}^{n} \frac{(-n)_k (n+\alpha+\beta+1)_k}{k!(\beta+1)_k} x^k$$
$$\times \left\{1 + \frac{\mu(k-n)(n+k+\alpha+\beta+1)}{2(k+1)(\beta+k+1)} {}_4F_3\left(\begin{matrix}k-n+1, n+k+\alpha+\beta+2, \beta+1, 1\\ k+2, \beta+k+2, \alpha+\beta+2\end{matrix};1\right)\right\},$$

and the spectral measure is given by

$$\phi'(x) = (1-x)^\alpha x^\beta \left|1 + \frac{\mu}{2x} {}_2F_1\left(\begin{matrix}1, \beta+1\\ \alpha+\beta+2\end{matrix};\frac{e^{i\pi}}{x}\right)\right|^{-2}. \tag{3.35}$$

The limit $\mu = 0$ leads back to the Jacobi polynomials.

The fourth-order differential equation satisfied by the co-recursive Laguerre polynomials can be factorized in the limit $c = 0$ to obtain, as in [27], the factorized (2+2) differential equation

$$\begin{aligned} 0 = &\left[(1-x^2)A(x)\mathrm{D}^2 + \left\{(\beta-\alpha-(\alpha+\beta+4)x)A(x) - (1-x^2)B(x)\right\}\mathrm{D}\right. \\ &+ \left\{n(n+\alpha+\beta+1) - (\alpha+\beta+2)\right\}A(x) \\ &\left. + \left\{\beta-\alpha-(\alpha+\beta+2)x\right\}B(x) + C(x)\right] \\ &\times \left[(1-x^2)\mathrm{D}^2 + \{(\alpha+\beta-2)x + \alpha - \beta\}\mathrm{D}\right. \\ &\left. + n(n+\alpha+\beta+1) + \alpha + \beta\right] R_n^{\alpha,\beta}(x;\mu), \end{aligned} \tag{3.36}$$

where

$$\begin{aligned} A(x) = {} & 2(\alpha+\beta)^2(2n+1)(n+\alpha+\beta+1/2)x^2 + 2(\alpha+\beta)(4n(n+\alpha+\beta+1) \\ & \times(-\mu(1+\alpha+\beta)+\alpha-\beta) + (1+\alpha+\beta)(-\mu(\alpha+\beta+2)+2\alpha-2\beta))x \\ & +4n(n+\alpha+\beta+1)(-\mu(1+\alpha+\beta)+\alpha-\beta)^2 \\ & -(\alpha+\beta)(-2\mu(1+\alpha+\beta)\times(\beta-\alpha) - 2(\beta-\alpha)^2 - 3\alpha - 3\beta) \\ B(x) = {} & -(\alpha+\beta)((\alpha+\beta)(8n(n+\alpha+\beta+1)+3\alpha+3\beta)x + 8n(n+\alpha+\beta+1) \end{aligned}$$

$$\times(-\mu(1+\alpha+\beta)+\alpha-\beta)-2(\alpha+\beta+2)(1+\alpha+\beta)\mu+(\alpha-\beta)(3\alpha+3\beta+4))$$
$$C(x)=-(\alpha+\beta)((\alpha+\beta)(\alpha+\beta+2)(2n(n+\alpha+\beta+1)+\alpha+\beta-1)x^2$$
$$+2(n(n+\alpha+\beta+1)(-\mu(\alpha+\beta+1)(\alpha+\beta-4)+2(\alpha-\beta)(\alpha+\beta-2))$$
$$+(\alpha+\beta)(\alpha-\beta)(\alpha+\beta-1))x-2n(n+\alpha+\beta+1)(-\mu(\alpha+\beta+1)(\beta-\alpha)$$
$$-(\alpha-\beta)^2+6\alpha+6\beta)+(\alpha+\beta)((\alpha-\beta)^2-3\alpha-3\beta-6)).$$

3.5.3. Limit $c=-\alpha-\beta$. Due to the T' invariance of (3.2) we obtain in this limit the special case of CAJ polynomials for which

$$R_n^{\alpha,\beta}(x;-\alpha-\beta,\mu)=R_n^{-\alpha,-\beta}(x;\mu). \tag{3.37}$$

All the results are obtained from §3.5.2 by changing α to $-\alpha$ and β to $-\beta$.

3.5.4. Limit $c=-\beta$. The explicit form (3.19) simplifies in the same way as in the case $c=0$. One obtains

$$R_n^{\alpha,\beta}(x;-\beta,\mu)=(-1)^n\frac{(\alpha-\beta+1)_n}{(\alpha+1)_n}\sum_{k=0}^{n}\frac{(-n)_k(n+\alpha-\beta+1)_k}{k!(1-\beta)_k}x^k$$
$$\times\left\{1+\frac{\mu(k-n)(n+k+\alpha-\beta+1)}{2(k+1)(1-\beta+k)}\right.$$
$$\left.\times{}_4F_3\left(\begin{matrix}k-n+1,n+k+\alpha-\beta+2,1-\beta,1\\k+2,2-\beta+k,\alpha-\beta+2\end{matrix};1\right)\right\}. \tag{3.38}$$

Comparing this form with the explicit form of the co-recursive Jacobi polynomials (3.34) one sees that

$$R_n^{\alpha,\beta}(x;-\beta,\mu)=\frac{n!(\alpha-\beta+1)_n}{(\alpha+1)_n(1-\beta)_n}R_n^{\alpha,-\beta}(x;\mu). \tag{3.39}$$

Of course, the spectral measure and the fourth-order differential equation are obtained from (3.35) and (3.36), changing β to $-\beta$.

3.5.5. Limit $c=-\alpha$. This case is the T' transform of the preceding case. All the results are obtained from §3.5.4 by changing α to $-\alpha$ and β to $-\beta$.

3.5.6. Limit $\mu=0$. In this limit we obtain the associated Jacobi polynomials studied in [31]. The form [31, eq. 28] is obtain directly using (3.18) but a slightly different form is

$$R_n^{\alpha,\beta}(x;c)=(1+T')\frac{(c+\alpha)(c+\alpha+\beta)}{\alpha(2c+\alpha+\beta)}\frac{(c+\alpha+1)_n}{(c+1)_n}$$
$$\times{}_2F_1\left(\begin{matrix}1-c-\alpha-\beta,c\\1-\alpha\end{matrix};1-x\right){}_2F_1\left(\begin{matrix}-n-c,n+c+\alpha+\beta+1\\1+\alpha\end{matrix};1-x\right). \tag{3.40}$$

The explicit form [31, eq. 19] is easily obtained starting from (3.19) with $G=2c+\alpha+\beta$, the ${}_5F_4$ reducing to a ${}_4F_3$. Obviously the limit $c=0$ leads back to the Jacobi polynomials.

The coefficients of the differential equation (2.38) satisfied by the associated Jacobi polynomials are

$$c_4=x^2(x-1)^2,\qquad c_3=5x(x-1)(2x-1), \tag{3.41a}$$

$$c_2 = \left(24 - (n+1)^2 - A\right) x(x-1) - Bx - \beta^2 + 4, \tag{3.41b}$$

$$c_1 = -3/2\left((3A + (n+3)(n-1))(2x-1) + B\right), \tag{3.41c}$$

$$c_0 = n(n+2)A, \tag{3.41d}$$

with

$$A = (C+n+1)(C+n-1), \quad B = (\alpha-\beta)(\alpha+\beta), \quad C = 2c+\alpha+\beta. \tag{3.42}$$

This result was first given by Hahn [9, eq. 20]. Note the T' invariance of A, B, and C leading to the invariance of the c_i, more obvious than in [31, eq. 47–48].

3.5.7. Limit $\mu = 2c(c+\alpha)/(2c+\beta+\alpha)(2c+\beta+\alpha+1)$. In this limit the symmetry T' is broken. We obtain the zero-related Jacobi polynomials studied in [15]. An explicit form is

(3.43)

$$\mathcal{R}_n^{\alpha,\beta}(x;c) = (-1)^n \frac{(2c+\alpha+\beta+1)_n(\beta+c+1)_n}{n!(c+\alpha+\beta+1)_n} \sum_{k=0}^{n} \frac{(-n)_k(n+2c+\alpha+\beta+1)_k}{(c+1)_k(c+\beta+1)_k} \times {}_4F_3\left(\begin{matrix} k-n, n+k+2c+\alpha+\beta+1, c, c+\beta+1 \\ c+k+1, c+\beta+k+1, 2c+\alpha+\beta+1 \end{matrix}; 1\right) x^k.$$

The limit $c = 0$ leads back to the Jacobi polynomials and the limit defined in (3.32) gives the zero-related Laguerre polynomials (2.4.4), using Kummer's transformations. The spectral measure is

$$\phi'(x) = (1-x)^\alpha x^{\beta+2c} \left| {}_2F_1\left(\begin{matrix} c, c+\beta+1 \\ 2c+\alpha+\beta+1 \end{matrix}; \frac{e^{i\pi}}{x}\right)\right|^{-2}. \tag{3.44}$$

The coefficients of the differential equation (2.38) satisfied by the polynomials $\mathcal{R}_n^{\alpha,\beta}(x;c)$ are

$$c_4 = x^2(x-1)^2(Ax+D), \tag{3.45a}$$

$$c_3 = x(x-1)\left(8Ax^2 - 3(A-3D)x - 4D\right), \tag{3.45b}$$

$$c_2 = -1/2A(A+2C^2-29)x^3 + \left(1/2A(A+2C^2-2B-23) - D(C^2-19)\right)x^2 - 1/4\left(A(D+1)(D-3) - 2D(2C^2-2B+D-35)\right)x - 1/4D(D^2-9), \tag{3.45c}$$

$$c_1 = -A(A+2C^2-5)x^2 + 1/4\left(A(A+2C^2-2B-5D-5) - 3D(4C^2-11)\right)x + 1/4D\left((D+3)A + 6C^2 - 6B + 3D - 15\right), \tag{3.45d}$$

$$c_0 = 2n(n+1)(C+n)(C+n+1)(Ax+3D), \tag{3.45e}$$

where B and C are defined in (3.42) and

$$A = (2n+1)(1+2C+2n), \qquad D = 1+2\beta. \tag{3.46}$$

3.5.8. Limit $\mu = 2(c+\beta)(c+\alpha+\beta)/(\beta+\alpha+2c)(\beta+\alpha+2c+1)$. This case is the T' transform of the case in §3.5.7. The explicit form is

(3.47)

$$\mathfrak{R}_n^{\alpha,\beta}(x;c) = (-1)^n \frac{(2c+\alpha+\beta+1)_n(\alpha+c+1)_n}{n!(c+1)_n} \sum_{k=0}^{n} \frac{(-n)_k(n+2c+\alpha+\beta+1)_k}{(c+\alpha+\beta+1)_k(c+\alpha+1)_k} \times {}_4F_3\left(\begin{matrix} k-n, n+k+2c+\alpha+\beta+1, c+\alpha+\beta, c+\alpha+1 \\ c+\alpha+\beta+k+1, c+\alpha+k+1, 2c+\alpha+\beta+1 \end{matrix}; 1\right) x^k.$$

The limit (3.32) leads back to the Laguerre case in §2.4.5 and the limit $c = 0$ to the co-recursive Jacobi polynomials with $\mu = 2\beta/(\alpha+\beta+1)$. The spectral measure is

$$\phi'(x) = (1-x)^{\alpha} x^{\beta+2c} \left| {}_2F_1\left(\begin{matrix} c+1, c+\beta \\ 2c+\alpha+\beta+1 \end{matrix}; \frac{e^{i\pi}}{x}\right)\right|^{-2}. \tag{3.48}$$

The coefficients of the differential equation (2.38) satisfied by the polynomials $\mathfrak{R}_n^{\alpha,\beta}(x;c)$ are obtained from (3.45), (3.46), changing only $D = 1+2\beta$ by $D = 1-2\beta$.

3.5.9. Limit $\mu = -2c(c+\beta)/(2c+\alpha+\beta)(2c+\alpha+\beta+1)$. The symmetry $\mathcal{T}'$ is also broken. We obtain a new simple case of CAJ polynomials. An explicit form is

(3.49)

$$\widetilde{\mathcal{R}}_n^{\alpha,\beta}(x;c) = (-1)^n \frac{(2c+\alpha+\beta+1)_n (\beta+c+1)_n}{n!(c+\alpha+\beta+1)_n} \sum_{k=0}^{n} \frac{(-n)_k (n+2c+\alpha+\beta+1)_k}{(c+1)_k (c+\beta+1)_k} \times {}_4F_3\left(\begin{matrix} k-n, n+k+2c+\alpha+\beta+1, c, c+\beta \\ c+k+1, c+\beta+k+1, 2c+\alpha+\beta+1 \end{matrix}; 1\right) x^k$$

and the spectral measure is

$$\phi'(x) = (1-x)^{\alpha} x^{\beta+2c} \left| {}_2F_1\left(\begin{matrix} c, c+\beta \\ 2c+\alpha+\beta+1 \end{matrix}; \frac{e^{i\pi}}{x}\right)\right|^{-2}. \tag{3.50}$$

The coefficients of the differential equation (2.38) satisfied by the polynomials $\widetilde{\mathcal{R}}_n^{\alpha,\beta}(x;c)$ are

$$c_4 = x^2(x-1)^2\left(A(x-1) - D\right), \tag{3.51a}$$

$$c_3 = x(x-1)\left(8Ax^2 - (13A+9D)x + 5A + 5D\right), \tag{3.51b}$$

$$\begin{aligned} c_2 = {} & -1/2A(A+2C^2-29)x^3 + \left(A(A+2C^2-B-32) + D(C^2-19)\right)x^2 \\ & -1/2\left(A(A+2C^2+2\beta^2-2B-43) + D(2C^2-2B-D-41)\right)x \\ & +(\beta^2-4)(A+D), \end{aligned} \tag{3.51c}$$

$$\begin{aligned} c_1 = {} & -2A(A+2C^2-5)x^2 \\ & +1/2\left(A(7A+14C^2-2B+5D-35) + D(3C^2-33)\right)x \\ & -A(3/2A+3C^2-B-2\alpha^2-7) - 3D(C^2-B-1/2D-3), \end{aligned} \tag{3.51d}$$

$$c_0 = n(n+1)(C+n)(C+n+1)(A(x-1) - 3D), \tag{3.51e}$$

where B and C are defined in (3.42) and

$$A = (2n+1)(1+2C+2n), \qquad D = 1+2\alpha. \tag{3.52}$$

3.5.10. Limit $\mu = -2(c+\alpha)(c+\alpha+\beta)/(2c+\alpha+\beta)(2c+\alpha+\beta+1)$. This case is the $\mathcal{T}'$ transform of the case in §3.5.9. The explicit form is

(3.53)

$$\widetilde{\mathfrak{R}}_n^{\alpha,\beta}(x;c) = (-1)^n \frac{(2c+\alpha+\beta+1)_n (\alpha+c+1)_n}{n!(c+1)_n} \sum_{k=0}^{n} \frac{(-n)_k (n+2c+\alpha+\beta+1)_k}{(c+\alpha+\beta+1)_k (c+\alpha+1)_k} \times {}_4F_3\left(\begin{matrix} k-n, n+k+2c+\alpha+\beta+1, c+\alpha+\beta, c+\alpha \\ c+\alpha+\beta+k+1, c+\alpha+k+1, 2c+\alpha+\beta+1 \end{matrix}; 1\right) x^k$$

and the spectral measure

$$\phi'(x) = (1-x)^{\alpha-2}x^{\beta+2c+2}\left|{}_2F_1\left(\begin{matrix}c+1, c+\beta+1\\ 2c+\alpha+\beta+1\end{matrix}; \frac{e^{i\pi}}{x}\right)\right|^{-2}. \tag{3.54}$$

The coefficients of the differential equation (2.38) satisfied by the polynomials $\widetilde{\mathfrak{R}}_n^{\alpha,\beta}(x;c)$ are obtained from (3.51), (3.52) by changing only $D = 1+2\alpha$ by $D = 1-2\alpha$.

3.6. Conclusion. We end with brief remarks. In this article we have studied properties of the co-recursive associated Laguerre and Jacobi polynomials that are of interest in the resolution of some birth and death processes with and without absorption. For a few values of the co-recursivity parameter we obtain polynomial families for which the results are of the same complexity as the corresponding associated polynomials. For the CAL polynomials we find the two expected cases corresponding to $\mu = \mu_0$ (zero-related polynomials) and the new dual case $\mu = \lambda_{-1}$ [20]. For the CAJ polynomials, due to the properties (3.3) and (3.5), we have two more cases corresponding to $\mu = -\mathcal{T}''\mu_0$ and $\mu = -\mathcal{T}''\lambda_{-1}$, where the transformation $\mathcal{T}''$ is defined by

$$\mathcal{T}''(c,\alpha,\beta) = (c+\alpha+\beta, -\beta, -\alpha). \tag{3.55}$$

In some cases the fourth-order differential equations satisfied by the polynomials studied above are factorizable (co-recursive and associated of order one), but we do not find factorization either for the co-recursive associated polynomials or for the associated one. Of course, this is not a proof that the conjectures on this factorizability made in [27] are wrong.

Acknowledgments. The author thanks Galliano Valent and Pascal Maroni for stimulating discussions during the preparation of this article.

REFERENCES

[1] G. E. Andrews and R. Askey, *Classical orthogonal polynomials*, in Polynômes Orthogonaux et Applications, C. Brezinski, A. Draux, A. Magnus, P. Maroni, and A. Ronveaux, eds., Vol. 1171, Springer-Verlag, Berlin, 1985, pp. 36–62.

[2] R. Askey and J. Wimp, *Associated Laguerre and Hermite polynomials*, Proc. Royal Soc. Edinburgh, 96A (1984), pp. 15–37.

[3] F. V. Atkinson and W. N. Everitt, *Orthogonal polynomials which satisfy second-order differential equations*, in Christoffel Festscrift, P. Butzer and F. Feher, eds., Birkhaüser-Verlag, Basel, 1981.

[4] W. N. Bailey, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, reprinted by Hafner, New York, 1972.

[5] S. Belmehdi and A. Ronveaux, *The fourth-order differential equation satisfied by the associated orthogonal polynomials*, Rendi. Mat. Roma (Serie VII), 11 (1991), pp. 313–326.

[6] B. W. Char, K. O. Geddes, G. H. Gonnet, M. B. Monagan, and S. M. Watt, MAPLE *reference manual*, WATCOM, University of Waterloo, Canada, 1988.

[7] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions*, vol. I, McGraw-Hill, New York, 1953.

[8] ———, *Higher Transcendental Functions*, vol. II, McGraw-Hill, New York, 1953.

[9] W. Hahn, *Über Orthogonalpolynome mit drei Parametern*, Vol. 5, Deutche. Math., Berlin, 1940–41.

[10] M. E. Ismail and D. H. Kelker, *Special functions, Stieltjes transforms and infinite divisibility*, SIAM J. Math. Anal., 10 (1979), pp. 884–901.

[11] M. E. Ismail, J. Letessier, D. Masson, and G. Valent, *Birth and death processes and orthogonal polynomials*, in Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., vol. 294, NATO ASI series C, Kluwer, Dordrecht, Boston, London, 1990, pp. 229–255.

[12] M. E. ISMAIL, J. LETESSIER, AND G. VALENT, *Linear birth and death models and associated Laguerre polynomials*, J. Approx. Theory, 56 (1988), pp. 337–348.

[13] ———, *Quadratic birth and death processes and associated continuous dual Hahn polynomials*, SIAM J. Math. Anal., 20 (1989), pp. 727–737.

[14] M. E. ISMAIL, J. LETESSIER, G. VALENT, AND J. WIMP, *Two families of associated Wilson polynomials*, Canad. J. Math., 42 (1990), pp. 659–695.

[15] M. E. ISMAIL AND D. R. MASSON, *Two families of orthogonal polynomials related to Jacobi polynomials*, Rocky Mountain J. Math., 21 (1991), pp. 884–901.

[16] S. KARLIN AND J. MCGREGOR, *The differential equations of birth and death processes and the Stieltjes moment problem*, Trans. Amer. Math. Soc., 85 (1958), pp. 489–546.

[17] S. KARLIN AND S. TAVARÉ, *A diffusion process with killing: the time to formation of recurrent deleterious mutant genes*, Stochastic Process. Appl., 13 (1982), pp. 249–261.

[18] ———, *Linear birth and death processes with killing*, J. Appl. Prob., 19 (1982), pp. 477–487.

[19] J. LABELLE, *Tableau d'Askey*, in Polynômes Orthogonaux et Applications, C. Brezinski, A. Draux, A. Magnus, P. Maroni, and A. Ronveaux, eds., Vol. 1171, Springer-Verlag, Berlin, 1985, p. 36.

[20] J. LETESSIER AND G. VALENT, *Dual birth and death processes and orthogonal polynomials*, SIAM J. Appl. Math., 46 (1986), pp. 393–405.

[21] Y. L. LUKE, *The special functions and their approximations*, Vol. I, Academic Press, New York, 1969.

[22] P. MARONI, *Une théorie algébrique des polynômes orthogonaux. Application aux polynômes orthogonaux semi-classique*, in Orthogonal Polynomials and Their Applications, C. Brezinski, L. Gori, and A. Ronveaux, eds., IMACS Vol. 9, 1991, pp. 95–130.

[23] W. M. ORR, *On the product $J_m(x)J_n(x)$*, Proc. Cambridge Philos. Soc., 10 (1900), pp. 93–100.

[24] G. E. REUTER, *Denumerable Markov processes and associated semigroups on l*, Acta Math., 97 (1957), pp. 1–46.

[25] A. RONVEAUX, *Fourth-order differential equations for numerator polynomials*, J. Phys. A: Math. Gen., 21 (1988), pp. L749–L753.

[26] ———, *4th-order differential equations and orthogonal polynomials of the Laguerre–Hahn class*, in Orthogonal Polynomials and Their Applications, C. Brezinski, L. Gori, and A. Ronveaux, eds., IMACS Vol. 9, 1991, pp. 379–385.

[27] A. RONVEAUX AND F. MARCELLAN, *Co-recursive orthogonal polynomials and fourth-order differential equation*, J. Comput Appl. Math., 25 (1989), pp. 105–109.

[28] J. A. SHOHAT AND J. D. TAMARKIN, *The problem of moments*, in Mathematical Surveys, Vol. 1, American Mathematical Society, Providence, RI, 1950.

[29] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, UK, 1966.

[30] G. N. WATSON, *Theory of Bessel functions*, Cambridge University Press, Cambridge, UK, 1944.

[31] J. WIMP, *Explicit formulas for the associated Jacobi polynomials and some applications*, Canad. J. Math., 39 (1987), pp. 983–1000.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 549–554, March 1994

020

BOUNDS AND MONOTONICITIES FOR THE ZEROS OF DERIVATIVES OF ULTRASPHERICAL BESSEL FUNCTIONS*

LEE LORCH† AND PETER SZEGO‡

Abstract. The positive zeros $p_{\nu k}^{(\ell)}$ of $[x^{-\nu+1}J_{\nu+\ell-1}(x)]'$, $\nu+\ell>0$, where $J_\nu(x)$ denotes the standard Bessel function, arise in the study of the eigenvalues of Neumann Laplacians in N dimensions [M. S. Ashbaugh and R. D. Benguria, *SIAM J. Math. Anal.*, 24 (1993), pp. 557–570]. The case $\ell=1$ is particularly relevant. To pave the way for these applications, the authors present here inter alia (i) lower and upper bounds for $p_{\nu 1}^{(\ell)}$ and (ii) an explicit representation for $dp_{\nu k}^{(\ell)}/d\nu$. The latter implies that $p_{\nu k}^{(\ell)}$ is increasing in ν for fixed k,ℓ, provided $\nu+\ell>1$.

Key words. eigenvalues, Bessel functions, zeros, monotonicity

AMS subject classification. 33C40

1. Background and statement of results. In the course of their study [1] of eigenvalues of Neumann Laplacians in higher dimensions, Ashbaugh and Benguria require information concerning the zeros of $[x^{-\nu+1}J_{\nu+\ell-1}(x)]'$, where $J_\nu(x)$ is the usual Bessel function of the first kind and order ν; ℓ is independent of ν. For convenience of reference, they propose calling $x^{-\nu+1}J_{\nu+\ell-1}(x)$ an "ultraspherical Bessel function," since this function becomes (apart from the multiplicative factor $\sqrt{2/\pi}$) the standard spherical Bessel function when $\nu=\frac{3}{2}$ and $\ell=0,1,2,3,\ldots$. We adopt their nomenclature here [1].

The purpose of this note is to establish results that Ashbaugh and Benguria have found relevant to their work. They have particular use for the case $\ell=1$, $\nu=\frac{1}{2}n$, $n=1,2,3,\ldots$, and so we provide explicit statements below for this case.

By $p_{\nu k}^{(\ell)}$ we mean the kth positive zero of $[x^{-\nu+1}J_{\nu+\ell-1}(x)]'$; $p_{\nu k}^{(1)}$ will be simply denoted by $p_{\nu k}$. Furthermore, we note that $p_{\nu k}^{(0)}=j_{\nu,k}$, where $j_{\nu,k}$ is the kth positive zero of $J_\nu(x)$, since [3, §3.2, p. 45] $[x^{-\nu}J_\nu(x)]'=-J_{\nu+1}(x)/x^\nu$.

We shall establish the following properties:

$$(1)\qquad \frac{2\ell(\nu+\ell)(\nu+\ell+1)}{\nu+2\ell+1}<\left[p_{\nu 1}^{(\ell)}\right]^2<2\ell(\nu+\ell),\qquad \ell>0,\ \nu>-1,\ \nu+\ell>0;$$

in particular,

$$(1')\qquad 2\nu+\frac{4}{\nu+3}<p_{\nu 1}^2<2\nu+2,\qquad \nu>-1.$$

We need (2) and (3) as an alternative lower bound for $[p_{\nu 1}^{(\ell)}]^2$. It is weaker than the lower bound provided in (1) for the case ($\ell=1$) required by Ashbaugh and Benguria, but stronger when $\ell>2$ and ν is sufficiently large.

$$(1'')\qquad \left[p_{\nu 1}^{(\ell)}\right]^2>2\nu\ell+\ell^2-2\ell,\quad \nu+\ell>0,\quad \ell>0.$$

* Received by the editors May 28, 1992; accepted for publication June 17, 1993. This work received partial support from the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics and Statistics, York University, North York, Ontario, Canada M3J 1P3.

‡ 75 Glen Eyrie Avenue, Apt. 19, San Jose, California 95125.

This inequality will show that the denominators in (2) and a fortiori in (2′) are positive.

With $p^{(\ell)} = p_{\nu k}^{(\ell)}$, $\nu + \ell > 1$, we have

$$\frac{dp^{(\ell)}}{d\nu} = \frac{p^{(\ell)}}{\left[p^{(\ell)}\right]^2 - ((2\nu-2)\ell + \ell^2)} \frac{1}{J_{\nu+\ell-1}^2(p^{(\ell)})} \cdot \left\{ 2(\nu - 1 + \ell) \int_0^{p^{(\ell)}} \frac{J_{\nu-1+\ell}^2(t)}{t} dt - J_{\nu-1+\ell}^2(p^{(\ell)}) \right\}, \tag{2}$$

when, with $p = p_{\nu k}$, $\nu > 0$,

$$\frac{dp}{d\nu} = \frac{p}{p^2 - (2\nu - 1)} \frac{1}{J_\nu^2(p)} \left\{ 2\nu \int_0^p \frac{J_\nu^2(t)}{t} dt - J_\nu^2(p) \right\}. \tag{2′}$$

From (1″) and (2) it follows, for each $k = 1, 2, \ldots$, that

$$\frac{dp_{\nu k}^{(\ell)}}{d\nu} > 0, \qquad \nu + \ell > 1. \tag{3}$$

From (1′) and (2′) it follows, for each $k = 1, 2, \ldots$, that

$$\frac{dp_{\nu k}}{d\nu} > 0, \qquad \nu > 0. \tag{3′}$$

As a consequence of (3), which establishes that $p_{\nu k}^{(\ell)}$ is an increasing function of ν for each fixed $k = 1, 2, 3, \ldots$, Ashbaugh and Benguria [1] can infer that each Neumann eigenvalue of the unit ball in n dimensions ($n \geq 2$) increases with the dimension.

2. A preliminary result. Common to the proofs of (1), (1′), (2), and (2′) will be the equation

$$p_{\nu k}^{(\ell)} J'_{\nu+\ell-1}\left(p_{\nu k}^{(\ell)}\right) = (\nu - 1) J_{\nu+\ell-1}\left(p_{\nu k}^{(\ell)}\right) \tag{4}$$

and the special case $\ell = 1$

$$p_{\nu k} J'_\nu(p_{\nu k}) = (\nu - 1) J_\nu(p_{\nu k}). \tag{4′}$$

This is an immediate consequence of the differentiation formula

$$x^\nu \frac{d}{dx} \left\{ \frac{J_{\nu+\ell-1}(x)}{x^{\nu-1}} \right\} = x J'_{\nu+\ell-1}(x) - (\nu - 1) J_{\nu+\ell-1}(x)$$

and the definition of $p_{\nu k}^{(\ell)}$.

3. Proof of (1) and (1′). First we establish the lower bounds in (1) and (1′). To do so, we consider (i) $J'_{\nu+\ell+1}(p_{\nu 1}^{(\ell)}) > 0$, a case which occurs whenever $\nu > 1$, as is clear from (4), and (ii) $J'_{\nu+\ell+1}(p_{\nu 1}^{(\ell)}) \leq 0$.

To case (i) we apply the unnumbered formula following [3, §15.3 (2), p. 486] with $x = p_{\nu 1}^{(\ell)}$. Taking (4) into account, this becomes

$$0 < p_{\nu 1}^{(\ell)} J'_{\nu+\ell+1}\left(p_{\nu 1}^{(\ell)}\right) = 2(\nu+\ell)\left\{1 - \frac{(\nu+\ell-1)(\nu+\ell+1)}{\left[p_{\nu 1}^{(\ell)}\right]^2}\right\} J_{\nu+\ell-1}\left(p_{\nu 1}^{(\ell)}\right)$$
$$- \left\{1 - \frac{2(\nu+\ell)(\nu+\ell+1)}{\left[p_{\nu 1}^{(\ell)}\right]^2}\right\} (\nu-1) J_{\nu+\ell-1}\left(p_{\nu 1}^{(\ell)}\right).$$

Recalling that $p_{\nu 1}^{(\ell)} < j_{\nu+\ell-1,1}$, so that $J_{\nu+\ell-1}\left(p_{\nu 1}^{(\ell)}\right) > 0$, this verifies the first inequality in (1) in case (i).

In case (ii), we have $p_{\nu 1}^{(\ell)} \geq j'_{\nu+\ell+1,1}$. But [3, §15.3 (3), p. 486] $(j'_{\nu 1})^2 > \nu^2 + 2\nu$, so that

$$\left[p_{\nu 1}^{(\ell)}\right]^2 > (\nu+\ell+1)(\nu+\ell+3),$$

which exceeds the lower bound stated in (1). This completes case (ii).

The upper bound in (1) can be inferred from the unnumbered formula immediately above [3, §15.3 (4), p. 486] with ν replaced by $\nu+\ell-1$, x by $p_{\nu 1}^{(\ell)}$, and then using (4), this gives

$$\frac{J_{\nu+\ell+1}\left(p_{\nu 1}^{(\ell)}\right)}{J_{\nu+\ell-1}\left(p_{\nu 1}^{(\ell)}\right)} = -1 + \frac{2(\nu+\ell-1)(\nu+\ell)}{\left[p_{\nu 1}^{(\ell)}\right]^2} - \frac{2(\nu-1)(\nu+\ell)}{\left[p_{\nu 1}^{(\ell)}\right]^2}.$$

The left member is positive, since $p_{\nu 1}^{(\ell)} < j_{\nu+\ell-1,1}$, when $\nu+\ell-1 > -1$, i.e., when $\nu+\ell > 0$.

The upper bounds in (1) and (1′) follow immediately.

Remark 1. From the upper bound in (1) we find that

$$\lim_{\ell \to 0} p_{\nu 1}^{(\ell)} = 0 \neq j_{\nu 1} = p_{\nu 1}^{(0)}, \qquad \nu > -1.$$

Thus, $p_{\nu 1}^{(\ell)}$ is not continuous at $\ell = 0$ for any fixed $\nu > -1$. From (1′) we see that $p_{\nu 1} \to 0$ as $\nu \to -1^+$, and, more generally, from (1) we see that $p_{\nu 1}^{(\ell)} \to 0$ as $\nu \to -\ell^+$.

Remark 2. Equation (4) implies that $J_{\nu+\ell-1}(p_{\nu k}^{(\ell)}) \neq 0$, $k = 1, 2, \ldots$, since the Bessel differential equation has a finite singularity only at $x = 0$, so that $J_{\nu+\ell-1}(x)$ and $J'_{\nu+\ell-1}(x)$ can vanish simultaneously only at $x = 0$.

4. Proof of (1″). The lower bound given by (1″) will be obtained from the differential equation, satisfied by $y = J_{\nu+\ell-1}(x)/x^{\nu-1}$:

$$x^2 y'' + (2\nu - 1) x y' + (x^2 - \ell^2 - 2\nu\ell + 2\ell) y = 0. \tag{5}$$

This is the form to which [3, §4.31(19), p. 98] specializes when in the latter $\psi(x)$ is taken to be x, μ to be $1-\nu$, and the ν occurring in [3] is replaced by $\nu+\ell-1$.

In the course of the proof, we shall use a simple and presumably already recorded property of a class of differential equations to which (5) belongs. We state and prove it here for completeness.

LEMMA. *Let $x = \xi$ yield a positive maximum of the function $y(x)$ that satisfies the differential equation*

$$a_0(x)y'' + a_1(x)y' + a_2(x)y = 0,$$

where $a_0(x), a_1(x), a_2(x)$ are differentiable functions such that $a_0(\xi) > 0$ and $a_2'(\xi) > 0$. Then $y''(\xi) < 0$.

Proof. Clearly, $y''(\xi) \leq 0$. Suppose, *par impossible*, that $y''(\xi) = 0$. Differentiating the given differential equation and then putting $x = \xi$ would give rise to the equation

$$a_0(\xi)y'''(\xi) + a_2'(\xi)y(\xi) = 0.$$

This equation shows, in the light of the lemma's hypotheses, that $y'''(\xi) < 0$. Thus, there exists $\delta > 0$ such that $y''(x)$ decreases in $\xi - \delta < x \leq \xi$ to 0. Hence $y''(x) > 0$, $\xi - \delta < x < \xi$, so that $y'(x)$ increases to 0, $\xi - \delta < x \leq \xi$, i.e., $y'(x) < 0$, $\xi - \delta < x < \xi$. But this implies that $y(x)$ *decreases* to its maximum $y(\xi)$, a manifest impossibility. This proves the lemma.

Now we revert to the special case (5) and note that its solution $y(x) = J_{\nu+\ell-1}(x)/x^{\nu-1} > 0$, $0 < x < j_{\nu+\ell-1,1}$, since (as assumed for (1″)) $\nu + \ell > 0$. Thus, $p_{\nu 1}^{(\ell)}$, the first positive zero of $y'(x)$, yields a positive maximum. From the lemma, $y''(p_{\nu 1}^{(\ell)}) < 0$. Putting $x = p_{\nu 1}^{(\ell)}$ in (5) concludes the proof of (1″), since $y(p_{\nu 1}^{(\ell)}) > 0$.

Remark 3. The lemma, which would otherwise be unnecessary, establishes *strict* inequality in (1″). This is required to permit the division in (2) and (2′) and the positivity asserted in (3) and (3′).

Remark 4. There are equations of type (5) possessing nontrivial solutions y having positive maxima $y(\xi)$ for which $y''(\xi) = 0$. With twice differentiable coefficients $a_0(x), a_1(x), a_2(x)$ such that $a_0(\xi) > 0$ as in the lemma, but with $a_2(\xi) = a_2'(\xi) = 0$ and $a_2''(\xi) > 0$, it follows that $y''(\xi) = y'''(\xi) = 0$ and $y^{(4)}(\xi) < 0$. Thus, it suffices to select the solution for which $y(\xi) = 1$, $y'(\xi) = 0$.

One such example is a special case of a transformation, due to Lommel [see 3, §4.31(9), p. 97], of Bessel's differential equation. This case gives the equation $y'' + 4x^2y = 0$. The initial conditions $y(0) = 2^{1/4}/\Gamma(\frac{3}{4})$, $y'(0) = 0$ yield the unique solution $y = |x|^{1/2}J_{-1/4}(x^2)$. Here $y''(0) = y'''(0) = 0$, $y^{(4)}(0) = -8 < 0$, so that $x = 0$ yields a positive maximum at $x = 0$ but with $y''(0) = 0$.

5. Proof of formulae (2) and (2′). Formula (2) was put in the form stated to facilitate the specialization into (2′) for which applications are at hand [1]. To simplify the notation employed in the proof of (2), we put

$$\mu = \nu - 1, \qquad q = p_{\nu k}^{(\ell)}$$

so that (2) can be stated, since $q^2 \neq 2\mu\ell + \ell^2$, $\mu + \ell > 0$, as

$$(2'') \qquad \frac{dq}{d\mu} = \frac{q}{q^2 - (2\mu\ell + \ell^2)} \frac{1}{J_{\mu+\ell}^2(q)} \left\{ 2(\mu + \ell) \int_0^q \frac{J_{\mu+\ell}^2(t)}{t}\, dt - J_{\mu+\ell}^2(q) \right\}.$$

Thus, (4) becomes

$$(4'') \qquad -\mu J_{\mu+\ell}(q) + qJ_{\mu+\ell}'(q) = 0.$$

Differentiating this with respect to μ yields

$$-J_{\mu+\ell}(q) - \mu \frac{\partial J_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q} + q \frac{\partial J'_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q} + (1-\mu)J'_{\mu+\ell}(q)\frac{dq}{d\mu} + q\frac{dq}{d\mu}J''_{\mu+\ell}(q) = 0,$$

or, collecting terms,

$$\frac{dq}{d\mu}\left\{(1-\mu)J'_{\mu+\ell}(q) + qJ''_{\mu+\ell}(q)\right\} = J_{\mu+\ell}(q) + \mu \frac{\partial J_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q} - q\frac{\partial J'_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q}. \tag{6}$$

In the left member of (6), we replace $J''_{\mu+\ell}(q)$ in terms of $J_{\mu+\ell}(q)$ and $J'_{\mu+\ell}(q)$ via Bessel's differential equation and then with $J'_{\mu+\ell}(q)$ in terms of $J_{\mu+\ell}(q)$ by means of (4″). Thus,

$$\frac{dq}{d\mu}\left\{(1-\mu)J'_{\mu+\ell}(q) + qJ''_{\mu+\ell}(q)\right\} = \frac{dq}{d\mu}\left\{\frac{2\mu\ell + \ell^2 - q^2}{q}J_{\mu+\ell}(q)\right\}. \tag{7}$$

The right member of (6) can be transformed via the identity [2, p. 247]

$$\frac{x}{2(\mu+\ell)}\left\{J_{\mu+\ell}(x)\frac{\partial J'_{\mu+\ell}(x)}{\partial \mu} - J'_{\mu+\ell}(x)\frac{\partial J_{\mu+\ell}(x)}{\partial \mu}\right\} = \int \frac{J^2_{\mu+\ell}(x)}{x}dx,$$

which we may evaluate between the limits of 0 and q. The lower limit on the left vanishes since $\mu + \ell > 0$. Hence,

$$\frac{q}{2(\mu+\ell)}\left\{J_{\mu+\ell}(q)\frac{\partial J'_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q} - J'_{\mu+\ell}(q)\frac{\partial J_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q}\right\} = \int_0^q \frac{J^2_{\mu+\ell}(x)}{x}\,dx.$$

In this equation, the factor $J'_{\mu+\ell}(q)$ can be expressed in terms of $J_{\mu+\ell}(q)$ by (4″). Hence, since $J_{\mu+\ell}(q) \neq 0$,

$$\mu \frac{\partial J_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q} - q\frac{\partial J'_{\mu+\ell}(x)}{\partial \mu}\bigg|_{x=q} = -\frac{2(\mu+\ell)}{J_{\mu+\ell}(q)}\int_0^q \frac{J^2_{\mu+\ell}(x)}{x}\,dx.$$

This evaluation, together with (6) and (7), establishes (2″) and also (2) and (2′).

6. Proof of (3) and (3′). Now that (1″), (2), and (2′) have been demonstrated, the signs in (3) and (3′) readily follow. The transition is accomplished with the use of the identity [3, §5.51 (5), p. 152], valid for $\nu + \ell - 1 > 0$,

$$2(\nu - 1 + \ell)\int_0^{p^{(\ell)}} \frac{J^2_{\nu+\ell-1}}{x}dx - J^2_{\nu+\ell-1}(p^{(\ell)}) = 2\sum_{n=1}^{\infty} J^2_{\nu+\ell-1+n}(p^{(\ell)}).$$

This is positive, so that the sign of $dp^{(\ell)}/d\nu$ is, according to (2), the same as the sign of $[p^{(\ell)}]^2 - ((2\nu-2)\ell + \ell^2)$, when $\nu + \ell > 1$, i.e., positive, from (1″).

Acknowledgment. We are grateful to a conscientious referee whose careful attention to detail corrected the erroneous value we had ascribed to the coefficient in (2), along with related slips. In addition, we thank M. S. Ashbaugh and R. D. Benguria for a preprint of [1].

REFERENCES

[1] MARK S. ASHBAUGH AND RAFAEL D. BENGURIA, *Universal bounds for the low eigenvalues of Neumann Laplacians in N dimensions*, SIAM J. Math. Anal., 24 (1993), pp. 557–570.
[2] FRANK W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[3] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 555–570, March 1994

021

BEST WEIGHTED POLYNOMIAL APPROXIMATION VIA JACOBI EXPANSIONS*

DORON S. LUBINSKY† AND VILMOS TOTIK‡

Abstract. Freud's method is modified to prove the Jackson–Favard estimate for weighted algebraic polynomials with Jacobi weights. This approach uses only classical results from the theory of orthogonal polynomials.

Key words. weighted polynomial approximation, orthogonal expansion, Jacobi polynomials

AMS subject classifications. 41A10, 42C10

1. Introduction.

Let f be a function on $[-1,1]$, and

$$E_n(f)_p = \inf_{\deg P_n \le n} \|f - P_n\|_{L^p[-1,1]}$$

be its best approximation by polynomials of degree at most n in the L^p metric. One of the central questions in the constructive theory of approximation is to characterize $E_n(f)_p$ in terms of structural properties of the function f. Even though the analogous problem for trigonometric approximation has been classical for a long time, the characterization problem for algebraic polynomial approximation was only solved in the eighties [15], [16], [17], [11] due to the complex nature of the behavior of the best approximating polynomials around the endpoints ± 1. Eventually, as the complete analogue of the trigonometric case, the following result was proved [11, Chap. 8]: let $\varphi(x) = \sqrt{1-x^2}$, and

$$\omega_\varphi^r(f,\delta)_p = \sup_{0\le h\le\delta} \|\Delta_{h\varphi}^r f\|_p,$$

where

$$\Delta_t^r f(x) = \sum_{k=0}^{r} (-1)^k \binom{r}{k} f\left(x + \left(\frac{r}{2} - k\right)t\right)$$

is the rth symmetric difference of f with increment t ($= h\varphi(x)$ in the definition of ω_φ^r) and where we set $\Delta_t^r f(x) = 0$ if any of the arguments in its defining expression does not belong to $[-1,1]$. Then for a positive integer r we have

$$E_n(f)_p \le C\omega_\varphi^r\left(f, \frac{1}{n}\right)_p$$

and its converse

$$\omega_\varphi^r\left(f, \frac{1}{n}\right)_p \le Cn^{-r}\sum_{k=1}^{n} k^{r-1}E_k(f)_p.$$

* Received by the editors June 19, 1992; accepted for publication (in revised form) February 8, 1993.

† Department of Mathematics, University of Witwatersrand, P.O. Box Wits 2050, Republic of South Africa.

‡ Bolyai Institute, Szeged, Aradi v. tere 1, 6720, Hungary and Department of Mathematics, University of South Florida, Tampa, Florida 33620. This work was supported by National Science Foundation grant DMS 9101380 and by Hungarian Science Foundation for Research grant 6529, and it was done while the author visited the University of the Witwatersrand, in the Spring of 1992.

In particular, for $0 < \alpha < r$,

$$E_n(f)_p = O(n^{-\alpha}) \Leftrightarrow \omega_\varphi^r(f,\delta)_p = O(\delta^\alpha).$$

The situation is even more complicated if we discuss weighted polynomial approximation:

$$E_n(f)_{w,p} = \inf_{\deg P_n \le n} \|w(f-P_n)\|_p,$$

namely the problem has only been solved for Jacobi-like weights ([11, Thms. 8.2.1 and 8.2.4]) with a weighted modulus of smoothness that is similar in spirit to the ω_φ^r above.

Let us only discuss the case of Jacobi weights

$$w(x) = (1-x)^\alpha (1+x)^\beta$$

with $\alpha, \beta > 0$. We set for $n = 1, 2, \dots$

$$w_n(x) = \left(\sqrt{1-x} + \frac{1}{n}\right)^{2\alpha} \left(\sqrt{1+x} + \frac{1}{n}\right)^{2\beta}$$

and

$$\varphi_n(x) = \sqrt{1-x^2} + \frac{1}{n}.$$

The crux of the matter is the Jackson–Favard type estimate

$$E_n(f)_{w,p} \le C\frac{1}{n^r}\|w_n \varphi_n^r f^{(r)}\|_p \tag{1}$$

with a matching converse Markov–Bernstein inequality

$$\|w_n \varphi_n^r P_n^{(r)}\|_p \le C n^r \|w_n P_n\|_p, \qquad \deg P_n \le n. \tag{2}$$

The rest belongs to the theory of interpolation spaces and to characterizations of weighted K-functionals (see [11, Chap. 8]). By now there are several different procedures known for proving (2) (see [21], [18], [19], and [11, Thm. 8.4.7]), but the only existing way to get to (1) is by transforming the problem to the trigonometric case and to use quite complicated estimates based on Riesz's interpolation formula [11, Chap. 8].

In the present paper we offer an alternative approach to (1). The method was developed by Freud [13], [14] in connection with exponential weights on the whole real line, in which case he based his approach on the orthogonal polynomials that now carry his name. We shall use Jacobi expansions in our method with respect to the weight w^2. Many steps will be the same as Freud's case, but, unlike Freud, we have to overcome the additional difficulty of the weight φ entering our formulas.

A crucial step in Freud's method is the boundedness of $(C,1)$ means of the expansions in question in the appropriate weighted spaces, which is also valid for Jacobi expansions for all parameter values $\alpha, \beta \ge 0$. This is where our work gets connected with the research of Askey, who has many remarkable discoveries concerning Cesaro means of Jacobi expansions (see [1]–[10]).

2. Some facts concerning Jacobi polynomials. Let

$$w(x) = (1-x)^{\alpha}(1+x)^{\beta}$$

be a Jacobi weight with $\alpha,\ \beta > -\frac{1}{2}$. We consider the Jacobi polynomials

$$p_n(x) = \gamma_n x^n + \cdots, \qquad \gamma_n > 0$$

associated with w^2 (sic!), i.e., for which

$$\int_{-1}^{1} p_n p_m w^2 = \delta_{nm}, \qquad n, m = 0, 1, \ldots.$$

We shall need a few classical results for them, all taken from Szegő's book [22].

In what follows we shall write $A \prec B$ if $A = O(B)$ and $A \sim B$ if $A = O(B)$ and $B = O(A)$.

First of all, uniformly in n and $x \in [-1, 1]$,

$$p_n(x) \prec 1 \Big/ \left(w_n(x)\varphi_n(x)^{1/2}\right) \tag{3}$$

(see [22, Thm. 7.32.2] and recall that $P_n^{(2\alpha,2\beta)}$ in [22] has to be normalized by the factor $h_n^{(2\alpha,2\beta)}$ of [22, eq. (4.3.3)] to get p_n). For any fixed $c > 0$,

$$p_n(\cos\theta) = b_n \left(\sin\frac{\theta}{2}\right)^{-2\alpha-1/2} \left(\cos\frac{\theta}{2}\right)^{-2\beta-1/2} \left\{\cos(N\theta+\gamma) + O((n\sin\theta)^{-1})\right\} \tag{4}$$

for

$$\frac{c}{n} \le \theta \le \pi - \frac{c}{n},$$

where

$$N = n + (2\alpha + 2\beta + 1)/2, \qquad \gamma = -\left(2\alpha + \frac{1}{2}\right)\frac{\pi}{2},$$

and b_n tends to a finite and positive limit as $n \to \infty$ (see [22, Thm. 8.21.13]); furthermore, say around $+1$,

$$p_n(\cos\theta) \sim n^{2\alpha+1/2}, \qquad 0 \le \theta \le \frac{c}{n} \tag{5}$$

for sufficiently small c (see [22, Thm. 8.21.12]).

These easily imply for the so-called Christoffel function

$$\lambda_n(w^2, x) := \lambda_n(x) := \left\{\sum_{k=0}^{n-1} p_k(x)^2\right\}^{-1} \tag{6}$$

the uniform estimate

$$\lambda_n(x) \sim \frac{w_n^2(x)\varphi_n(x)}{n} \tag{7}$$

(cf. also [20]).

We shall also need that for $\alpha,\ \beta > -1$,

$$\lambda_n(w_n,x) \sim \frac{w_n(x)\varphi_n(x)}{n}. \tag{8}$$

To prove this we have to recall (see [12]) that for any weight v on $[-1,1]$,

$$\lambda_n(v,x) = \inf_{\deg P_n \le n-1} \frac{1}{P_n^2(x)} \int_{-1}^{1} P_n^2 v.$$

Since $w \le w_n$, this formula immediately gives

$$\lambda_n(w_n,x) \ge \lambda_n(w,x) \ge c\frac{w_n(x)\varphi_n(x)}{n},$$

where, at the last step we applied (7) for w rather than w^2.

Equation (7) also shows that if $\{p_k^{**}\}$ are the orthonormal Jacobi polynomials with respect to the weight

$$w(x)^{1/2}/\varphi(x) = (1-x)^{(\alpha-1)/2}(1+x)^{(\beta-1)/2},$$

then

$$R_{n/2}(x) = \frac{1}{n}\sum_{k=0}^{n/4} p_k^{**}(x)^2$$

is a polynomial of degree at most $n/2$ with

$$R_{n/2}^2(x) \sim \frac{1}{w_n(x)}.$$

Hence

$$\begin{aligned}
\lambda_n(w_n,x) &\prec \inf_{\deg Q_{n/2}<n/2} \frac{1}{R_{n/2}^2(x)Q_{n/2}^2(x)} \int_{-1}^{1} (R_{n/2}Q_{n/2})^2 w_n \\
&\sim w_n(x) \inf_{\deg Q_{n/2}<n/2} \frac{1}{Q_{n/2}^2(x)} \int_{-1}^{1} Q_{n/2}^2 \\
&\sim w_n(x)\lambda_{n/2}(1,x) \sim w_n(x)\frac{\varphi_n(x)}{n},
\end{aligned}$$

where at the very last step we again used (7) for the Legendre weight

$$1 = (1-x)^0(1+x)^0.$$

Finally, it easily follows from the formula [22, eq. (4.5.5)] that

$$\begin{aligned}
p_{k-1}(x) - p_{k+1}(x) = d_k(1-x^2)p_{k-1}^*(x) \\
+ O\left(\frac{1}{k}(|p_{k-1}(x)| + |p_k(x)| + |p_{k+1}(x)|)\right)
\end{aligned}$$

uniformly in n and $x \in [-1,1]$, where $\{d_k\}$ is a positive sequence converging to 1, and $p_{k-1}^*(x)$ is the $(k-1)$st Jacobi polynomial associated with the weight

$$w^2(x)\varphi^2(x) = (1-x)^{2\alpha+1}(1+x)^{2\beta+1}.$$

This formula immediately implies

$$\sum_{k=1}^{n}(p_{k-1}(x)-p_{k+1}(x))^2 \prec (1-x^2)^2\lambda_n^*(x)^{-1}+\sum_{k=0}^{n+1}\frac{1}{(k+1)^2}p_k^2(x), \tag{9}$$

where λ_n^* is the Christoffel function associated with the system $\{p_k^*(x)\}$. The first term on the right is (see (7))

$$(1-x^2)^2\lambda_n^*(x)^{-1} \prec \frac{n\varphi_n(x)}{w_n^2(x)}, \tag{10}$$

while for the second term we easily get from (3) that it is

$$\prec \sum_{k=1}^{n+1}\frac{1}{k^2w_k^2(x)\varphi_k(x)}.$$

For $0 \le x < 1$ we break the last sum into two parts with the ranges $1 \le k \le \varphi_n^{-1}(x)$ and the rest. When $0 < \alpha \le \frac{1}{4}$ this yields

$$\sum_{k=1}^{n+1}\frac{1}{k^2w_k^2(x)\varphi_k(x)} \prec \sum_{k=1}^{\varphi_n^{-1}(x)}\frac{1}{k^{1-4\alpha}}\frac{1}{(k\sqrt{1-x}+1)^{4\alpha+1}}+\sum_{k>\varphi_n^{-1}(x)}\frac{1}{k^2}\frac{1}{w_n^2(x)\varphi_n(x)}$$

because for $n+1 \ge k > \varphi_n^{-1}(x)$ we have

$$\sqrt{1-x}+\frac{1}{k} \sim \sqrt{1-x}+\frac{1}{n},$$

and so $w_k^2(x)\varphi_k(x) \sim w_n^2(x)\varphi_n(x)$. Finally, both sums on the right are of the order

$$\sum_{k=1}^{\varphi_n^{-1}(x)}\frac{1}{k^{1-4\alpha}} \sim \frac{1}{w_n^2(x)\varphi_n(x)}\sum_{k>\varphi_n^{-1}(x)}\frac{1}{k^2} \sim w_n^{-2}(x).$$

When $\alpha > \frac{1}{4}$ similar consideration gives

$$\begin{aligned}\sum_{k=1}^{n+1}\frac{1}{k^2w_k^2(x)\varphi_k(x)} &\prec \frac{1}{(\sqrt{1-x}+\frac{1}{n})^{4\alpha-1}}\sum_{k=1}^{\varphi_n^{-1}(x)}1+\frac{1}{w_n^2(x)\varphi_n(x)}\sum_{k>\varphi_n^{-1}(x)}\frac{1}{k^2}\\ &\prec \frac{1}{w_n^2(x)}.\end{aligned}$$

The argument is similar when $-1 \le x \le 0$.

Our estimates proved so far show that

$$\sum_{k=0}^{n+1}\frac{1}{(k+1)^2}p_k^2(x) \prec \frac{1}{w_n^2(x)}. \tag{11}$$

This, (9), and (10) finally yield

$$\sum_{k=1}^{n}(p_{k-1}(x)-p_{k+1}(x))^2 \prec \frac{n\varphi_n(x)}{w_n^2(x)}. \tag{12}$$

The last fact we need on the Jacobi polynomials is that the leading coefficients γ_n satisfy

$$\left|\frac{\gamma_k}{\gamma_{k+1}} - \frac{1}{2}\right| \prec \frac{1}{k^2} \tag{13}$$

(see [22, eq. (4.5.1), eq. (3.2.2)]).

3. Boundedness of $(C,1)$ means in weighted norms. Let f be a function such that the integrals for the Fourier coefficients

$$c_n(f) = \int_{-1}^{1} fw^2$$

exist. These give rise to the expansion

$$f(x) \sim \sum_{n=0}^{\infty} c_n(f)p_n(x)$$

with partial sums

$$S_k(f,x) = \sum_{j=0}^{k} c_j(f)p_j(x).$$

The $(C,1)$ means are then defined as

$$\sigma_n(f,x) = \frac{1}{n}\sum_{k=1}^{n} S_k(f,x).$$

We shall need the following in the proof of (1).

STATEMENT 3.1. *If $w(x) = (1-x)^\alpha(1+x)^\beta$ with $\alpha, \beta > 0$, then for any $1 \le p \le \infty$*

$$\|w_n\sigma_n(f)\|_p \le C\|wf\|_p \tag{14}$$

with a constant C depending only on α and β.

Proof. First we consider the case $p = \infty$. From here the Statement for $p = 1$ will follow by standard duality argument, although it causes some difficulty that on the left of (14) not the weight w but w_n appears. The case $1 < p < \infty$ follows then by interpolation.

3.1. The case $p = \infty$. Suppose that $\|wf\|_{[-1,1]} \le 1$, and $x \in [-1,1]$ is arbitrary. We write

$$f_n(t) = f_{n,x}(t) = \begin{cases} f(t) & \text{if } |x-t| \le \varphi_n(x)/n, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$F_n(t) = F_{n,x}(t) = \begin{cases} f(t)/(x-t) & \text{if } |x-t| > \varphi_n(x)/n, \\ 0 & \text{otherwise.} \end{cases}$$

If

$$K_k(x,t) = \sum_{j=0}^{k} p_j(x)p_j(t) = \frac{\gamma_k}{\gamma_{k+1}}\frac{p_{k+1}(x)p_k(t) - p_k(x)p_{k+1}(t)}{x-t}$$

is the reproducing kernel for the system $\{p_j\}$, then

$$S_k(f,x) = \int f_n(t)K_k(x,t)w^2(t)dt + \frac{\gamma_k}{\gamma_{k+1}}(c_k(F_n)p_{k+1}(x) - c_{k+1}(F_n)p_k(x)) = I_k(x) + J_k(x),$$

where

$$c_k(F_n) = \int F_n(t)p_k(t)w^2(t)dt$$

are the Fourier coefficients of F_n with respect to $\{p_k\}$. We estimate the two terms in the expression of S_k separately.

For $I_k(x)$ we can write in view of $\|wf\| \le 1$

$$\begin{aligned} |I_k(x)| &\le \int_{|t-x|\le\varphi_n(x)/n} |K_k(x,t)|w(t)dt \\ &\le 2\left(\frac{\varphi_n(x)}{n}\right)^{1/2}\left(\int K_k^2(x,t)w^2(t)dt\right)^{1/2} \\ &= 2\left(\frac{\varphi_n(x)}{n}\right)^{1/2}\lambda_{k+1}^{-1/2}(x) \prec \frac{1}{w_n(x)}, \end{aligned} \tag{15}$$

provided $k \le n$, where in the last step we used (7).

For γ_k/γ_{k+1} in J_k we use (13) to obtain

$$J_k(x) = \frac{1}{2}(c_k(F_n)p_{k+1}(x) - c_{k+1}(F_n)p_k(x)) + O\left(|c_k(F_n)|\frac{|p_{k+1}(x)|}{(k+1)^2} + |c_{k+1}(F_n)|\frac{|p_k(x)|}{(k+1)^2}\right) = J_k^{(1)}(x) + J_k^{(2)}(x).$$

In

$$\sigma_n(f,x) = \frac{1}{n}\sum_{k=1}^{n} S_k(f,x)$$

the sum of the "errors" $J_k^{(2)}(x)$ can be estimated as

$$\sum_{k=1}^{n}|J_k^{(2)}(x)| \le \left(\sum_k c_k(F_n)^2\right)^{1/2}\left(\sum_{k=1}^{n+1}\frac{p_k^2(x)}{k^4}\right)^{1/2}.$$

Making use of Bessel's inequality, $\|wf\| \le 1$ and (11), this can be continued as

$$\begin{aligned} &\prec\left(\int F_n(t)^2w(t)^2dt\right)^{1/2}\frac{1}{w_n(x)} \\ &= \left(\int_{|x-t|\ge\varphi_n(x)/n}\frac{1}{|x-t|^2}dt\right)^{1/2}\frac{1}{w_n(x)} \\ &\le 2\left(\frac{n}{\varphi_n(x)}\right)^{1/2}\frac{1}{w_n(x)}. \end{aligned} \tag{16}$$

The sum of the first terms in $J_k(x)$ is half of

$$2\sum_{k=1}^{n} J_k^{(1)}(x) = c_1(F_n)p_2(x) - c_{n+1}(F_n)p_n(x) + \sum_{k=2}^{n} c_k(F_n)(p_{k+1}(x) - p_{k-1}(x)).$$

For the last expression we obtain from the Schwarz inequality, Bessel inequality, and (12)

$$\left|\sum_{k=2}^{n} c_k(F_n)(p_{k+1}(x) - p_{k-1}(x))\right| \prec \left(\int F_n(t)^2 w(t)^2 dt\right)^{1/2} \frac{(n\varphi_n(x))^{1/2}}{w_n(x)} \prec \frac{n}{w_n(x)}.$$

A similar argument gives that

$$|c_1(F_n)p_2(x)| \prec \frac{n}{w_n(x)}.$$

As for the term $|c_{n+1}(F_n)||p_n(x)|$, we write with (3)

$$\begin{aligned} |c_{n+1}(F_n)||p_n(x)| &\prec \frac{\|wf\|}{w_n(x)\varphi_n^{1/2}(x)} \left(\int_{|x-t|\geq\varphi_n(x)/n} \frac{|p_{n+1}(t)|w(t)}{|x-t|}\, dt\right) \\ &\prec \frac{1}{w_n(x)\varphi_n^{1/2}(x)} \int_{|x-t|>\varphi_n(x)/n} \frac{1}{|x-t|\varphi_n^{1/2}(t)}\, dt, \end{aligned} \tag{17}$$

and so the inequality

$$\int_{|x-t|>\varphi_n(x)/n} \frac{1}{|x-t|\varphi_n^{1/2}(t)}\, dt \prec n\varphi_n^{1/2}(x), \tag{18}$$

which we prove in a short while, gives the estimate

$$|c_{n+1}(F_n)||p_n(x)| \prec \frac{n}{w_n(x)}.$$

Collecting our estimates we finally arrive at

$$|\sigma_n(f;x)| \prec \frac{1}{w_n(x)} \tag{19}$$

provided $\|wf\| \leq 1$, by which the inequality

$$\|w_n\sigma_n(f)\|_\infty \leq C\|wf\|_\infty \tag{20}$$

has been verified.

It remains to show (18) to complete the case $p = \infty$. Suppose for example that $0 \leq x < 1$. If $x > 1 - 4/n^2$, then the integral is obviously bounded by

$$\prec \int_0^{1-n^{-2}} \frac{1}{(1-t)^{5/4}}\, dt \prec n^{1/2} \prec n\varphi_n^{1/2}(x).$$

If, however, $0 \le x < 1 - 4/n^2$, then we write for the the integral

$$\int_{\varphi_n(x)/n \le |x-t| \le (1-x)/2} + \int_{t-x>(1-x)/2} + \int_{x-t>(1-x)/2}$$
$$\prec \frac{1}{\varphi_n^{1/2}(x)} \int_{\varphi_n(x)/n \le |x-t| \le (1-x)/2} \frac{1}{|x-t|} dt + \frac{1}{(1-x)} \int_{t-x>(1-x)/2} \frac{1}{(1-t)^{1/4}}\, dt$$
$$+ \int_{x-t>(1-x)/2} \frac{1}{(1-t)^{5/4}}\, dt \prec \frac{1}{\varphi_n^{1/2}(x)} \log \frac{n(1-x)}{2\varphi_n(x)} + \frac{(1-x)^{3/4}}{(1-x)} + \frac{1}{(1-x)^{1/4}}$$
$$\prec n\varphi_n^{1/2}(x),$$

where we used that $n\varphi_n(x) \ge 1$, and $(\log u)/u$ is bounded for $u \in [1, \infty)$.

We shall also need that

$$\|w_n \varphi_n \sigma_n(f)\|_\infty \le C \|w\varphi_n f\|_\infty. \tag{21}$$

If we go through the preceding proof we can see that there are three places where the assumption $\|wf\|_\infty \le 1$ is used, and these are when we estimated the integrals

$$I_k(x) = \int_{|x-t| \le \varphi_n(x)/n} f(t) K_k(x,t) w^2(t) dt,$$
$$B_n(x) := \left(\int_{|x-t| > \varphi_n(x)/n} \frac{f(t)^2 w(t)^2}{(x-t)^2} dt \right)^{1/2},$$

and

$$R_n(x) := \int_{|x-t| \ge \varphi_n(x)/n} \frac{|f(t) p_{n+1}(t)| w^2(t)}{|x-t|}\, dt,$$

and for these we got the estimates

$$|I_k(x)| \prec \|wf\| \left(\frac{\varphi_n(x)}{n} \right)^{1/2} \lambda_{k+1}^{-1/2}(x), \tag{22}$$

$$|B_n(x)| \prec \|wf\| \left(\frac{n}{\varphi_n(x)} \right)^{1/2}, \tag{23}$$

and

$$R_n(x) \prec \|wf\| n\varphi_n^{1/2}(x) \tag{24}$$

(see (15)–(18)).

Since

$$\varphi_n(x) \sim \varphi_n(t) \quad \text{when } |x-t| < \frac{\varphi_n(x)}{n},$$

together with the first one we can also get

$$\varphi_n(x) |I_k(x)| \prec \|w_n \varphi_n f\| \left(\frac{\varphi_n(x)}{n} \right)^{1/2} \lambda_{k+1}^{-1/2}(x). \tag{25}$$

To prove the analogue of (23) we write

$$\varphi_n(x)B_n(x) \le \|w_n\varphi_n f\| \left(\int_{|x-t|>\varphi_n(x)/n} \frac{1}{(x-t)^2\varphi_n(t)^2} dt \right)^{1/2} \varphi_n(x). \tag{26}$$

Here the estimate of the integral is similar to that of (18). In fact, suppose $0 \le x < 1$. Then the integral can be restricted to $0 \le t \le 1$, because the integral over this range majorizes that of over the range $-1 \le t \le 0$. If $x > 1 - 4/n^2$, then the integral is bounded by

$$2\frac{n}{\varphi_n(x)}n^2 \le 2\frac{n}{\varphi_n(x)}\frac{1}{\varphi_n^2(x)}.$$

Similarly, for all $0 \le x < 1$,

$$\int_{\substack{1-4/n^2\le t\le 1\\ |x-t|\ge\varphi_n(x)/n}} \frac{1}{(x-t)^2\varphi_n(t)^2}\, dt \prec \frac{1}{n^2}\frac{1}{\varphi_n^4(x)}n^2 \prec \frac{n}{\varphi_n(x)}\frac{1}{\varphi_n^2(x)}$$

(check this separately for $x > 1 - 8n^{-2}$ and for $x \le 1 - 8n^{-2}$). If, however, $0 \le x < 1 - 4/n^2$, then for the rest of the integral we write

$$\begin{aligned} &= \int_{\substack{\varphi_n(x)/n\le|x-t|\le(1-x)/2\\ 0\le t\le 1-4/n^2}} + \int_{\substack{t-x>(1-x)/2\\ 0\le t\le 1-4/n^2}} + \int_{\substack{x-t>(1-x)/2\\ 0\le t\le 1-4/n^2}} \\ &\prec \frac{1}{\varphi_n^2(x)} \int_{\varphi_n(x)/n\le|x-t|\le(1-x)/2} \frac{1}{(x-t)^2}\, dt + \frac{1}{(1-x)^2}\int_{\substack{t-x>(1-x)/2\\ 0\le t\le 1-4/n^2}} \frac{1}{1-t}dt \\ &\quad + \int_{\substack{x-t>(1-x)/2\\ 0\le t\le 1-4/n^2}} \frac{1}{(1-t)^3}dt \prec \frac{n}{\varphi_n(x)}\frac{1}{\varphi_n^2(x)} + \frac{1}{(1-x)^2}\log(n^2(1-x)) + \frac{1}{(1-x)^2}. \end{aligned}$$

Since $0 \le x \le 1 - 4/n^2$, we have

$$\frac{1}{(1-x)^2}\log(n^2(1-x)) \prec \frac{1}{\varphi_n(x)^4}\log(n^2\varphi_n^2(x)) \prec \frac{n}{\varphi_n(x)}\frac{1}{\varphi_n^2(x)}$$

because $n\varphi_n(x) \ge 1$, and $(\log u)/u$ is bounded for $u \in [1, \infty)$.

Thus, we get

$$\varphi_n(x)B_n(x) \le \|w_n\varphi_n f\| \left(\frac{n}{\varphi_n(x)}\right)^{1/2}. \tag{27}$$

Finally, the analogue of (24), namely, that

$$\varphi_n(x) \int_{|x-t|\ge\varphi_n(x)/n} \frac{|f(t)p_{n+1}(t)|w^2(t)}{|x-t|}\, dt \prec \|w\varphi_n f\| n\varphi_n^{1/2}(x), \tag{28}$$

can be verified with the method that we used in the proof of (18).

Using (25), (27), and (28) instead of (22)–(24) we can prove (21) with the same argument that we used for (20). In fact, the rest of the proof is the same, word for word, as before.

3.2. The case $p = 1$. To obtain (14) for $p = 1$ from the same inequality with $p = \infty$, we use a standard duality argument. But first we need the following: if P_m is a polynomial of degree at most m, then for every $c > 0$ there is a constant C such that

$$\|w_m P_m\|_p \le C\|w_m P_m\|_{L^p[-1+cm^{-2},1-cm^{-2}]}. \tag{29}$$

Actually we shall need this inequality only for some $c > 0$, but once it is verified with a $c > 0$, we can get it for larger and larger c's (and eventually for all c) by dilation and iteration. Thus, let $c > 0$ be a small number.

The Chebyshev–Markov inequality (2) implies

$$\|w_m P_m'\|_p \le C_1 m^2 \|w_m P_m\|_p. \tag{30}$$

We write

$$Q_m(x) := P_m(x) - P_m(x(1 - cm^{-2})) = \int_{x(1-cm^{-2})}^{x} P_m'(t)\, dt,$$

and use the fact that

$$\left|\frac{w_m(x)}{w_m(y)} - 1\right| \le \eta(c) \quad \text{if } |x - y| \le \frac{c}{m^2}, \tag{31}$$

where $\eta(c) \to 0$ as $c \to 0$. We can easily get from this and (30) combined with Hölder's inequality, that

$$\|w_m Q_m\|_p \le \varepsilon(c)\|w_m P_m\|_p,$$

where $\varepsilon(c) \to 0$ as $c \to 0$. This, together with (31), implies

$$\left\|w_m(x(1 - cm^{-2}))P_m(x(1 - cm^{-2}))\right\|_p > \tfrac{1}{2}\|w_m P_m\|_p$$

for sufficiently small c, and this is the same as (29).

Let us now return to (14) with $p = 1$. We use (29), and observe that on the interval $[-1 + c/n^2, 1 - c/n^2]$ we have $w(x) \sim w_n(x)$. Hence

$$\begin{aligned}\|w_n\sigma_n(f)\|_1 &\le C\int_{-1+c/n^2}^{1-c/n^2} w(t)|\sigma_n(f,t)|dt \\ &= C\sup_{\|wg\|_\infty \le 1}\int_{-1}^{1}\sigma_n(f,t)g(t)w^2(t)dt,\end{aligned}$$

where the supremum is taken for all g that vanish outside the interval

$$[-1 + c/n^2, 1 - c/n^2]$$

and satisfy $\|wg\|_\infty \le 1$. Continuing this inequality, we get from (14) with $p = \infty$

$$\begin{aligned}\|w_n\sigma_n(f)\|_1 &\le C\sup_g \int_{-1}^{1} f(t)\sigma_n(g,t)w^2(t)dt \\ &\le C\sup_g \|w\sigma_n(g)\|_\infty \left(\int w|f|\right) \\ &\le C\sup_g \|wg\|_\infty \|wf\|_1 \le C\|wf\|_1.\end{aligned}$$

3.3. Completion of the proof. From the cases $p = 1$ and $p = \infty$ we get (14) for all $1 \le p \le \infty$ by the Riesz–Thorin theorem. □

We shall also need the delayed arithmetic means of the Jacobi expansion of a function f defined as

$$\rho_n(f,x) = 2\sigma_{2n}(f,x) - \sigma_n(f,x). \tag{32}$$

For any polynomial P_n of degree at most n we have

$$\rho_n(P_n,x) \equiv P_n(x), \tag{33}$$

and for any f the function $\rho_n(f,x)$ is a polynomial of degree at most $2n$. Furthermore, we get from (14)

$$\|w_n\rho_n(f)\|_p \le C\|wf\|_p \tag{34}$$

for all f and p.

4. Proof of the Jackson–Favard type estimate (1). We closely follow the method of Freud [13], [14]. This consists of the following steps built on one another.

Step 1. Weighted approximation in L^1 metric of a piecewise constant function with a single jump.

Step 2. Approximation of functions of bounded variation in L^1 metric.

Step 3. A Bohr-type inequality.

Step 4. The L^∞ case of (1), and a similar estimate for some linear means.

Step 5. Interpolation of the L^1 and L^∞ cases, and iteration of the obtained estimate.

The individual steps are not too long, and they go as follows.

4.1. Step 1. Approximation of Γ_ξ in L^1. Let

$$\Gamma_\xi(t) = \begin{cases} 0 & \text{if } -1 \le t \le \xi, \\ 1 & \text{if } \xi < t < 1. \end{cases}$$

It is well known (see [12]) that

$$E_n(\Gamma_\xi)_{w_n,1} \le \lambda_{[n/2]+2}(w_n,\xi) = \int \lambda_{[n/2]+2}(w_n,t)d\Gamma_\xi(t). \tag{35}$$

4.2. Step 2. Approximation of functions of bounded variation in L^1. If f is a right continuous function of bounded variation, and

$$f_m(x) = f(-1) + \sum_{k=1}^{2m}\left[f\left(-1+\frac{k}{m}\right) - f\left(-1+\frac{k-1}{m}\right)\right]\Gamma_{-1+k/m}(x),$$

then $f_m \to f$ in $L^1[-1,1]$ and

$$\int \lambda_{[n/2]+2}(w_n,t)|df_m(t)| \to \int \lambda_{[n/2]+2}(w_n,t)|df(t)|$$

as $m \to \infty$. By (35),

$$\begin{aligned} E_n(f_m)_{w_n,1} &\le \sum_{k=1}^{2m}\left|f\left(-1+\frac{k}{m}\right) - f\left(-1+\frac{k-1}{m}\right)\right| E_n(\Gamma_{-1+k/m})_{w_n,1} \\ &\le \int \lambda_{[n/2]+2}(w_n,t)|df_m(t)|, \end{aligned}$$

and for $m \to \infty$ we obtain

$$E_n(f)_{w_n,1} \le \int \lambda_{[n/2]+2}(w_n,t)|df(t)| \prec \frac{1}{n}\int w_n(t)\varphi_n(t)|df(t)|, \tag{36}$$

where, at the very last step we made use of (8).

4.3. Step 3. A Bohr-type inequality. We show that if

$$\int_{-1}^{1} gP_n w^2 = 0$$

for every polynomial P_n of degree at most n, then

$$w_n(x)\left|\int_0^x g(t)dt\right| \prec \frac{1}{n}\|w_n\varphi_n g\|_\infty. \tag{37}$$

Without loss of generality let $0 \le x < 1$, and consider first the case $0 \le x \le 1 - 1/n^2$. Let

$$\phi_x(t) = \begin{cases} w^{-2}(t) & \text{if } t \in [0,x], \\ 0 & \text{otherwise.} \end{cases}$$

With an appropriate polynomial P_n of degree at most n

$$\left|\int_0^x g(t)dt\right| = \left|\int g(t)(\phi_x(t) - P_n(t))w^2(t)dt\right| \le \|w\varphi_n g\|_\infty E_n(\phi_x)_{w_n/\varphi_n,1},$$

and if we use (36) for the weight $w_n/\varphi_n \sim (w/\varphi)_n$ rather than for w_n, then we can continue this as

$$\prec \|w\varphi_n g\|_\infty \frac{1}{n}\int w_n(t)|d\phi_x(t)|$$

$$\prec \|w\varphi_n g\|_\infty \frac{1}{n}\left(\frac{1}{w_n(x)} + \frac{1}{w_n(0)} + \int_0^x w_n(t)|(w^{-2}(t))'|dt\right).$$

Here

$$\left|\left(\frac{1}{w^2(t)}\right)'\right| \prec \frac{1}{w^2(t)\varphi^2(t)};$$

hence for $0 \le x \le 1 - 1/n^2$,

$$\int_0^x w_n(t)|(w^{-2}(t))'|dt \prec \int_0^x \frac{1}{w(t)\varphi^2(t)}dt \prec \int_0^x \frac{1}{(1-t)^{\alpha+1}}dt \prec \frac{1}{w(x)} \prec \frac{1}{w_n(x)}.$$

Since $w_n(x) \prec w_n(0)$, (37) follows from the previous computations.

To prove (37) for the case $1 - 1/n^2 \le x \le 1$, as well, we write

$$\left|\int_0^x g(t)dt\right| \prec \left|\int_0^{1-1/n^2} + \int_{1-1/n^2}^1\right| \prec \frac{1}{n}\frac{\|w_n\varphi_n g\|_\infty}{w_n(1-1/n^2)} + \|w_n\varphi_n g\|_\infty \int_{1-1/n^2}^1 n^{2\alpha+1}$$
$$\prec \frac{1}{n}\|w_n\varphi_n g\|\frac{1}{w_n(x)},$$

which is what we needed to prove.

4.4. Step 4. The L^∞ case of (1). Set

$$g = f' - \rho_{[n/2]-1}(f', x),$$

where the $\rho_k(f;x)$'s are the means defined in (32). We obtain from (33) that the preceding lemma can be applied to g with n replaced by $[n/2]-1$. This yields

$$E_n(f)_{w_n,\infty} \le \|w_n(x)\int_0^x g(t)dt\|_\infty$$
$$\prec \frac{1}{n}\|w_n\varphi_n(f' - \rho_{[n/2]-1}(f', x))\|_\infty$$
$$\prec \frac{1}{n}\|w_n\varphi_n f'\|_\infty,$$

where at the last step we used (21).

Since $\rho_n(P_n) \equiv P_n$ for every polynomial of degree at most n, we also get from the preceding estimate with some appropriate P_n that

$$\|w_n(f-\rho_n(f))\|_\infty = \|w_n\left((f-P_n) - \rho_n(f-P_n)\right)\|_\infty \tag{38}$$
$$\prec E_n(f)_{w_n,\infty} \prec \frac{1}{n}\|w_n\varphi_n f'\|_\infty,$$

where in the last but one step we applied the main result of §3, namely (14).

4.5. Step 5. Interpolation of the L^1 and L^∞ cases and iteration of the estimate obtained. We know from Step 2 that with some appropriate P_n,

$$\|w_n(f-\rho_n(f))\|_1 = \|w_n\left((f-P_n) - \rho_n(f-P_n)\right)\|_1 \tag{39}$$
$$\prec E_n(f)_{w_n,1} \prec \frac{1}{n}\|w_n\varphi_n f'\|_1,$$

where we once more utilized (14).

From (38) and (39) we get for any $1 \le p \le \infty$ by interpolation,

$$\|w_n(f-\rho_n(f))\|_p \prec \frac{1}{n}\|w_n\varphi_n f'\|_p;$$

apply interpolation to the operator defined for $h = w_n\varphi_n f'$ as

$$Th(x) = (f(x) - \rho_n(f,x))w_n(x)$$
$$= \left(\int_0^x \frac{h(t)}{w_n(t)\varphi_n(t)}dt - \rho_n\left(\int_0 \frac{h(t)}{w_n(t)\varphi_n(t)}dt; x\right)\right)w_n(x).$$

This yields

$$E_{2n}(f)_{w_n,p} \prec \frac{1}{n}\|w_n\varphi_n f'\|_p,$$

and hence

$$E_n(f)_{w_n,p} \le E_{2[n/2]}(f)_{w_{[n/2],p}} \prec \frac{1}{n}\|w_{[n/2]}\varphi_{[n/2]}f'\| \prec \frac{1}{n}\|w_n\varphi_n f'\|,$$

which in turn implies by subtracting an appropriate polynomial of degree at most $n-1$ from f that

$$E_n(f)_{w_n,p} \prec \frac{1}{n}E_{n-1}(f')_{w_n\varphi_n,p}.$$

This is already in a form that can be iterated (note that $w_n(x) \sim w_{n-1}(x)$), and we obtain

$$E_n(f)_{w_n,p} \prec \frac{1}{n^m}E_{n-m}(f^{(m)})_{w_n\varphi_n^m,p}$$

for every $m \ge 1$, which shows that

$$E_n(f)_{w_n,p} \prec \frac{1}{n^m}\|w_n\varphi_n^m f^{(m)}\|_p,$$

as we claimed in (1).

REFERENCES

[1] R. ASKEY, *Jacobi polynomial expansions with positive coefficients and imbeddings of projective spaces*, Bull. Amer. Math. Soc., 74 (1968), pp. 301–304.

[2] ———, *Mean convergence of orthogonal series and Lagrange interpolation.* Acta Math. Acad. Sci. Hungar., 23 (1972), pp. 71–85.

[3] ———, *Summability of Jacobi series*, Trans. Amer. Math. Soc., 179 (1973), pp. 71–84.

[4] ———, *Jacobi polynomials* I., *New proofs of Koornwinder's Laplace type integral representation and Bateman's bilinear sum*, SIAM J. Math. Anal., 5 (1974), pp. 119–124.

[5] R. ASKEY AND J. FITCH, *Positivity of the Cotes numbers for some ultraspherical abscissas*, SIAM J. Numer. Anal., 5 (1968), pp. 198–201.

[6] ———, *Integral representations for Jacobi polynomials and some applications*, J. Math. Anal. Appl., 26 (1969), pp. 11–437.

[7] R. ASKEY AND G. GASPER, *Linearization of the product of Jacobi polynomials* III, Canad. J. Math., 23 (1971), pp. 332–338.

[8] ———, *Jacobi polynomial expansions of Jacobi polynomials with non-negative coefficients*, Proc. Cambridge Philos. Soc., 70 (1971), pp. 243–255.

[9] R. ASKEY AND I. I. HIRSCHMAN, *Mean summability for ultrasperical polynomials*, Math. Scand., 12 (1963), pp. 167–177.

[10] R. ASKEY AND S. WAINGER, *A convolution structure for Jacobi series*, Amer. J. Math., 91 (1969), pp. 463–485.

[11] Z. DITZIAN AND V. TOTIK, *Moduli of Smoothness*, Springer Ser. Comput. Math. 9, Springer-Verlag, New York, 1987.

[12] G. FREUD, *Orthogonal Polynomials*, Akadémiai Kiadó/Pergamon Press, Budapest, 1971.

[13] ———, *Markov–Bernstein type inequalities in L_p*, in Approximation Theory II, Academic Press, San Diego, CA, 1976, pp. 369–377.

[14] ———, *On Markov–Bernstein-type inequalities and their applications*, J. Approx. Theory, 19 (1977), pp. 22–37.

[15] K. G. IVANOV, *Direct and converse theorems for the best algebraic approximation in $C[-1,1]$ and $L_p[-1,1]$*, C. R. Acad. Bulgare Sci., 33 (1980), pp. 1309–1312.

[16] ———, *Direct and converse theorems for the Best algebraic approximation in $C[-1,1]$ and $L_p[-1,1]$*, in Functions, Series and Operators, Coll. Math. Soc., János Bolyai, 35, Budapest, 1980, pp. 675–682.

[17] ———, *Some characterizations of the best algebraic approximation in $L_p[-1,1]$, $1 \le p \le \infty$*, C. R. Acad. Bulgare Sci., 34 (1981), pp. 1229–1232.

[18] B. A. KHALILOVA, *On some estimates for polynomials (Russian)*, Izv. Akad. Nauk Azerbaidzhan SSR, 2 (1974), pp. 46–55.

[19] S. V. KONJAGIN, *Bounds on the derivatives of polynomials* (*Russian*), Dokl. Akad. Nauk SSSR, 243 (1978), pp. 1116–1118, English transl.: Soviet Math. Dokl., 19 (1978), pp. 1477–1480.
[20] P. NEVAI, *Orthogonal Polynomials*, Mem. Amer. Math. Soc., 213, Providence, RI, 1979.
[21] M. K. POTAPOV, *Some inequalities for polynomials and their derivatives* (*Russian*), Vestnik Moscow Univ., 2 (1960), pp. 10–19.
[22] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., 23, American Mathematical Society, Providence, RI, 1939, fourth ed., 1975.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 571–595, March 1994

022

THE C_ℓ ROGERS–SELBERG IDENTITY*

STEPHEN C. MILNE†

Abstract. In this paper the classical Rogers–Selberg identity is extended to the setting of multiple basic hypergeometric series very well poised on symplectic C_ℓ groups. The C_ℓ Rogers–Selberg identity is deduced from the C_ℓ q-Whipple transformation. Schur functions and q-Kostka polynomials are then used to simplify the balanced "sum side" of these identities. A study of several special limiting cases then provides an elegant generalization of how the classical Rogers–Selberg identity is simplified termwise in the standard analytical proofs of the Rogers–Ramanujan identities. For C_2, explicit Rogers–Ramanujan-type identities are obtained. One of these gives a new expansion, involving q-Kostka polynomials, of the product side of one of Bressoud's (mod 6) Rogers–Ramanujan identities. It is also found that a similar analysis applied to the C_ℓ terminating ${}_6\phi_5$ summation theorem leads to a C_ℓ extension of Sylvester's identity. Special limiting C_2 and C_3 cases of this identity give new expansions of the products in the simplest classical Kac–Peterson identities. The work in this paper is motivated by the unitary A_ℓ case and the classical case corresponding to A_1 or, equivalently, to $U(2)$.

Key words. multiple basic hypergeometric series, very well poised on symplectic groups C_ℓ, C_ℓ terminating ${}_6\phi_5$ summation theorem, C_ℓ q-Whipple transformation, Rogers–Ramanujan identities, Kac–Peterson identities, Schur functions, q-Kostka polynomials

AMS subject classifications. primary 33D70, 05A19; secondary 05E05, 05A17

1. Introduction.

The purpose of this paper is to extend the classical Rogers–Selberg identity to the setting of multiple basic hypergeometric series very well poised on symplectic C_ℓ groups [Gus89], [LM91], [ML92a], [ML92b], and then to study several special limiting cases. The motivation for this work was to extend the analysis in Watson's [Wat29a] proof of the Rogers–Ramanujan identities to the C_ℓ case. This program depends on the C_ℓ q-Whipple transformation from [ML92b], which we stated in Theorem A.3 of Appendix A. Our analysis provides an elegant C_ℓ generalization of how the classical Rogers–Selberg identity is simplified termwise in the standard analytical proofs of the Rogers–Ramanujan identities. For C_2 we are able to obtain explicit Rogers–Ramanujan-type identities. One of these gives a new expansion, involving q-Kostka polynomials, of the product side of one of Bressoud's (mod 6) Rogers–Ramanujan identities [Bre80].

Our work on the C_ℓ case is motivated by the previous analysis of the unitary A_ℓ or, equivalently, the $U(\ell+1)$ case from [Mil88b], [Mil89], [Mil92]. The classical case of this work corresponds to A_1 or, equivalently, to $U(2)$. The ordinary ($q=1$) case of some of the multiple series in [Mil88a], [Mil88b], [Mil89] first appeared in certain applications of mathematical physics and the unitary groups $U(n+1)$ or, equivalently, A_n. This earlier work on the theory of Wigner coefficients for $SU(n)$ was due to Biedenharn, Holman, and Louck [BL68], [BL81a], [BL81b], [Hol80], [HBL76]. They showed in [Hol80], [HBL76] how the classical work on ordinary hypergeometric series is intimately related to the irreducible representations of the compact group $SU(2)$.

*Received by the editors September 24, 1992; accepted for publication December 10, 1992.

†Department of Mathematics, Ohio State University, Columbus, Ohio 43210 (milne@math.ohio-state.edu) . This research was partially supported by National Security Agency grant MDA 904-91-H-0055.

At this point it is useful to recall the classical Rogers–Selberg identity. Let q be a complex number such that $|q| < 1$. Define

$$(\alpha)_\infty \equiv (\alpha;q)_\infty := \prod_{k\geq 0}(1-\alpha q^k), \tag{1.1a}$$

$$(\alpha)_n \equiv (\alpha;q)_n := (\alpha)_\infty / (\alpha q^n)_\infty . \tag{1.1b}$$

For products of q-shifted factorials we use the more compact notation

$$(a_1,a_2,\dots,a_m;q)_\infty = (a_1;q)_\infty (a_2;q)_\infty \cdots (a_m;q)_\infty . \tag{1.2}$$

We then have the classical Rogers–Selberg identity [GR90, p. 37, eq. (2.7.6)] in

$$\begin{aligned} &1+\sum_{k=1}^{\infty} \frac{(aq;q)_{k-1}\,(1-aq^{2k})}{(q;q)_k}\,(-1)^k\,a^{2k}\,q^{k(5k-1)/2} \\ &\qquad = (aq;q)_\infty \sum_{k=0}^{\infty} \frac{a^k q^{k^2}}{(q;q)_k}. \end{aligned} \tag{1.3}$$

Most of the standard analytical proofs [And76], [And86], [Bai35], [GR90], [HW79], [Sla66], [Wat29a] of the Rogers–Ramanujan identities first established (1.3). Watson [Wat29a] obtains (1.3) as a special limiting case of his q-analog of Whipple's [Whi24], [Whi26] classical transformation of a very well poised ${}_7F_6(1)$ into a balanced ${}_4F_3(1)$. Setting $a=1$ or $a=q$ in (1.3), they next observe that the left side of (1.3) simplifies termwise into a theta function, which can be summed by Jacobi's [Jac29] well-known triple-product identity in Theorem A.5 of Appendix A. They then immediately obtain the classical Rogers–Ramanujan identities [And76, Chap. 7], [Rog94]

$$\sum_{k=0}^{\infty} \frac{q^{k^2}}{(q;q)_k} = \frac{1}{(q;q^5)_\infty\,(q^4;q^5)_\infty} \tag{1.4}$$

and

$$\sum_{k=0}^{\infty} \frac{q^{k^2+k}}{(q;q)_k} = \frac{1}{(q^2;q^5)_\infty\,(q^3;q^5)_\infty}, \tag{1.5}$$

respectively.

In Theorem 2.1 of §2 we obtain the C_ℓ Rogers–Selberg identity from a limiting case of the C_ℓ q-Whipple transformation in Theorem A.3. The single parameter a in (1.3) becomes the parameters $x_1,\dots,x_\ell$ in Theorem 2.1. We then appeal to the symmetric function and q-difference-equation techniques of [Mil92] to rewrite the balanced multiple sum (2.2c) on the right side of Theorem 2.1 as the sum of products of q-Kostka polynomials $K_{\mu(2^m)}(q)$ and Schur functions $s_\mu(x_1,\dots,x_\ell)$ in Corollary 2.21. This facilitates specializing the $x_1,\dots,x_\ell$ in Theorem 2.1.

The $q\mapsto q^\ell$, $x_k \mapsto q^{k-1}$ specialization of the C_ℓ Rogers–Selberg identity in §3 generalizes what happens to (1.3) when $a=1$. The resulting termwise simplification of (2.22a) is given by Lemma 3.2. Keeping in mind the standard formula for

the specialized Schur function $s_\mu(1, q, \dots, q^{\ell-1})$, we then arrive at the specialized C_ℓ Rogers–Selberg identity in Theorem 3.23. For $\ell = 2$ we are able to express (3.24a) as a double Laurent series, which is then factored into the product of two one-dimensional theta functions. Each of these is then summed by the Jacobi triple-product identity (A.6). That is, for C_2 we obtain the explicit Rogers–Ramanujan-type identity in Theorem 3.26. The sum in (3.27a) gives a new expansion, involving q-Kostka polynomials $K_{(2m-y,y)(2^m)}(q^2)$, of one of Bressoud's (mod 6) Rogers–Ramanujan identities [Bre80]. Theorem 3.26 is the C_2 generalization of the first Rogers–Ramanujan identity in (1.4). Other specializations of the $\ell = 2$ case of (2.22a), such as $\{q \mapsto q^2, x_1 \mapsto q, x_2 \mapsto q^2\}$ and $\{q \mapsto q, x_1 = x_2 = q\}$, do not factor into a single, simple infinite product. Thus (1.5) does not extend to the C_2 case. Moreover, the multiple sum (3.24a) does not appear to factor into a single, simple infinite product when $\ell \geq 3$.

We find in §4 that a limiting case of the C_ℓ terminating ${}_6\phi_5$ summation in Theorem A.1 gives the C_ℓ Sylvester identity in Theorem 4.1. The same calculation that is performed in the proof of Lemma 3.2 then specializes Theorem 4.1 into the C_ℓ Euler pentagonal-number theorem in (4.4). The $\ell = 2$ and $\ell = 3$ cases of Theorem 4.3 lead to new expansions of the infinite products $(q;q)_\infty$, $(q^2;q^2)_\infty$, and $(q;q)_\infty^2$, which appear in the simplest classical Kac–Peterson identities [Rog94], [Hec59], [KP80], [And84a], [Bre86].

In §5 we show that the $\{q \mapsto q, x_1 = x_2 = 1\}$ specialization of the $\ell = 2$ case of (2.22a) can be transformed into the one-dimensional Laurent series (5.13), which is then summed by appealing to the derivative of the quintuple-product identity (A.8). We then have the $x_1 = x_2 = 1$ C_2 Rogers–Ramanujan identity in Theorem 5.15. Just as for (3.24a), the $x_1 = \cdots = x_\ell = 1$ case of (2.22a) does not appear to factor into a single, simple infinite product when $\ell \geq 3$. However, as illustrated by Corollary 5.20, this specialization of (2.22a) should be equivalent to multiple Laurent series similar to those in [Mac72, Appendix I].

2. C_ℓ Rogers–Selberg identities. Motivated by Watson's [Wat29a] proof of the Rogers–Ramanujan identities, we begin this section with the following theorem.

THEOREM 2.1 (first version of C_ℓ Rogers–Selberg identity). *Let $x_1, \dots, x_\ell$ be indeterminate, $\ell \geq 1$, $0 < |q| < 1$, and suppose that none of the denominators in* (2.2) *vanishes. Then*

$$\begin{aligned}
&\sum_{\substack{y_k \geq 0 \\ k=1,2,\dots,\ell}} \left\{ \prod_{1 \leq r < s \leq \ell} \left[\frac{1 - \frac{x_r}{x_s} q^{y_r - y_s}}{1 - \frac{x_r}{x_s}} \, \frac{1 - x_r x_s q^{y_r + y_s}}{1 - x_r x_s} \right] \right. \\
&\qquad \times \prod_{k=1}^{\ell} \left[\frac{1 - x_k^2 q^{2y_k}}{1 - x_k^2} \right] \prod_{r,s=1}^{\ell} \left[\frac{(x_r x_s)_{y_r}}{\left(q \frac{x_r}{x_s}\right)_{y_r}} \right] \\
&\qquad \times \left[(-1)^{\ell(y_1 + \cdots + y_\ell)} \prod_{k=1}^{\ell} x_k^{(\ell+4)y_k - (y_1 + \cdots + y_\ell)} \right] \\
&\qquad \left. \times \left[q^{y_2 + 2y_3 + \cdots + (\ell-1)y_\ell} \, q^{((\ell+4)/2)(y_1^2 + \cdots + y_\ell^2) - (\ell/2)(y_1 + \cdots + y_\ell)} \right] \right\} \qquad \text{(2.2a)} \\
&\qquad = \left\{ \prod_{k=1}^{\ell} (q x_k^2)_\infty \prod_{1 \leq r < s \leq \ell} (q x_r x_s)_\infty \right\} \qquad \text{(2.2b)}
\end{aligned}$$

$$\times \sum_{\substack{m_k \geq 0 \\ k=1,2,\dots,\ell}} \left\{ \prod_{1 \leq r < s \leq \ell} \left[\frac{1 - \frac{x_r}{x_s} q^{m_r - m_s}}{1 - \frac{x_r}{x_s}} \right] \prod_{r,s=1}^{\ell} \left(q \frac{x_r}{x_s} \right)_{m_r}^{-1} \right.$$

$$\times \left[(-1)^{(\ell+1)(m_1+\cdots+m_\ell)} \prod_{k=1}^{\ell} x_k^{(\ell+2)m_k - (m_1+\cdots+m_\ell)} \right]$$

$$\times \left[q^{m_2+2m_3+\cdots+(\ell-1)m_\ell} \, q^{(\ell+2)\left[\binom{m_1}{2}+\cdots+\binom{m_\ell}{2}\right]} \right]$$

$$\left. \times \left[q^{m_1+\cdots+m_\ell} \, q^{-\binom{m_1+\cdots+m_\ell}{2}} \right] \right\}. \tag{2.2c}$$

Proof. Apply the relation $(a)_n = (-a)^n q^{\binom{n}{2}} (a^{-1}q^{1-n})_n$ to the appropriate factors in (A.4), simplify, and then take the limits $b \to 0$, $\beta \to 0$, $a \to \infty$, $\alpha \to \infty$, and $N_1 \to \infty, \dots, N_\ell \to \infty$ in Theorem A.3, while appealing to the dominated convergence theorem. To check the convergence of (2.2a), first observe that by using the product formula for a Vandermonde determinant and some algebra,

$$\prod_{1 \leq r < s \leq \ell} \left[1 - \frac{x_r}{x_s} q^{y_r - y_s} \right]$$

$$= \prod_{k=1}^{\ell} x_k^{1-k} \sum_{\sigma \in \mathcal{S}_\ell} \varepsilon(\sigma) \prod_{k=1}^{\ell} x_{\sigma(k)}^{k-1} \prod_{k=1}^{\ell} q^{(\sigma^{-1}(k)-k)y_k}. \tag{2.3}$$

Then, interchange summation and apply the multiple power-series-ratio test [AK26], [Hor89], [MS73] to each of the resulting $\ell!$ inner multiple sums.

For (2.2c) we use the comparison test and the same argument applied to the dominating multiple series determined by replacing

$$q^{(\ell+2)\left[\binom{m_1}{2}+\cdots+\binom{m_\ell}{2}\right]} \, q^{-\binom{m_1+\cdots+m_\ell}{2}} \tag{2.4a}$$

by

$$q^{(m_1^2+\cdots+m_\ell^2)-((\ell+1)/2)(m_1+\cdots+m_\ell)}. \tag{2.4b}$$

This last step depends on the identity

$$(\ell+2)\left[\binom{m_1}{2}+\cdots+\binom{m_\ell}{2}\right] - \binom{m_1+\cdots+m_\ell}{2}$$

$$= (m_1^2+\cdots+m_\ell^2) - \tfrac{(\ell+1)}{2}(m_1+\cdots+m_\ell) + \tfrac{1}{2} \sum_{1 \leq r < s \leq \ell} (m_r - m_s)^2. \qquad \square \tag{2.5}$$

Remark. The $\ell = 1$ case of (2.2) is the classical Rogers–Selberg identity in (1.3), in which $a = x_1^2$.

The analytical and combinatorial techniques in [Mil92] lead to a substantial simplification of the multiple sum in (2.2c), once we sum over the diagonals

$$m_1 + \cdots + m_\ell = m \quad \text{for} \quad m \geq 0. \tag{2.6}$$

For convenience, we work with the slightly more general diagonal sum determined by

(2.7)

$$
\begin{aligned}
&[G]_m^{(\ell)}(q;b;x_1,\dots,x_\ell)\\
&\quad:=\left[(-1)^{(\ell+1)m}\,(x_1\cdots x_\ell)^{-m}\,q^m\,q^{-\binom{m}{2}}\right]\\
&\qquad\times\sum_{\substack{m_k\geq 0\\ k=1,2,\dots,\ell\\ m_1+\cdots+m_\ell=m}}\left\{\prod_{1\leq r<s\leq\ell}\left[\frac{1-\frac{x_r}{x_s}q^{m_r-m_s}}{1-\frac{x_r}{x_s}}\right]\prod_{r,s=1}^{\ell}\left(q\frac{x_r}{x_s}\right)_{m_r}^{-1}\right.\\
&\qquad\qquad\left.\times\left[q^{m_2+2m_3+\cdots+(\ell-1)m_\ell}\,q^{(\ell+b)\left[\binom{m_1}{2}+\cdots+\binom{m_\ell}{2}\right]}\prod_{k=1}^{\ell}x_k^{(\ell+b)m_k}\right]\right\}.
\end{aligned}
$$

It is clear that (2.2c) equals

$$
\sum_{m=0}^{\infty}[G]_m^{(\ell)}(q;2;x_1,\dots,x_\ell). \tag{2.8}
$$

The first step in simplifying (2.8) is to show that $[G]_m^{(\ell)}(q;b;x_1,\dots,x_\ell)$ satisfies the q-difference equation in the following lemma.

LEMMA 2.9. *Let $[G]_m^{(\ell)}(q;b;x_1,\dots,x_\ell)$ be determined by* (2.7), *with $\ell\geq 1$ and $m\geq 1$. We then have*

$$
\begin{aligned}
&[G]_m^{(\ell)}(q;b;x_1,\dots,x_\ell) \qquad\qquad (2.10a)\\
&\quad=\sum_{p=1}^{\ell}\left[(-1)^{\ell-1}\,\frac{q}{1-q^m}\,x_p^{\ell+b-1}\prod_{\substack{r=1\\ r\neq p}}^{\ell}(x_r-x_p)^{-1}\right]\\
&\qquad\times[G]_{m-1}^{(\ell)}(q;b;x_1q^{\delta_{1,p}},\dots,x_\ell q^{\delta_{\ell,p}}), \qquad\qquad (2.10b)
\end{aligned}
$$

where $\delta_{r,s}$ is 1 if $r=s$ and is 0 otherwise.

Proof. The analysis is identical to that in [Mil92, §3]. Substitute (2.7) into (2.10b), and simplify. Group together all terms such that $(m_1,\dots,m_p+1,\dots,m_\ell)$ is the same ℓ-tuple as $(w_1,\dots,w_p,\dots,w_\ell)$ for some p. We then obtain the double sum

(2.11a)

$$
\begin{aligned}
&\left[(-1)^{(\ell+1)m}\,(x_1\cdots x_\ell)^{-m}\,q^m\,q^{-\binom{m}{2}}\,\frac{1}{1-q^m}\right]\\
&\qquad\times\sum_{\substack{w_k\geq 0\\ k=1,2,\dots,\ell\\ w_1+\cdots+w_\ell=m}}\left\{\prod_{1\leq r<s\leq\ell}\left[\frac{1-\frac{x_r}{x_s}q^{w_r-w_s}}{1-\frac{x_r}{x_s}}\right]\prod_{r,s=1}^{\ell}\left(q\frac{x_r}{x_s}\right)_{w_r}^{-1}\right.
\end{aligned}
$$

(2.11b)

$$
\left.\times\left[q^{w_2+2w_3+\cdots+(\ell-1)w_\ell}\,q^{(\ell+b)\left[\binom{w_1}{2}+\cdots+\binom{w_\ell}{2}\right]}\prod_{k=1}^{\ell}x_k^{(\ell+b)w_k}\right]\right\}
$$

(2.11c)

$$\times \sum_{p=1}^{\ell}(1-q^{w_p}) \prod_{\substack{k=1\\k\neq p}}^{\ell}\left[\frac{1-\frac{x_k}{x_p}q^{w_k}}{1-\frac{x_k}{x_p}}\right].$$

It is immediate that the sum in (2.11c) is the $y_i = q^{w_i}$ case of

$$1-y_1y_2\cdots y_\ell = \sum_{p=1}^{\ell}(1-y_p) \prod_{\substack{k=1\\k\neq p}}^{\ell}\left[\frac{x_p-x_ky_k}{x_p-x_k}\right], \tag{2.12}$$

where the $\{x_i\}$ are distinct. Direct, elementary proofs of (2.12) appear in [Mil88a, §8]. Hence the sum in (2.11c) equals

$$1-q^{w_1+\cdots+w_\ell} = 1-q^m,$$

and the proof of Lemma 2.9 is complete. □

We next use [Mil92] to write the solution of the q-difference equation (2.10) in terms of Schur functions $s_\lambda(x_1,\dots,x_\ell)$ and q-Kostka polynomials $K_{\lambda\mu}(q)$.

Let $\lambda = (\lambda_1,\lambda_2,\dots,\lambda_r,\dots)$ be a partition of nonnegative integers in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \cdots$, such that only finitely many of the λ_i are nonzero. The sum of the nonzero parts λ_i is denoted by $|\lambda|$, and the length $\ell(\lambda)$ is the number of nonzero parts of λ. If $|\lambda| = n$, we write $\lambda \vdash n$. The partition with exactly m parts, each equal to b, is (b^m). The conjugate partition to λ is denoted by λ', where $\lambda' = (\lambda_1',\lambda_2',\dots,\lambda_{(\lambda_1)}')$ and λ_i' is the number of parts λ_j in λ that are $\geq i$.

Given a partition $\lambda = (\lambda_1,\dots,\lambda_\ell)$ of length $\leq \ell$,

$$s_\lambda(x) \equiv s_\lambda(x_1,\dots,x_\ell) := \frac{\det(x_i^{\lambda_j+\ell-j})}{\det(x_i^{\ell-j})} \tag{2.13}$$

is the Schur function [Mac79] corresponding to the partition λ. (Here, $\det(a_{ij})$ denotes the determinant of an $\ell\times\ell$ matrix with (i,j)th entry a_{ij}). The Schur function $s_\lambda(x)$ is a symmetric polynomial in $x_1,\dots,x_\ell$ with nonnegative integer coefficients.

The q-Kostka polynomials $K_{\lambda\mu}(q)$ are the entries of the connection coefficient matrix $K(q)$ between Schur functions $s_\lambda(x)$ and Hall–Littlewood polynomials $P_\lambda(x;q)$. The matrix $K(q)$ is strictly upper unitriangular with rows and columns indexed by partitions in reverse lexicographic order, so that (n) comes first and (1^n) comes last. The entries $K_{\lambda\mu}(q)$ of $K(q)$ are polynomials in q with nonnegative integer coefficients. We have

$$s_\lambda(x) = \sum_{\mu\vdash|\lambda|} K_{\lambda\mu}(q)P_\mu(x;q). \tag{2.14}$$

It turns out that $K(0)$ is the identity matrix and that $K(1) \equiv K$ is the Kostka connection coefficient matrix between Schur functions and monomial symmetric functions. Many additional properties of $K(q)$ can be found in [Mac79, Chap. 3].

If (2.10) is kept in mind, it follows that the $\lambda = (b^m)$, $n = \ell$, $z_i = x_i$, $\gamma_i = 0$ case of [Mil92, Thms. 4.29 and 4.34] applied to

$$[F]_m^{(\ell)}(q;b;x_1,\dots,x_\ell) := q^{-m}\ (q)_m\ [G]_m^{(\ell)}(q;b;x_1,\dots,x_\ell), \tag{2.15}$$

leads to the following lemma.

LEMMA 2.16. *Let* $[G]_m^{(\ell)}(q;b;x_1,\dots,x_\ell)$ *be determined by* (2.7), *with* $\ell \geq 1$ *and* $m \geq 1$. *We then have*

$$[G]_m^{(\ell)}(q;b;x_1,\dots,x_\ell) = \frac{q^m}{(q)_m} \sum_{\substack{\mu \geq (b^m) \\ |\mu| = bm \\ \ell(\mu) \leq \ell}} K_{\mu(b^m)}(q)\, s_\mu(x_1,\dots,x_\ell), \tag{2.17}$$

where $\mu \geq (b^m)$ *denotes the dominance of the partial sums of the parts of* μ *over those of* (b^m).

Remark. If $\ell(\mu) > \ell$, then $s_\mu(x_1,\dots,x_\ell) = 0$. Also note that $\{K_{\mu(b^m)}(q)\}$ is a column of $K(q)$ corresponding to (b^m).

Remark. By equation (4.37b) and [Mil92, Thm. 4.34], combined with [Mac79, Ex. 1, p. 18], we have

$$[F]_m^{(\ell)}(1;b;1,\dots,1) = (h_b(1,\dots,1))^m = \binom{\ell+b-1}{b}^m, \tag{2.18}$$

where $h_b(x_1,\dots,x_\ell)$ is the bth homogeneous symmetric function of $x_1,\dots,x_\ell$.

Remark. If $\ell = 1$, then the sum in (2.17) is taken over the single partition $\mu = (bm)$. From of [Mac79, ex. 1, p. 130] we have

$$K_{(bm)(b^m)}(q) = q^{b\binom{m}{2}}, \tag{2.19}$$

and, consequently,

$$[G]_m^{(1)}(q;b;x_1) = \frac{q^{m+b\binom{m}{2}}}{(q)_m}\, x_1^{bm}. \tag{2.20}$$

Relation (2.8) and the $b = 2$ case of Lemma 2.16 now immediately give the following corollary.

COROLLARY 2.21 (second version of C_ℓ Rogers–Selberg identity). *Let* $x_1,\dots,x_\ell$ *be indeterminate,* $\ell \geq 1$, $0 < |q| < 1$, *and suppose that none of the denominators in* (2.22) *vanishes. Then*

$$\begin{aligned}
&\sum_{\substack{y_k \geq 0 \\ k=1,2,\dots,\ell}} \left\{ \prod_{1 \leq r < s \leq \ell} \left[\frac{1 - \frac{x_r}{x_s} q^{y_r - y_s}}{1 - \frac{x_r}{x_s}} \, \frac{1 - x_r x_s q^{y_r + y_s}}{1 - x_r x_s} \right] \right. \\
&\qquad \times \prod_{k=1}^{\ell} \left[\frac{1 - x_k^2 q^{2y_k}}{1 - x_k^2} \right] \prod_{r,s=1}^{\ell} \left[\frac{(x_r x_s)_{y_r}}{\left(q \frac{x_r}{x_s}\right)_{y_r}} \right] \\
&\qquad \times \left[(-1)^{\ell(y_1+\cdots+y_\ell)} \prod_{k=1}^{\ell} x_k^{(\ell+4)y_k - (y_1+\cdots+y_\ell)} \right] \\
&\qquad \left. \times \left[q^{y_2+2y_3+\cdots+(\ell-1)y_\ell}\, q^{((\ell+4)/2)(y_1^2+\cdots+y_\ell^2) - (\ell/2)(y_1+\cdots+y_\ell)} \right] \right\} \qquad \text{(2.22a)} \\
&= \left\{ \prod_{k=1}^{\ell} (q x_k^2)_\infty \prod_{1 \leq r < s \leq \ell} (q x_r x_s)_\infty \right\} \qquad \text{(2.22b)}
\end{aligned}$$

$$\times \sum_{m=0}^{\infty} \frac{q^m}{(q)_m} \sum_{\substack{\mu \geq (2^m) \\ |\mu|=2m \\ \ell(\mu)\leq \ell}} K_{\mu(2^m)}(q)\; s_\mu(x_1,\dots,x_\ell), \tag{2.22c}$$

where $\mu \geq (2^m)$ *denotes the dominance of the partial sums of the parts of* μ *over those of* (2^m).

Remark. It is clear from (2.20), with $b=2$, that the $\ell = 1$ case of (2.22) is also the classical Rogers–Selberg identity in (1.3) in which $a = x_1^2$.

For illustration purposes we conclude this section with the $m = 0,1,2,3$ cases of the inner sum in (2.22c). That is, we give $[F]_m^{(\ell)}(q;2;x)$ for $m = 0,1,2,3$. We use the notation $s_\mu(x) \equiv s_\mu(x_1,\dots,x_\ell)$ and the tables of the q-Kostka polynomials $K_{\mu\lambda}(q)$ from [Mac79, pp. 126–127]. We have the following:

$$[F]_0^{(\ell)}(q;2;x) = 1, \tag{2.23a}$$

$$[F]_1^{(\ell)}(q;2;x) = s_2(x), \tag{2.23b}$$

$$[F]_2^{(\ell)}(q;2;x) = \big\{q^2 s_4(x) + q s_{3,1}(x) + s_{2^2}(x)\big\}, \tag{2.23c}$$

$$\begin{aligned}[F]_3^{(\ell)}(q;2;x) = \big\{&q^6 s_6(x) + (q^4+q^5)s_{5,1}(x) \\ &+ (q^2+q^3+q^4)s_{4,2}(x) + q^3 s_{4,1^2}(x) \\ &+ q^3 s_{3^2}(x) + (q+q^2)s_{3,2,1}(x) + s_{2^3}(x)\big\}.\end{aligned} \tag{2.23d}$$

By [Mil92, Thm. 4.34], $[F]_m^{(\ell)}(q;b;x_1,\dots,x_\ell)$ is the $\lambda = (b^m)$ case of $H_\lambda(x;\emptyset,q)$, which is a natural q-analog of the complete homogeneous symmetric function $h_\lambda(x)$.

3. The specialized C_ℓ Rogers–Selberg identity. Motivated by [Mil92, Thms. 1.19 and 1.21] we rewrite the

$$q \mapsto q^\ell \quad \text{and} \quad x_k \mapsto q^{k-1} \quad \text{for } k = 1,2,\dots,\ell \tag{3.1}$$

case of Corollary 2.21. This generalizes a key simplification of the Rogers–Selberg identity in the standard analytical proof of the Rogers–Ramanujan identities. We first simplify (2.22a) and then deal with (2.22b) and (2.22c).

We begin with the following lemma.

LEMMA 3.2 (C_ℓ Rogers–Selberg product-side simplification). *Let* $0 < |q| < 1$ *and* $\ell \geq 1$. *Then the* (3.1) *case of* (2.22a) *can be termwise rewritten as*

$$\begin{aligned}\sum_{\substack{y_2,\dots,y_\ell \geq 0 \\ -\infty < y_1 < \infty}} \Bigg\{ &\prod_{k=1}^{\ell-1} (q;q)_{2k}^{-1} \prod_{k=2}^{\ell} (1+q^{k-1+\ell y_k}) \\ &\times \prod_{1\leq r<s\leq \ell} \left[(q^{\ell y_r} - q^{s-r+\ell y_s})\,(1-q^{r+s-2+\ell y_r+\ell y_s})\right] \\ &\times \left[(-1)^{\ell(y_1+\cdots+y_\ell)}\, q^{(\ell+4)(y_2+2y_3+\cdots+(\ell-1)\,y_\ell)}\right] \\ &\times \left[q^{(\ell(\ell+4)/2)(y_1^2+\cdots+y_\ell^2)}\, q^{(\frac{\ell}{2}-\ell^2)(y_1+\cdots+y_\ell)+\ell y_1}\right]\Bigg\}.\end{aligned} \tag{3.3}$$

Proof. We first find that the (3.1) case of the $y_1, \dots, y_\ell > 0$ term in (2.22a) simplifies to

$$\begin{aligned}(3.4)\qquad &\Bigg\{\prod_{k=1}^{\ell-1}(q;q)_{2k}^{-1}\prod_{k=1}^{\ell}(1+q^{k-1+\ell y_k})\\ &\times\prod_{1\le r<s\le\ell}\left[(q^{\ell y_r}-q^{s-r+\ell y_s})\,(1-q^{r+s-2+\ell y_r+\ell y_s})\right]\\ &\times\left[(-1)^{\ell(y_1+\cdots+y_\ell)}\,q^{(\ell+4)(y_2+2y_3+\cdots+(\ell-1)\,y_\ell)}\right]\\ &\times\left[q^{(\ell(\ell+4)/2)(y_1^2+\cdots+y_\ell^2)}\,q^{(\frac{\ell}{2}-\ell^2)(y_1+\cdots+y_\ell)}\right]\Bigg\}.\end{aligned}$$

Before applying (3.1), we used

$$(3.5)\qquad \frac{(x_1^2)_{y_1}}{(1-x_1^2)}=(qx_1^2)_{y_1-1}$$

and noted that

$$(3.6)\qquad q^{y_2+2y_3+\cdots+(\ell-1)\,y_\ell}\prod_{1\le r<s\le\ell}\left[\frac{1-\frac{x_r}{x_s}q^{y_r-y_s}}{1-\frac{x_r}{x_s}}\right]=\prod_{1\le r<s\le\ell}\left[\frac{q^{y_r}-\frac{x_s}{x_r}q^{y_s}}{1-\frac{x_s}{x_r}}\right].$$

We then used, in the following order, the relations

$$(3.7)\qquad \prod_{1\le r<s\le\ell}(1-q^{s-r})^{-1}=\prod_{k=1}^{\ell-1}(q;q)_k^{-1};$$

$$(3.8)\qquad \prod_{1\le r<s\le\ell}(1-q^{r+s-2})^{-1}=\prod_{k=1}^{\ell-1}(q^k;q)_k^{-1};$$

$$(3.9)\qquad \prod_{r,s=1}^{\ell}(q^{\ell+r-s};q^\ell)_{y_r}^{-1}=\prod_{k=1}^{\ell}(q^k;q)_{\ell y_k}^{-1};$$

$$(3.10a)\qquad (q^\ell;q^\ell)_{y_1-1}\prod_{s=2}^{\ell}(q^{s-1};q^\ell)_{y_1}=(q;q)_{\ell y_1-1};$$

$$(3.10b)\qquad \prod_{s=1}^{\ell}(q^{k+s-2};q^\ell)_{y_k}=(q^{k-1};q)_{\ell y_k}\quad\text{if } k=2,3,\dots,\ell;$$

$$(3.11)\qquad \prod_{k=1}^{\ell}(1-q^{2k-2+2\ell y_k})=\prod_{k=1}^{\ell}\left[(1+q^{k-1+\ell y_k})\,(1-q^{k-1+\ell y_k})\right].$$

At this point a crucial simplification was provided by

$$(3.12a)\qquad \left[(1-q^{\ell y_1})\,(q;q)_{\ell y_1-1}\right]/(q;q)_{\ell y_1}=1$$

and

$$\left[(1-q^{k-1+\ell y_k})\,(q^{k-1};q)_{\ell y_k}\right] / (q^k;q)_{\ell y_k} = (1-q^{k-1}) \tag{3.12b}$$

if $k = 2, 3, \ldots, \ell$. The classical case of (3.12) just amounts to setting $\ell = 1$ in (3.12a).

Finally, we arrived at (3.4) by observing that

$$\frac{(1-q^k)}{(q;q)_k\,(q^k;q)_k\,(1-q^{2k})} = (q;q)_{2k}^{-1} \tag{3.13}$$

for $k = 1, 2, \ldots, \ell - 1$.

A similar calculation shows that (3.4) also holds for $y_2, \ldots, y_\ell \geq 0$ and $y_1 > 0$.

If $y_1 = 0$, we do not need (3.5). In this case, we obtain (3.4) multiplied by $\frac{1}{2}$.

Let $(1 + q^{\ell y_1})T(y_1, \ldots, y_\ell)$ be given by (3.4). Then the (3.1) case of (2.22a) becomes

$$\sum_{y_2,\ldots,y_\ell \geq 0} T(0, y_2, \ldots, y_\ell) + \sum_{\substack{y_2,\ldots,y_\ell \geq 0 \\ y_1 > 0}} (1 + q^{\ell y_1})\, T(y_1, \ldots, y_\ell). \tag{3.14}$$

It is not hard to see that

$$q^{-\ell y_1} T(-y_1, y_2, \ldots, y_\ell) = T(y_1, \ldots, y_\ell). \tag{3.15}$$

Thus (3.14) equals

$$\sum_{\substack{y_2,\ldots,y_\ell \geq 0 \\ -\infty < y_1 < \infty}} q^{\ell y_1} T(y_1, y_2, \ldots, y_\ell), \tag{3.16}$$

and the proof is complete. □

We now consider the products in (2.22b). An elementary calculation shows that

$$\prod_{k=1}^{\ell} (q^{\ell+2k-2};q^\ell)_\infty \prod_{1 \leq r < s \leq \ell} (q^{\ell+r+s-2};q^\ell)_\infty \tag{3.17a}$$

$$= \begin{cases} \displaystyle\prod_{k=m}^{2m} (q^{1+2k};q)_\infty & \text{for } \ell = 1+2m,\ m \geq 0, \\ \displaystyle (q^{2m};q^2)_\infty \prod_{k=m}^{2m-1} (q^{1+2k};q)_\infty & \text{for } \ell = 2m,\ m \geq 1. \end{cases} \tag{3.17b}$$

To simplify the (3.1) case of (2.22c) we need an explicit formula for the specialized Schur function $s_\lambda(1, q, \ldots, q^{\ell-1})$, which appears in [Mac79], [Sta71].

Consider the Ferrers diagram of λ in which the rows and columns are arranged as in a matrix with the ith row consisting of λ_i cells. For a given cell $x = (i,j) \in \lambda$ we define the hook length $h(x)$ and content $c(x)$ as follows:

$$h(x) \equiv h(i,j) := (\lambda_i - i) + (\lambda'_j - j) + 1, \tag{3.18}$$

where λ' is the partition conjugate to λ and

$$c(x) := j - i. \tag{3.19}$$

Note that $h(x)$ is the number of cells to the right of, on, or below the cell in the (i,j) position and that $c(x)$ measures how far the cell (i,j) is from the main diagonal.

Given (3.18) and (3.19), we have

$$s_\lambda(1,q,\dots,q^{\ell-1}) = q^{n(\lambda)} \prod_{x\in\lambda} \frac{(1-q^{\ell+c(x)})}{(1-q^{h(x)})}, \tag{3.20}$$

where

$$n(\lambda) = \sum_{i\geq 1}(i-1)\lambda_i. \tag{3.21}$$

We are now ready to rewrite the (3.1) case of Corollary 2.21, in which we have multiplied both sides by

$$\prod_{k=1}^{\ell-1} (q;q)_{2k}\,. \tag{3.22}$$

Lemma 3.2 and equations (3.17) and (3.20) immediately imply that (2.22) is transformed into the following theorem.

THEOREM 3.23 (specialized C_ℓ Rogers–Selberg identity). *Let* $0<|q|<1$ *and* $\ell\geq 1$. *We then have*

$$\begin{aligned}
&\sum_{\substack{y_2,\dots,y_\ell\geq 0\\ -\infty<y_1<\infty}} \Bigg\{ \prod_{k=2}^{\ell}(1+q^{k-1+\ell y_k}) \\
&\qquad \times \prod_{1\leq r<s\leq \ell} \left[(q^{\ell y_r}-q^{s-r+\ell y_s})\,(1-q^{r+s-2+\ell y_r+\ell y_s})\right] \\
&\qquad \times \left[(-1)^{\ell(y_1+\cdots+y_\ell)}\, q^{(\ell+4)(y_2+2y_3+\cdots+(\ell-1)\,y_\ell)}\right] \\
&\qquad \times \left[q^{(\ell(\ell+4)/2)(y_1^2+\cdots+y_\ell^2)}\, q^{(\frac{\ell}{2}-\ell^2)(y_1+\cdots+y_\ell)+\ell y_1}\right]\Bigg\}. \qquad \text{(3.24a)} \\
&\quad = P(\ell;q) \sum_{m=0}^{\infty} \frac{q^{\ell m}}{(q^\ell;q^\ell)_m} \sum_{\substack{\mu\geq(2^m)\\ |\mu|=2m\\ \ell(\mu)\leq\ell}} q^{n(\mu)}\, K_{\mu(2^m)}(q^\ell) \\
&\qquad\qquad \times \prod_{x\in\mu} \frac{(1-q^{\ell+c(x)})}{(1-q^{h(x)})}, \qquad \text{(3.24b)}
\end{aligned}$$

where

$$P(\ell;q) = \begin{cases} (q;q)_\infty^{m+1} \displaystyle\prod_{k=1}^{m-1} (q;q)_{2k} & \text{for } \ell = 1+2m,\ m\geq 0, \\ (q^2;q^2)_\infty\ (q;q)_\infty^m \displaystyle\prod_{k=1}^{m-1} (q;q)_{2k-1} & \text{for } \ell = 2m,\ m\geq 1. \end{cases} \tag{3.25}$$

Remark. The products in (3.25) are the product side of the specialized C_ℓ Euler pentagonal number theorem 4.3 in §4.

Remark. The $\ell=1$ case of (3.24a) is [GR90, eq. (2.7.7), p. 37], which is summed by the Jacobi triple-product identity in (A.6) in Appendix A to complete one of the standard proofs of the first Rogers–Ramanujan identity in (1.4).

The $\ell=2$ case of (3.24) leads to the following theorem.

THEOREM 3.26 ($q \mapsto q^2$, $x_1 \mapsto 1$, $x_2 \mapsto q$ C_2 Rogers–Ramanujan identity). *Let* $0 < |q| < 1$. *We then have*

$$1 + \sum_{m=1}^{\infty} \frac{q^{2m}}{(q^2;q^2)_m} \sum_{y=0}^{m} \frac{q^y(1-q^{2m-2y+1})}{(1-q)} K_{(2m-y,y)(2^m)}(q^2) \tag{3.27a}$$

$$= \prod_{\substack{n=1 \\ n\not\equiv 0,\pm 1 \pmod 6}}^{\infty} (1-q^n)^{-1}. \tag{3.27b}$$

Proof. The sum in (3.27a) is immediate from the $\ell = 2$ case of (3.24b) once we note that the set of $m+1$ partitions in the inner sum of (3.24b) is given by

$$\{(2m,0),(2m-1,1),\ldots,(m,m)\}. \tag{3.28}$$

The $m = 1$ case of (3.28) works since $K_{(1^2)(2)}(q) = 0$.

The product in (3.27b) is a consequence of

$$(-q^3,-q^9,q^{12};q^{12})_\infty = \frac{(q^6;q^6)_\infty}{(q^3;q^6)_\infty} \tag{3.29}$$

and the simplification in the following lemma.

LEMMA 3.30. *The* $\ell = 2$ *case of* (3.24a) *can be factored into the infinite product*

$$(-q^3,-q^9,q^{12};q^{12})_\infty\,(q;q)_\infty\,. \tag{3.31}$$

Proof. Write the $\ell = 2$ case of (3.24a) as

$$\sum_{\substack{y_2\geq 0 \\ -\infty<y_1<\infty}} (1+q^{1+2y_2})S(y_1,y_2) \tag{3.32}$$

$$= \sum_{\substack{y_2>0 \\ -\infty<y_1<\infty}} S(y_1,y_2) + \sum_{\substack{y_2>0 \\ -\infty<y_1<\infty}} q^{1+2y_2}S(y_1,y_2) + \sum_{-\infty<y_1<\infty} (1+q)S(y_1,0). \tag{3.33}$$

It is not hard to see that

$$S(y_1,-y_2-1) = q^{1+2y_2}S(y_1,y_2) \quad \text{for } y_2 \geq 0. \tag{3.34}$$

Thus, applying (3.34) to (3.33) gives

$$\sum_{-\infty<y_1,y_2<\infty} S(y_1,y_2),$$

which, in turn, is equal to

$$\sum_{-\infty<y_1,y_2<\infty} q^{6y_1^2+6y_2^2+y_1+3y_2} - \sum_{-\infty<y_1,y_2<\infty} q^{6y_1^2+6y_2^2+3y_1+5y_2+1} \tag{3.35a}$$

$$+ \sum_{-\infty<y_1,y_2<\infty} q^{6y_1^2+6y_2^2+y_1+7y_2+2} - \sum_{-\infty<y_1,y_2<\infty} q^{6y_1^2+6y_2^2-y_1+5y_2+1}. \tag{3.35b}$$

The two double Laurent series in (3.35b) cancel termwise under the transformation

$$y_1 \mapsto -y_1 \quad \text{and} \quad y_2 \mapsto -y_2 - 1,$$

leaving us with (3.35a). Interchanging y_1 and y_2 in the first series in (3.35a) enables us to factor (3.35a) into the product

$$\left\{ \sum_{y_1=-\infty}^{\infty} q^{6y_1^2+3y_1} \right\} \left\{ \sum_{y_2=-\infty}^{\infty} q^{6y_2^2+y_2}(1-q^{1+4y_2}) \right\}. \tag{3.36}$$

The first factor in (3.36) can be summed by the $q \mapsto q^{12}$ and $z \mapsto -q^3$ case of the Jacobi triple product identity in (A.6) in Appendix A to give

$$(-q^3, -q^9, q^{12}; q^{12})_\infty . \tag{3.37}$$

The second factor in (3.36) is just the sum side of the Euler pentagonal number theorem in which the terms are grouped into even or odd index of summation. That is, the $q \mapsto q^3$ and $z \mapsto q^{1/2}$ case of (A.6) implies that the second factor in (3.36) is

$$(q;q)_\infty . \tag{3.38}$$

Gordon [Gor61] has observed that the second factor in (3.36) can also be summed directly by the $q \mapsto q^4$ and $z \mapsto -q$ case of the quintuple-product identity in (A.8) of Appendix A.

Multiplying (3.37) and (3.38) completes the proof. □

The proof of Theorem 3.26 is also complete. □

The product in (3.27b) also appears in the $r = 1$ and $k = 3$ case of Bressoud's multisum Rogers–Ramanujan identity in [Bre80, eq. (3.4)]. This case of Bressoud's identity is

$$\prod_{\substack{n=1 \\ n\not\equiv 0,\pm 1 \pmod 6}}^{\infty} (1-q^n)^{-1} = \sum_{m_1,m_2 \geq 0} \frac{q^{(m_1+m_2)^2+m_2^2+m_1+2m_2}}{(q;q)_{m_1}\,(q^2;q^2)_{m_2}}. \tag{3.39}$$

Clearly, the sums in (3.27a) and (3.39) are equal. It is also possible that the mth diagonal sum in (3.39) over $m_1 + m_2 = m$ equals the inner sum in (3.27a). We have verified this equality for $m = 1, 2, 3$. If true, it appears to be nontrivial. For future reference we note the following conjecture.

CONJECTURE 3.40. *Let m be any positive integer. We then have*

$$q^{2m^2} \sum_{y=0}^{m} q^{-my+\binom{y}{2}}\, (-q^{-m};q)_y \begin{bmatrix} m \\ y \end{bmatrix}_q \tag{3.41a}$$

$$= \sum_{y=0}^{m} \frac{q^y(1-q^{2m-2y+1})}{(1-q)}\, K_{(2m-y,y)(2^m)}(q^2), \tag{3.41b}$$

where $\begin{bmatrix} m \\ y \end{bmatrix}_q$ is the polynomial in q known as the q-binomial coefficient

$$\begin{bmatrix} m \\ y \end{bmatrix}_q = \frac{(q;q)_m}{(q;q)_y\,(q;q)_{m-y}}. \tag{3.42}$$

Equation (3.41) is quite striking in view of Kirillov's formula [Kir88, Thm. 3.1] for the general q-Kostka polynomial $K_{\lambda\setminus\rho,\mu|\eta}(q)$ as a multiple sum of products of q-binomial coefficients. A special case of Kirillov's formula gives $K_{(2m-y,y)(2^m)}(q^2)$ as a sum over the set of partitions $\{\lambda\,|\lambda\vdash y,\ 2\lambda_1\le m\}$, where the term corresponding to λ is a suitable power of q times a product of $\ell(\lambda)$ q-binomial coefficients, each in base q^2.

It is not hard to see that the limit as $q\to 1^-$ of both sides of (3.41) is 3^m.

Bressoud also provided an elegant combinatorial interpretation of his general Rogers–Ramanujan identity in [Bre80, eq. (3.4)]. The special case corresponding to (3.39) is as follows:

For all positive integers n, the partitions of n into parts not congruent to $0,\pm1$ (mod 6) on the product side are equinumerous with the partitions of n into parts, say, $b_1+\cdots+b_s$, such that $b_i\ge b_{i+1}$, $b_i-b_{i+2}\ge 2$; if $b_i-b_{i+1}\le 1$, then b_i+b_{i+1} is even and all parts are ≥ 2 on the sum side.

The same combinatorial interpretation holds for (3.27).

Just as in the classical A_1 case [And84b] it should be possible to iterate the C_2 Bailey lemma in [ML92a] and embed Theorem 3.26 into an infinite family of such identities.

For $\ell\ge 3$ it is not possible to write (3.24a) as an ℓ-dimensional Laurent series. After multiplying out

$$\prod_{k=2}^{\ell}(1+q^{k-1+\ell y_k}),$$

we do have the termwise symmetry determined by

$$y_k\mapsto -y_k-\frac{2(k-1)}{\ell}\quad\text{for } k=1,2,\dots,\ell \tag{3.43}$$

in (3.24a), but (3.43) gives all integer shifts only for $\ell=1$ and 2.

The multiple sum (3.24a) does not appear to factor into a single, simple infinite product when $\ell\ge 3$. We have found, for $\ell=3$ and 4, that Euler's infinite-product-representation algorithm (EIPRA) [And86, p. 104] applied to the power series in q determined by (3.24a) yields an infinite product

$$\prod_{n=1}^{\infty}(1-q^n)^{-a_n} \tag{3.44}$$

in which the a_n's have no simple pattern and $|a_n|\to\infty$. Any power series with constant term 1 and integer coefficients can be uniquely expressed by (3.44). It is still possible that (3.24a) can by written as a finite linear combination of simple infinite products when $\ell\ge 3$.

A further combinatorial examination of (3.24a) should be facilitated by applying the $x_k=-q^{k-1+\ell y_k}$ case of the Weyl denominator formula [Bre87], [Mac79, pp. 46–47] for the root system B_ℓ to expand the product

$$\prod_{k=1}^{\ell}(1+q^{k-1+\ell y_k})\prod_{1\le r<s\le\ell}\left[(q^{\ell y_r}-q^{s-r+\ell y_s})\,(1-q^{r+s-2+\ell y_r+\ell y_s})\right] \tag{3.45}$$

into an alternating sum of $2^\ell\ell!$ monomials.

4. The C_ℓ Euler pentagonal number theorem. In this section we study C_ℓ generalizations of the classical Euler pentagonal number theorem. We also consider the C_2 cases of these general identities.

In the same way that we derived Theorem 2.1 from Theorem A.3 in Appendix A we find that taking the limit $b \to 0$, $a \to \infty$, and $N_1 \to \infty, \dots, N_\ell \to \infty$ in Theorem A.1 leads to the following theorem.

THEOREM 4.1 (the unspecialized C_ℓ Sylvester's identity). *Let $x_1, \dots, x_\ell$ be indeterminate, $\ell \geq 1$, $0 < |q| < 1$, and suppose that none of the denominators in* (4.2) *vanishes. Then*

$$\begin{aligned}\sum_{\substack{y_k \geq 0 \\ k=1,2,\dots,\ell}} \Bigg\{ & \prod_{1 \leq r < s \leq \ell} \left[\frac{1 - \frac{x_r}{x_s} q^{y_r - y_s}}{1 - \frac{x_r}{x_s}} \, \frac{1 - x_r x_s q^{y_r + y_s}}{1 - x_r x_s} \right] \\ & \times \prod_{k=1}^{\ell} \left[\frac{1 - x_k^2 q^{2y_k}}{1 - x_k^2} \right] \prod_{r,s=1}^{\ell} \left[\frac{(x_r x_s)_{y_r}}{\left(q \frac{x_r}{x_s}\right)_{y_r}} \right] \\ & \times \left[(-1)^{\ell(y_1 + \cdots + y_\ell)} \prod_{k=1}^{\ell} x_k^{(\ell+2)y_k - (y_1 + \cdots + y_\ell)} \right] \\ & \times \left[q^{y_2 + 2y_3 + \cdots + (\ell-1)\, y_\ell} \, q^{((\ell+2)/2)(y_1^2 + \cdots + y_\ell^2) - (\ell/2)(y_1 + \cdots + y_\ell)} \right] \Bigg\} \end{aligned} \tag{4.2a}$$

$$= \prod_{k=1}^{\ell} (q x_k^2)_\infty \prod_{1 \leq r < s \leq \ell} (q x_r x_s)_\infty . \tag{4.2b}$$

Remark. The $\ell = 1$ case of (4.2) is Sylvester's identity in [And76, Thm. 9.2, p. 140] in which we take $x = x_1^2 q^{-1}$.

The same calculation that was performed in the proof of Lemma 3.2, followed by (3.17), enables us to transform the (3.1) case of (4.2) into the following lemma.

THEOREM 4.3 (first C_ℓ Euler pentagonal number theorem). *Let $0 < |q| < 1$ and $\ell \geq 1$. We then have*

$$\begin{aligned}\sum_{\substack{y_2, \dots, y_\ell \geq 0 \\ -\infty < y_1 < \infty}} \Bigg\{ & \prod_{k=2}^{\ell} (1 + q^{k-1+\ell y_k}) \\ & \times \prod_{1 \leq r < s \leq \ell} \left[(q^{\ell y_r} - q^{s-r+\ell y_s}) \, (1 - q^{r+s-2+\ell y_r + \ell y_s}) \right] \\ & \times \left[(-1)^{\ell(y_1 + \cdots + y_\ell)} \, q^{(\ell+2)(y_2 + 2y_3 + \cdots + (\ell-1)\, y_\ell)} \right] \\ & \times \left[q^{(\ell(\ell+2)/2)(y_1^2 + \cdots + y_\ell^2)} \, q^{(\frac{\ell}{2} - \ell^2)(y_1 + \cdots + y_\ell) + \ell y_1} \right] \Bigg\}. \end{aligned} \tag{4.4a}$$

$$= \begin{cases} (q;q)_\infty^{m+1} \prod_{k=1}^{m-1} (q;q)_{2k} & \text{for } \ell = 1 + 2m, \ m \geq 0, \\ (q^2;q^2)_\infty \, (q;q)_\infty^m \prod_{k=1}^{m-1} (q;q)_{2k-1} & \text{for } \ell = 2m, \ m \geq 1. \end{cases} \tag{4.4b}$$

Remark. The $\ell = 1$ case of (4.4) is Euler's pentagonal number theorem in [And76, Cor. 1.7, p. 11].

Remark. The B_ℓ Weyl denominator formula can also be used to expand the products in the sum in (4.4a).

For $\ell = 2$ we are able to write (4.4a) as an alternating sum of four double Laurent series. Two of these series cancel termwise. This is the same analysis that was performed in the proof of Lemma 3.30 that expresses (3.24a) as the difference of two double Laurent series. Keeping in mind the $\ell = 2$ case of (4.4b), we then obtain the following corollary.

COROLLARY 4.5 (second C_2 Euler pentagonal number theorem). *Let* $0 < |q| < 1$. *We then have*

$$(q;q)_\infty (q^2;q^2)_\infty = \sum_{-\infty<y_1,y_2<\infty} q^{4y_1^2+4y_2^2+y_1+y_2} - q \sum_{-\infty<y_1,y_2<\infty} q^{4y_1^2+4y_2^2+3y_1+3y_2}. \tag{4.6}$$

Each double Laurent series in (4.6) is the square of a theta function, which can be summed by the Jacobi triple-product identity (A.6) in Appendix A. It is then clear that (4.6) becomes the following corollary.

COROLLARY 4.7. *Let* $0 < |q| < 1$. *Then*

$$(q,q^2,q^3,q^4,q^5,q^6,q^7;q^8)_\infty (q^2,q^4,q^6;q^8)_\infty \tag{4.8a}$$

$$= (-q^3,-q^5;q^8)_\infty^2 - q(-q,-q^7;q^8)_\infty^2. \tag{4.8b}$$

The product side of Corollary 4.5 also appears in the Kac–Peterson-type identities

$$(q;q)_\infty (q^2;q^2)_\infty = \sum_{\substack{y_1,y_2=-\infty \\ y_2\geq 3|y_1|}}^{\infty} (-1)^{y_2} q^{(y_2^2-8y_1^2+y_2)/2}, \tag{4.9a}$$

$$(q;q)_\infty (q^2;q^2)_\infty = \sum_{\substack{y_1,y_2=-\infty \\ y_2\geq 2|y_1|}}^{\infty} (-1)^{y_1+y_2} q^{(y_2^2-2y_1^2+y_2)/2}, \tag{4.9b}$$

$$(q;q)_\infty (q^2;q^2)_\infty = \sum_{\substack{y_1,y_2=-\infty \\ y_2\geq |y_1|}}^{\infty} (-1)^{y_2} q^{(2y_2^2-y_1^2+2y_2+y_1)/2}. \tag{4.9c}$$

Equation (4.9a) is the original Kac–Peterson identity in [KP80, final equation]. Andrews [And84a] derived (4.9a) and (4.9b) by computing the constant term in the series expansion of a certain infinite product by two very different methods. Bressoud [Bre86] used q-Hermite polynomial expansion techniques of L. J. Rogers to prove (4.9b) and (4.9c). The sums in (4.9a) and (4.9b) are termwise transformed into each other in [And84a, §4]. This is useful since (4.9b) is much easier to prove than is (4.9a).

We give much simpler termwise transformations between the sums in (4.9a) and (4.9b), and those in (4.9b) and (4.9c).

We first transform (4.9b) into (4.9a) by considering a suitable mapping T_1 of one index of summation set into the other.

If $y_1 \geq 0$ is even, then T_1 maps the vertical half-line $\{(y_1, \alpha+2y_1) \,|\alpha \geq 0\}$ onto the vertical half-line $\{(y_1/2, \alpha+2y_1) \,|\alpha \geq 0\}$, and if $y_1 > 0$ is odd, then T_1 maps the vertical half-line $\{(y_1, \alpha+2y_1) \,|\alpha \geq 0\}$ onto the half-line $\{(\alpha+(1+y_1)/2, 3\alpha+1+2y_1) \,|\alpha \geq 0\}$ of slope 3. Moreover, T_1 preserves symmetry with respect to the y_2-axis.

It is not hard to see that T_1 maps $\{(y_1,y_2)\,|y_2 \geq 2|y_1|\}$ one-to-one onto $\{(y_1,y_2)\,|y_2 \geq 3|y_1|\}$.

An explicit formula for T_1 is given by

$$T_1(y_1,y_2) := \begin{cases} \left(\frac{y_1}{2}, y_2\right) & \text{if } y_1 \geq 0 \text{ is even,} \\ \left(y_2 - \frac{3}{2}y_1 + \frac{1}{2}, 3y_2 - 4y_1 + 1\right) & \text{if } y_1 > 0 \text{ is odd,} \end{cases} \tag{4.10a}$$

and

$$T_1(-y_1,y_2) := (-w_1,w_2) \quad \text{if} \quad T_1(y_1,y_2) = (w_1,w_2). \tag{4.10b}$$

By checking the same cases and symmetry needed to define T_1, it is not hard to see that

$$F_2(y_1,y_2) = F_3(T_1(y_1,y_2)), \tag{4.11}$$

where the general terms in the sums in (4.9b) and (4.9a) are denoted by $F_2(y_1,y_2)$ and $F_3(y_1,y_2)$, respectively. That is, the sums in (4.9b) and (4.9a) are termwise equivalent.

We next transform (4.9b) into (4.9c). If $y_1 > 0$, let T_2 map the vertical half-line $\{(y_1,\alpha+2y_1)\,|\alpha \geq 0\}$ onto the half-line $\{(\alpha+1,\alpha+y_1)\,|\alpha \geq 0\}$, and if $y_1 \leq 0$, let T_2 map the vertical half-line $\{(y_1,\alpha-2y_1)\,|\alpha \geq 0\}$ onto the half-line $\{(-\alpha,\alpha-y_1)\,|\alpha \geq 0\}$.

It is clear that T_2 maps $\{(y_1,y_2)\,|y_2 \geq 2|y_1|\}$ one-to-one onto $\{(y_1,y_2)\,|y_2 \geq |y_1|\}$.

The transformation T_2 is given explicitly by

$$T_2(y_1,y_2) := \begin{cases} (1+y_2-2y_1, y_2-y_1) & \text{if } y_1 > 0, \\ (-2y_1-y_2, y_1+y_2) & \text{if } y_1 \leq 0. \end{cases} \tag{4.12}$$

Just as for T_1, it easily follows that

$$F_2(y_1,y_2) = F_1(T_2(y_1,y_2)), \tag{4.13}$$

where the general terms in the sums in (4.9b) and (4.9c) are $F_2(y_1,y_2)$ and $F_1(y_1,y_2)$, respectively. Thus the sums in (4.9b) and (4.9c) are termwise equivalent.

The collection of all the terms in the sum side of (4.6), before any cancellation, is not a subset of all of the terms in any of (4.9a)–(4.9c), and vice versa. This is quite a contrast to the fact that the sums in (4.9a)–(4.9c) are termwise equivalent to one another. The equality of the sum side of (4.6) with the sums in (4.9) is not trivial.

The $\ell = 3$ case of Theorem 4.3 gives an expansion of $(q;q)_\infty^2$ that is not termwise equivalent to the sum in the identity

$$(q;q)_\infty^2 = \sum_{\substack{y_1,y_2=-\infty \\ y_2 \geq 2|y_1|}}^{\infty} (-1)^{y_1+y_2} q^{(y_2^2-3y_1^2+y_2+y_1)/2}. \tag{4.14}$$

Equation (4.14) was first stated and proved by Rogers in [Rog94, p. 323], and was subsequently re-proved in different ways by Hecke [Hec59, eq. (7), p. 425], Kac and Peterson [KP80], Andrews [And84a], and Bressoud [Bre86].

5. The $x_1 = x_2 = 1$ C_2 Rogers–Ramanujan identity. In this section we consider the $x_1 = x_2 = 1$ and $\ell = 2$ specialization of Corollary 2.21 and Theorem 4.1 and the $x_1 = x_2 = x_3 = 1$ and $\ell = 3$ specialization of Corollary 2.21.

It is useful to first observe that the limit as $x_1, x_2 \to 1$ of

$$q^{y_2}(1 - x_1x_2)[1 - (x_1/x_2)q^{y_1-y_2}]/(1 - (x_1/x_2))$$

is $(q^{y_1} - q^{y_2})$, provided that $y_1, y_2 > 0$ and $y_1 \neq y_2$. It is then not hard to see that we have the following lemma.

LEMMA 5.1. *Let $R(y_1, \dots, y_\ell)$ be the general term in either* (2.22a) *or* (4.2a), *where $\ell \geq 2$. If $y_1, y_2 > 0$, let*

$$L(y_1, y_2) = \lim_{x_1, x_2 \to 1} R(y_1, \dots, y_\ell), \tag{5.2}$$

where the notation $L(y_1, y_2)$ suppresses dependence on $y_3, \dots, y_\ell$, as well as $x_3, \dots, x_\ell$. Then $L(y_1, y_2)$ exists, $L(y_1, y_1) = 0$, and

$$L(y_1, y_2) = (q^{y_1} - q^{y_2})Q(y_1, y_2) \quad \text{if} \quad y_1 \neq y_2, \tag{5.3}$$

where $Q(y_1, y_2)$ is symmetric in y_1 and y_2. Thus we also have

$$L(y_1, y_2) + L(y_2, y_1) = 0 \quad \text{if} \quad y_1 \neq y_2. \tag{5.4}$$

It is clear from Lemma 5.1 that if $R(y_1, \dots, y_\ell)$ is the general term in either (2.22a) or (4.2a), with $\ell \geq 2$, then

$$\lim_{\substack{x_k \to 1 \\ k=1,2,\dots,\ell}} \sum_{\substack{y_k \geq 0 \\ k=1,2,\dots,\ell}} R(y_1, \dots, y_\ell) = \lim_{\substack{x_k \to 1 \\ k=1,2,\dots,\ell}} \sum_{\substack{y_1 \text{ or } y_2 = 0 \\ y_3, \dots, y_\ell \geq 0}} R(y_1, \dots, y_\ell). \tag{5.5}$$

As the first application of (5.5) we have the following lemma.

LEMMA 5.6. *The $x_1 = x_2 = 1$ and $\ell = 2$ case of* (2.22a) *can be factored into the infinite product*

$$(q^2; q^2)_\infty^3 \, (q^2; q^4)_\infty^2 . \tag{5.7}$$

Proof. By (5.5) we have to consider only the terms $R(y_1, y_2)$ of (2.22a) in which $y_1 = y_2 = 0$, $y_1 = 0$, or $y_2 = 0$. That is, we can start with

$$1 + \sum_{y \geq 1} [R(y, 0) + R(0, y)], \tag{5.8}$$

with $x_1 = 1$. Simplifying (5.8) gives

$$1 + \sum_{y \geq 1} \left[\frac{(qx_2)_y}{(q)_y} \frac{q^{3y^2 - y}}{(q/x_2)_{y-1} \, (qx_2)_{y-1}} \right] \tag{5.9a}$$

$$\times \left\{ \frac{1}{(x_2 - 1)} \left[(q)_{y-1} \, (qx_2)_{y-1} \, (1 - q^{2y}) \, x_2^{1-y} \right. \right.$$

$$\left. \left. - (qx_2^2)_{y-1} \, (q/x_2)_{y-1} \, (1 - x_2^2 q^{2y}) \, x_2^{5y} \right] \right\}. \tag{5.9b}$$

The limit as $x_2 \to 1$ of the factors in (5.9a) is

$$\frac{q^{3y^2-y}}{(q)_{y-1}^2}. \tag{5.10}$$

An elementary calculation involving l'Hôpital's rule and log differentiation shows that the limit as $x_2 \to 1$ of the factors in (5.9b) is

$$(q)_{y-1}^2\,(1-q^{2y})\left[(1-6y)+\frac{2q^{2y}}{(1-q^{2y})}\right]. \tag{5.11}$$

Thus, multiplying (5.10) and (5.11), we see that the limit of the term in (5.9) is

$$(6y+1)q^{3y^2+y}+(-6y+1)q^{3y^2-y}. \tag{5.12}$$

It is now clear that the $x_1 = x_2 = 1$ and $\ell = 2$ case of (2.22a) is the Laurent series

$$\sum_{y=-\infty}^{\infty}(6y+1)q^{3y^2+y}. \tag{5.13}$$

Taking $q \mapsto q^2$ in the quintuple product identity (A.8) in Appendix A, differentiating both sides with respect to z, setting $z=-1$, and simplifying yields

$$\sum_{y=-\infty}^{\infty}(6y+1)q^{3y^2+y} = (q^2;q^2)_\infty^3\,(q^2;q^4)_\infty^2. \tag{5.14}$$ □

Remark. Equation (5.14), with $q^2 \mapsto q$, was proved, substantially as in the preceeding, by Gordon in [Gor61, eq. (11)]. A more recent exposition appears in [BB87, pp. 306–307]. The $q^2 \mapsto q$ case of (5.14) is also equivalent to Macdonald's BC_1 identity in [Mac72, p. 93 and Appendix I, Type BC_1, eq. 6(a)].

It is now not hard to see that the $x_1 = x_2 = 1$ and $\ell = 2$ case of Corollary 2.21 can be written as the following theorem.

THEOREM 5.15 (the $x_1 = x_2 = 1$ C_2 Rogers–Ramanujan identity). *Let* $0 < |q| < 1$. *We then have*

$$\frac{(-q;q^2)_\infty^2}{(q;q^2)_\infty} = 1+\sum_{m=1}^{\infty}\frac{q^m}{(q)_m}\sum_{y=0}^{m}(2m-2y+1)\,K_{(2m-y,y)(2^m)}(q). \tag{5.16}$$

Proof. Apply Lemma 5.6 to (2.22a), recall (3.28), note the $q=1$ and $\ell=2$ case of (3.20), and then simplify. □

The product side of (5.16) also arises in $(q;q)_\infty^{-1}$ times the simple specialization of the Jacobi triple-product identity, (A.6) of Appendix A in which $q \mapsto q^2$ and $z=-1$. We have

$$\frac{(-q;q^2)_\infty^2}{(q;q^2)_\infty} = \frac{1}{(q;q)_\infty}\sum_{m=-\infty}^{\infty}q^{m^2}. \tag{5.17}$$

Equation (5.17) is just equivalent to a classical identity of Gauss in [And76, eq. (2.2.12), p. 23].

Equating the sum sides of (5.16) and (5.17) immediately gives

$$1+2\sum_{m=1}^{\infty}q^{m^2}=(q)_\infty\left\{1+\sum_{m=1}^{\infty}\frac{q^m}{(q)_m}\sum_{y=0}^{m}(2m-2y+1)\,K_{(2m-y,y)(2^m)}(q)\right\}. \tag{5.18}$$

The $x_1=x_2=1$ and $\ell=2$ specialization of Theorem 4.1 turns out to be

$$(q)_\infty^3=1+\sum_{y\geq 1}(4y+1)q^{2y^2+y}+\sum_{y\geq 1}(-4y+1)q^{2y^2-y}, \tag{5.19}$$

which is equivalent to both Jacobi's [HW79, Thm. 357, p. 285] expansion for $(q)_\infty^3$ and also to Macdonald's [Mac72, Eq. (0.7), p. 93]. The derivation of (5.19) from Theorem 4.1 is exactly the same as the analysis, up to (5.13), in the proof of Lemma 5.6.

The $x_1=\cdots=x_\ell=1$ case of Theorem 4.1 provides an expansion for $(q)_\infty^{\binom{\ell+1}{2}}$. It should be possible to show directly that such expansions are equivalent to the $C_\ell^\vee$ or $B_\ell^\vee$ identities in [Mac72, Appendix I].

We conclude this section with the $x_1=x_2=x_3=1$ and $\ell=3$ specialization of Corollary 2.21 in the following corollary.

COROLLARY 5.20 (the $x_1=x_2=x_3=1$ C_3 Rogers–Selberg identity). *Let* $0<|q|<1$. *We then have*

$$\sum_{-\infty<y_1,y_2<\infty}(-1)^{y_1+y_2}q^{(7y_1^2+7y_2^2-3y_1-y_2)/2}$$

$$\times\left[\frac{(7y_1-1)(7y_1-2)}{2}-\frac{7y_2(7y_2-1)}{2}\right] \tag{5.21a}$$

$$=(q)_\infty^6\sum_{m=0}^{\infty}\frac{q^m}{(q)_m}\sum_{\substack{\mu\geq(2^m)\\ |\mu|=2m\\ \ell(\mu)\leq 3}}K_{\mu(2^m)}(q)\;s_\mu(1,1,1). \tag{5.21b}$$

Proof. Equation (5.21b) is immediate. To establish (5.21a) we carry out a very long elementary calculation that is analogous to the proof of Lemma 5.6, up to (5.13).

First, by Lemma 5.1 we only need to look at the terms in (2.22a) corresponding to the following cases:

$$y_1=0 \quad\text{and}\quad y_2,y_3>0, \tag{5.22a}$$

$$y_2=0 \quad\text{and}\quad y_1,y_3>0. \tag{5.22b}$$

$$y_1>0 \quad\text{and}\quad y_2=y_3=0, \tag{5.23a}$$

$$y_2>0 \quad\text{and}\quad y_1=y_3=0, \tag{5.23b}$$

$$y_3>0 \quad\text{and}\quad y_1=y_2=0. \tag{5.23c}$$

$$y_1=y_2=y_3=0. \tag{5.24}$$

We compute the limit as $x_1,x_2,x_3\to 1$ of each of the cases (5.22)–(5.24) separately. First, relabel the terms in (5.22a) and (5.22b) by $y_2\mapsto y_1$, $y_3\mapsto y_2$, and $y_1\mapsto y_1$, $y_3\mapsto y_2$, respectively. We then compute the limit of the sum of the terms in (5.22). Set $x_1=1$ and then calculate the limit as $x_2\to 1$. We obtain an expression depending on x_3, where $0<y_1\leq y_2$. At this point we can just set $x_3=1$ whenever

$0 < y_1 = y_2$. We obtain four terms, which correspond to (5.21a) with $y_1 = y_2 \neq 0$ or $y_1 = -y_2 \neq 0$. Otherwise, if $0 < y_1 \neq y_2$, we compute the limit as $x_3 \to 1$ of the sum of our expressions corresponding to (y_1, y_2) and (y_2, y_1). After a lengthly calculation we obtain eight terms, each indexed by $0 < y_1 < y_2$, which account for all the terms in (5.21a) except those in which $y_1 = 0$, $y_2 = 0$, or $y_1 = \pm y_2$.

Next, consider the limit of the sum of the three terms in (5.23), once we have relabeled each by $y_i \mapsto y$. Set $x_1 = 1$, and then compute the limit as $x_2, x_3 \to 1$. We end up with four terms, indexed by $0 < y$, which account for the terms in (5.21a) in which $y_1 = 0$, or $y_2 = 0$, but $(y_1, y_2) \neq (0, 0)$.

Finally, just note that the $y_1 = y_2 = y_3 = 0$ case of (2.22a) is 1. □

Just as for (3.24a), the double Laurent series (5.21a) does not factor into a simple infinite product. We obtain (3.44) in which the a_n's alternate in sign, they have no simple pattern, and $|a_n| \to \infty$. A similar situation probably holds when $\ell \geq 3$.

Nonetheless, the $x_1 = \cdots = x_\ell = 1$ case of (2.22a) should be multiple Laurent series similar to those in [Mac72, Appendix I].

Appendix A: Background information. The main results in this paper depend on the C_ℓ terminating very well poised ${}_6\phi_5$ summation theorem from [Gus89], [LM91] and the C_ℓ q-Whipple transformation from [ML92b].

We start with the following theorem.

THEOREM A.1 (the C_ℓ terminating ${}_6\phi_5$ summation theorem). *Let a, b and $x_1, \ldots, x_\ell$ be indeterminate, let N_i be nonnegative integers for $i = 1, 2, \ldots, \ell$, with $\ell \geq 1$, and suppose that none of the denominators in* (A.2) *vanishes. Then*

$$\sum_{\substack{0 \leq y_k \leq N_k \\ k=1,2,\ldots,\ell}} \left\{ \prod_{k=1}^{\ell} \left[\frac{1 - x_k^2 q^{2y_k}}{1 - x_k^2} \right] \prod_{1 \leq r < s \leq \ell} \left[\frac{1 - \frac{x_r}{x_s} q^{y_r - y_s}}{1 - \frac{x_r}{x_s}} \, \frac{1 - x_r x_s q^{y_r + y_s}}{1 - x_r x_s} \right] \right.$$

$$\times \prod_{r,s=1}^{\ell} \left[\frac{\left(\frac{x_r}{x_s} q^{-N_s}\right)_{y_r} (x_r x_s)_{y_r}}{\left(q \frac{x_r}{x_s}\right)_{y_r} (q x_r x_s q^{N_s})_{y_r}} \right] \prod_{k=1}^{\ell} \left[\frac{(a x_k)_{y_k} (q x_k b^{-1})_{y_k}}{(b x_k)_{y_k} (q x_k a^{-1})_{y_k}} \right]$$

$$\left. \times \, q^{(N_1 + \cdots + N_\ell)(y_1 + \cdots + y_\ell)} \, q^{y_2 + 2y_3 + \cdots + (\ell-1) y_\ell} \left(\frac{b}{a}\right)^{y_1 + \cdots + y_\ell} \right\} \tag{A.2a}$$

$$= \prod_{k=1}^{\ell} \left[\frac{(q x_k^2)_{N_k}}{(b x_k)_{N_k} (q a^{-1} x_k)_{N_k}} \right] \prod_{1 \leq r < s \leq \ell} \left[\frac{(q x_r x_s)_{N_r}}{(q x_r x_s q^{N_s})_{N_r}} \right]$$

$$\times \left(\frac{b}{a}\right)_{N_1 + \cdots + N_\ell}. \tag{A.2b}$$

Proof. We derive (A2) in [LM91] from Gustafson's C_ℓ ${}_6\psi_6$ summation theorem [Gus89, Thm. 5.1]. Specializations serve to terminate Gustafson's C_ℓ ${}_6\psi_6$ summation from below and then from above. These yield the C_ℓ ${}_6\phi_5$ summation theorem and then (A2), respectively.

A summary of the substitutions that transform Gustafson's C_ℓ ${}_6\psi_6$ into Theorem A.1 is given by

$$a_i \mapsto a_i q^{-z_i} \mapsto q^{-N_i} q^{-z_i} \mapsto q^{-N_i} x_i^{-1} \qquad \text{for} \quad i = 1, 2, \ldots, \ell;$$
$$a_{\ell+1} \mapsto a;$$

$$b_i \mapsto b_i q^{-z_i} \mapsto q^{1-z_i} \mapsto q x_i^{-1} \qquad \text{for} \quad i = 1, 2, \dots, \ell;$$
$$b_{\ell+1} \mapsto b;$$
$$q^{z_i} \mapsto x_i. \qquad \square$$

Remark. The $\ell = 1$ case of (A2) is the classical terminating ${}_6\phi_5$ summation in [GR90, eq. (II.21), p. 238] in which $a \mapsto x_1^2$, $n \mapsto N_1$, $b \mapsto a x_1$, and $c \mapsto q x_1 b^{-1}$. That is, they are equivalent.

See [LM91, §2] for the detailed proof of Theorem 2.5.

The unspecialized C_ℓ Sylvester identity in Theorem 4.1 of §4 is a special limiting case of Theorem A.1.

One of the most important applications of Theorem A.1 and the C_ℓ Bailey pair inversion theorem of [LM91], [ML92a] is the following theorem.

THEOREM A.3 (The C_ℓ q-Whipple transformation). *Let a, b, α, β, and $x_1, \dots, x_\ell$ be indeterminate, let N_i be nonnegative integers for $i = 1, 2, \dots, \ell$, with $\ell \geq 1$, and suppose that none of the denominators in* (A.4) *vanishes. Then*

$$\begin{aligned}
&\sum_{\substack{0 \leq y_k \leq N_k \\ k=1,2,\dots,\ell}} \left\{ \prod_{k=1}^{\ell} \left[\frac{1 - x_k^2 q^{2y_k}}{1 - x_k^2} \right] \prod_{1 \leq r < s \leq \ell} \left[\frac{1 - \frac{x_r}{x_s} q^{y_r - y_s}}{1 - \frac{x_r}{x_s}} \, \frac{1 - x_r x_s q^{y_r + y_s}}{1 - x_r x_s} \right] \right. \\
&\quad \times \prod_{k=1}^{\ell} \left[\frac{(a x_k)_{y_k} (q x_k b^{-1})_{y_k}}{(b x_k)_{y_k} (q x_k a^{-1})_{y_k}} \, \frac{(\alpha x_k)_{y_k} (q x_k \beta^{-1})_{y_k}}{(\beta x_k)_{y_k} (q x_k \alpha^{-1})_{y_k}} \right] \\
&\quad \times \prod_{r,s=1}^{\ell} \left[\frac{\left(\frac{x_r}{x_s} q^{-N_s}\right)_{y_r} (x_r x_s)_{y_r}}{\left(q \frac{x_r}{x_s}\right)_{y_r} (q x_r x_s q^{N_s})_{y_r}} \right] q^{(N_1 + \cdots + N_\ell)(y_1 + \cdots + y_\ell)} \\
&\quad \left. \times \left[q^{y_2 + 2y_3 + \cdots + (\ell-1) y_\ell} \left(\frac{b\beta}{a\alpha} \right)^{y_1 + \cdots + y_\ell} \right] \right\} \qquad \text{(A.4a)} \\
&= \prod_{k=1}^{\ell} \left[\frac{(q x_k^2)_{N_k}}{(b x_k)_{N_k} (q a^{-1} x_k)_{N_k}} \right] \prod_{1 \leq r < s \leq \ell} \left[\frac{(q x_r x_s)_{N_r}}{(q x_r x_s q^{N_s})_{N_r}} \right] \left(\frac{b}{a} \right)_{N_1 + \cdots + N_\ell} \\
&\quad \times \sum_{\substack{0 \leq m_k \leq N_k \\ k=1,2,\dots,\ell}} \left\{ \prod_{1 \leq r < s \leq \ell} \left[\frac{1 - \frac{x_r}{x_s} q^{m_r - m_s}}{1 - \frac{x_r}{x_s}} \right] q^{m_1 + 2m_2 + \cdots + \ell m_\ell} \right. \\
&\quad \times \prod_{r,s=1}^{\ell} \left[\frac{\left(\frac{x_r}{x_s} q^{-N_s}\right)_{m_r}}{\left(q \frac{x_r}{x_s}\right)_{m_r}} \right] \prod_{k=1}^{\ell} \left[\frac{(a x_k)_{m_k} (q x_k b^{-1})_{m_k}}{(\beta x_k)_{m_k} (q x_k \alpha^{-1})_{m_k}} \right] \\
&\quad \left. \times \left[\left(\frac{\beta}{\alpha} \right)_{m_1 + \cdots + m_\ell} \left(\frac{a}{b} q^{1 - (N_1 + \cdots + N_\ell)} \right)_{m_1 + \cdots + m_\ell}^{-1} \right] \right\}. \qquad \text{(A.4b)}
\end{aligned}$$

Proof. We establish [ML92b, eq. (A.4)] by extending the analysis of the classical ($\ell = 1$) case of [Wat29a], [GR90]. First, apply the C_ℓ Bailey pair inversion theorem of [Lil91], [LM91], [ML92a] to the C_ℓ Bailey pair determined by Theorem A.1. This gives the C_ℓ terminating balanced ${}_3\phi_2$ summation in [ML92b, Thm. 4.4].

The rest of the proof is very similar to the A_ℓ case. Multiply each term in (A.2a) by a suitable rewriting of the product side of [ML92b, Thm. 4.4]. The resulting sum is rewritten as (A.4a). Then use [ML92b, Thm. 4.4] to replace the factors just added by the corresponding sum. At this point, interchange summation and manipulate the resulting inner multiple sum termwise until a shifted C_ℓ very well poised ${}_6\phi_5$ sum is obtained. Use Theorem A.1 to sum this inner sum. Finally, simplify the resulting single multiple sum termwise to obtain (A.4b). □

Remark. The $\ell = 1$ case of (A.4) is [GR90, eq. (III.18), p. 242], in which $a \mapsto x_1^2$, $n \mapsto N_1$, $b \mapsto \alpha x_1$, $c \mapsto q x_1 \beta^{-1}$, $d \mapsto a x_1$, and $e \mapsto q x_1 b^{-1}$. That is, they are equivalent.

The first version of the C_ℓ Rogers–Selberg identity in Theorem 2.1 of §2 is a special limiting case of Theorem A.3.

The derivation of the first C_2 Rogers–Ramanujan identity in Theorem 3.26 and the theta-function identity in Corollary 4.7 require Jacobi's [Jac29] well-known triple product identity.

THEOREM A.5 (Jacobi triple product identity). *Let* $|q| < 1$ *and* $z \neq 0$. *Then*

$$\left(zq^{\frac{1}{2}}, q^{\frac{1}{2}}/z, q; q\right)_\infty = \sum_{m=-\infty}^{\infty} (-1)^m q^{m^2/2} z^m. \tag{A6}$$

See [GR90, p. 12] for an analytical proof of (A.6), and see [And76], [And86], [Bai35], [GR90], [HW79], [Sla66], [Wat29a] for applications of (A.6) to classical Rogers–Ramanujan-type identities.

The final part of the proof of the $x_1 = x_2 = 1$ C_2 Rogers–Ramanujan identity in Theorem 5.15 relies on the following theorem.

THEOREM A.7 (quintuple-product identity). *Let* $0 < |q| < 1$ *and* $z \neq 0$. *Then*

$$\begin{aligned} &\sum_{m=-\infty}^{\infty} (-1)^m q^{m(3m-1)/2} z^{3m} (1 + zq^m) \\ &\qquad = (q, -z, -q/z; q)_\infty \, (qz^2, q/z^2; q^2)_\infty . \end{aligned} \tag{A.8}$$

The identity (A.8) was discovered by Watson [Wat29b] and was rediscovered by Gordon [Gor61], and an equivalent identity is explicitly given by Ramanujan in [Ram88, p. 207]. A more recent exposition can be found in [BB87], [GR90]. The identity (A.8) is also the BC_1 Macdonald identity [Mac72].

Equation (A.8) is the formulation given by [GR90, Ex. 5.6, p. 134].

REFERENCES

[And76] G. E. ANDREWS, *The Theory of Partitions*, Encyclopedia of Mathematics and Its Applications, (G.-C. Rota, ed.), Vol. 2, Addison-Wesley, Reading, MA, 1976; reissued by Cambridge University Press, Cambridge, England, 1985.

[And84a] ———, *Hecke modular forms and the Kac–Peterson identities*, Trans. Amer. Math. Soc., 283 (1984), pp. 451–458.

[And84b] ———, *Multiple series Rogers–Ramanujan type identities*, Pacific J. Math., 114 (1984), pp. 267–283.

[And86] ———, *q-Series: Their development and application in analysis, number theory, combinatorics, physics and computer algebra,* in National Science Foundation Conference Board of the Mathematical Sciences Regional Conference Series, Vol. 66, 1986, pp. 1–130.

[AK26] P. APPELL AND J. KAMPÉ DE FÉRIET, *Fonctions hypergéométriques et hypersphériques; polynomes d'hermites*, Gauthier-Villars, Paris, 1926.

[Bai35] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, England, 1935; reprinted by Stechert-Hafner, New York, 1964.

[BL68] L. C. BIEDENHARN AND J. D. LOUCK, *A pattern calculus for tensor operators in the unitary groups*, Comm. Math. Phys. 8 (1968), pp. 89–131.

[BL81a] ———, *Angular Momentum in Quantum Physics: Theory and Applications*, Encyclopedia of Mathematics and Its Applications, (G.-C. Rota, ed.), Vol. 8, Addison-Wesley, Reading, MA, 1981.

[BL81b] ———, *The Racah–Winger Algebra in Quantum Theory*, Encyclopedia of Mathematics and Its Applications, (G.-C. Rota, ed.), Vol. 9, Addison-Wesley, Reading, MA, 1981.

[BB87] J. M. BORWEIN AND P. B. BORWEIN, *Pi and the AGM–A Study in Analytic Number Theory and Computational Complexity*, Canad. Math. Soc. Ser. Monographs Adv. Texts, John Wiley, New York, 1987.

[Bre80] D. M. BRESSOUD, *Analytic and combinatorial generalizations of the Rogers–Ramanujan identities*, Mem. Amer. Math. Soc., 24 (1980), pp. 1–54.

[Bre86] ———, *Hecke modular forms and q-Hermite polynomials*, Illinois J. Math., 30 (1986), pp. 185–196.

[Bre87] ———, *Colored tournaments and Weyl's denominator formula*, European J. Combin., 8 (1987), pp. 245–255.

[GR90] G. GASPER AND M. RAHMAN, *Basic Hypergeometric Series*, Encyclopedia of Mathematics and Its Applications, (G.-C. Rota, ed.), Vol. 35, Cambridge University Press, Cambridge, England, 1990.

[Gor61] B. GORDON, *Some identities in combinatorial analysis*, Quart. J. Math. Oxford Ser. (2), 12 (1961), pp. 285–290.

[Gus89] R. A. GUSTAFSON, *The Macdonald identities for affine root systems of classical type and hypergeometric series very well-poised on semi-simple Lie algebras*, in Proc. Ramanujan International Symposium on Analysis, Dec. 26–28, 1987, Pune, India, (N. K. Thakare, ed.), 1989, pp. 187–224.

[HW79] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, fifth ed., Oxford University Press, Oxford, England, 1979.

[Hec59] E. HECKE, *Über einen Zusammenhang zwischen elliptischen Modulfunktionen und indefiniten quadratischen Formen*, in Mathematische Werke, Vandenhoeck and Ruprecht, Göttingen, Germany, 1959, pp. 418–427.

[Hol80] W. J. HOLMAN III, *Summation theorems for hypergeometric series in $U(n)$*, SIAM J. Math. Anal., 11 (1980), pp. 523–532.

[HBL76] W. J. HOLMAN III, L. C. BIEDENHARN, AND J. D. LOUCK, *On hypergeometric series well-poised in $SU(n)$*, SIAM J. Math. Anal., 7 (1976), pp. 529–541.

[Hor89] J. HORN, *Ueber die Convergenz der hypergeometrische Reihen zweier und dreier Veränderlichen*, Math. Ann., 34 (1889), pp. 544–600.

[Jac29] C. G. J. JACOBI, *Fundamenta Nova Theoriae Functionum Ellipticarum*, Regiomonti. Sumptibus fratrum Borntträger; 1829, reprinted in Jacobi's Gesammelte Werke 1, Reimer, Berlin, 1881–1891, pp. 49–239; reprinted by Chelsea, New York, 1969.

[KP80] V. G. KAC AND D. H. PETERSON, *Affine Lie algebras and Hecke modular forms*, Bull. Amer. Math. Soc. N.S., 3 (1980), pp. 1057–1061.

[Kir88] A. N. KIRILLOV, *On the Kostka–Green–Foulkes polynomials and Clebsch–Gordan numbers*, J. Geom. Phys., 5 (1988), pp. 365–389.

[Lil91] G. M. LILLY, *The C_ℓ Generalization of Bailey's Transform and Bailey's Lemma*, Ph.D. thesis,, University of Kentucky, Lexington, KY, 1991.

[LM91] G. M. LILLY AND S. C. MILNE, *The C_ℓ Bailey transform and Bailey lemma*, Constr. Approx., 9 (1993), pp. 473–500.

[Mac72] I. G. MACDONALD, *Affine root systems and Dedekind's η-function*, Invent. Math., 15 (1972), pp. 91–143.

[Mac79] ———, *Symmetric Functions and Hall Polynomials*, Oxford University Press, Oxford, England, 1979.

[MS73] A. M. MATHAI AND R. K. SAXENA, *Generalized Hypergeometric Functions with Applications in Statistics and Physical Sciences*, Lecture Notes in Math., Vol. 348, Springer-Verlag, New York, 1973.

[Mil88a] S. C. MILNE, *A q-analog of the Gauss summation theorem for hypergeometric series in $U(n)$*, Adv. in Math., 72 (1988), pp. 59–131.

[Mil88b] ———, *Multiple q-series and $U(n)$ generalizations of Ramanujan's ${}_1\Psi_1$ sum*, in Ramanujan Revisited, G. E. Andrews, R. A. Askey, B. C. Berndt, K. G. Ramanathan, and R. A. Rankin, eds., Academic Press, New York, 1988, pp. 473–524.

[Mil89] ———, *The multidimensional ${}_1\Psi_1$ sum and Macdonald identities for $A_\ell^{(1)}$*, in Proc. Symposia in Pure Mathematics, Vol 49 (Part 2), L. Ehrenpreis and R. C. Gunning, eds., American Mathematical Society, Providence, RI, 1989, pp. 323–359.

[Mil92] ———, *Classical partition functions and the $U(n+1)$ Rogers–Selberg identity*, Discrete Math., 99 (1992), pp. 199–246.

[ML92a] S. C. MILNE AND G. M. LILLY, *The A_ℓ and C_ℓ Bailey transform and lemma*, Bull. Amer. Math. Soc. N.S., 26 (1992), pp. 258–263.

[ML92b] ———, *Consequences of the A_ℓ and C_ℓ Bailey transform and Bailey lemma*, Discrete Math., to appear.

[Ram88] S. RAMANUJAN, *The Lost Notebook and Other Unpublished Papers*, Narosa, New Delhi, 1988.

[Rog94] L. J. ROGERS, *Second memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 25 (1894), pp. 318–343.

[Sla66] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, London, 1966.

[Sta71] R. P. STANLEY, *Theory and applications of plane partitions*, Stud. Appl. Math., 50 (1971), pp. 167–188, 259–279.

[Wat29a] G. N. WATSON, *A new proof of the Rogers–Ramanujan identities*, J. London Math. Soc., 4 (1929), pp. 4–9.

[Wat29b] ———, *Theorems stated by Ramanujan* (VII): *Theorems on continued fractions*, J. London Math. Soc., 4 (1929), pp. 39–48.

[Whi24] F. J. W. WHIPPLE, *On well-poised series, generalized hypergeometric series having parameters in pairs, each pair with the same sum*, Proc. London Math. Soc., 24 (1924), pp. 247–263.

[Whi26] ———, *Well-poised series and other generalized hypergeometric series*, Proc. London Math. Soc., 25 (1926), pp. 525–544.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 596–601, March 1994

023

POSITIVITY OF INTEGRALS OF BESSEL FUNCTIONS*

JOLANTA K. MISIEWICZ† AND DONALD ST. P. RICHARDS‡

Abstract. Using results on multiply monotone functions, we establish the positivity of integrals of Bessel functions of the form

$$\int_0^x (x^\mu - t^\mu)^\lambda t^\alpha J_\beta(t)dt, \qquad x > 0,$$

where $0 < \mu \leq 1 \leq \lambda$ and α, β satisfy various conditions. In particular, the result holds if $-\frac{1}{2} \leq \alpha = \beta \leq \frac{3}{2}$ or if $\frac{3}{2} = \alpha \leq \beta$.

Key words. Bessel functions, characteristic functions, fractional integration, multiply monotone functions, Pólya's criterion, Riesz means

AMS subject classifications. primary 33A45; secondary 60E10

1. Introduction.

The problem of proving positivity of the Bessel function integrals,

$$\int_0^x (x^\mu - t^\mu)^\lambda t^\alpha J_\beta(t)dt, \qquad x > 0, \tag{1.1}$$

is an old one. We refer to Askey [1] and Gasper [5] for reviews of the techniques developed for these problems, and for references to applications of those results. For our purposes it is enough to say that much is known about the positivity of (1.1) when $\mu = 1$ or 2, but that very little seems to be available for other values of μ. In particular, the explicit formulas for the classical hypergeometric series that proved to be extremely useful when $\mu = 1$ or 2 do not appear to be as helpful for noninteger values of μ. This has, perhaps, been the main difficulty in the analysis of (1.1), despite the need (cf. [1, p. 90]) for positivity results when $0 < \mu < 1$.

Here, we obtain positivity results for (1.1) when $0 < \mu < 1$. Our starting point is the observation of Wintner [13, pp. 126 *bis, tre*], that if t_+^λ is t^λ or 0 according as $t > 0$ or $t \leq 0$, respectively, then for $0 < \mu \leq 1$, the function $\phi_\mu : \mathbf{R} \to \mathbf{R}$ given by $\phi_\mu(t) = (1 - |t|^\mu)_+$ is a characteristic function; that is, ϕ_μ is the Fourier transform of a probability measure. (As Loren Pitt reminded us, the fact that ϕ_μ is a characteristic function follows immediately from Pólya's criterion [4]: If a function $\phi : \mathbf{R} \to \mathbf{R}$ is symmetric, convex, and satisfies $\phi(0) = 1$, then ϕ is a characteristic function.)

Since ϕ_μ is integrable on $\mathbf{R}$, then its inverse Fourier transform exists and is nonnegative. Therefore if $0 < \mu \leq 1$, then by the symmetry of ϕ_μ,

$$\int_0^1 (1 - t^\mu) \cos xt \; dt \geq 0, \qquad x > 0, \tag{1.2}$$

---

*Received by the editors March 2, 1992; accepted for publication (in revised form) February 12, 1993.

†Institute of Mathematics, Technical University of Wrocław, Wrocław 50-370, Poland. This research was supported in part by the Ministerstwo Edukacij Narodowej grant DNS-P/05/028/90-2.

‡Department of Mathematics, University of Virginia, Charlottesville, Virginia 22903. This research was supported in part by National Science Foundation grant DMS-9101740.

or

$$\int_0^x (x^\mu - t^\mu) \cos t \, dt \geq 0, \qquad x > 0.$$

Wintner actually proves that ϕ_μ is a characteristic function by directly establishing (1.2). Moreover, he proves that if $\mu > 1$ then (1.2) fails for some $x > 0$. His proof is so simple that we cannot resist the temptation to reproduce it: After an integration by parts, we find that (1.2) is equivalent to

$$\int_0^x t^{\mu-1} \sin t \, dt \geq 0, \qquad x > 0. \tag{1.3}$$

Decompose the interval $(0, x)$ into subintervals $((n-1)\pi, n\pi)$, $n = 1, 2, \ldots$. Then the absolute value of the contribution of the nth interval to the integral in (1.3) is an increasing or decreasing function of n according to whether the function $t^{\mu-1}$ is increasing or decreasing; that is, as $\mu > 1$ or $0 < \mu < 1$, respectively. Since $(-1)^{n-1} \sin t$ is positive if $(n-1)\pi < t < n\pi$, then the nth contribution has the same sign as $(-1)^{n-1}$ and is therefore positive for $n = 1$. Consequently, the integral (1.3) is negative for some $x > 0$ if $\mu > 1$, and is positive for all $x > 0$ if $0 < \mu < 1$. If $\mu = 1$, then (1.3) is easily seen to be valid, so that (1.3) holds for all $0 < \mu \leq 1$.

We also note that (1.3) (and hence (1.2)) can be derived from an expansion as a sum of squares, obtained by applying the expansion formula of Gasper [5, eq. (3.2)] to a special case (with $\lambda = 0$) of [5, eq. (2.20)].

2. Results. In our first extension of (1.2), we replace the function ϕ_μ by more general convex functions.

PROPOSITION 2.1. *Suppose that $\phi : (0, \infty) \to \mathbf{R}$ is nonnegative, nonincreasing, convex, and satisfies $\phi(1) = 0$. Then*

$$\int_0^1 \phi(t) \cos xt \, dt \geq 0, \qquad x > 0. \tag{2.1}$$

Proof. We use the results of Williamson [12], on multiply monotone functions, to derive (2.1). In the terminology of [12], the function ϕ is 2-monotone. By [12], Theorem 1, there exists a unique, nondecreasing function, F, which is bounded from below, such that

$$\phi(t) = \int_0^\infty (1 - ut)_+ \, dF(u), \qquad t > 0. \tag{2.2}$$

Since $\phi(1) = 0$, then it follows from (2.2) that

$$0 = \int_0^\infty (1-u)_+ \, dF(u) = \int_0^1 (1-u)_+ \, dF(u).$$

Therefore $dF(u) = 0$ for $u \leq 1$.

Substituting (2.2) into the integral in (2.1), using Fubini's theorem to reverse the

order of integration, and denoting $u^{-1}dF(u^{-1})$ by $dG(u)$, we have

$$\begin{aligned}\int_0^1 \phi(t)\cos xt\ dt &= \int_0^1 \int_1^\infty (1-ut)_+ \cos xt\ dF(u)\ dt \\ &= \int_0^1 \int_0^1 \left(1-\frac{t}{u}\right)_+ \cos xt\ dF(u^{-1})\ dt \\ &= \int_0^1 \int_0^1 (u-t)_+ \cos xt\ dG(u)\ dt \\ &= \int_0^1 \int_t^1 (u-t)\cos xt\ dG(u)\ dt \\ &= \int_0^1 \int_0^u (u-t)\cos xt\ dt\ dG(u).\end{aligned}$$

Since

$$\int_0^u (u-t)\cos xt\ dt = x^{-2}(1-\cos(ux)) \geq 0, \qquad u>0, x>0,$$

then (2.1) is immediately seen to be nonnegative. □

The proof of Proposition 2.1 has the advantage of explicitly representing the integral (2.1) as a nonnegative function. By choosing $\phi(t) = (1-t^\mu)_+^\lambda$, where $0 < \mu \leq 1 \leq \lambda$, we get the following (portion of a) result of Kuttner [8].

COROLLARY 2.2 (Kuttner [8]). *For* $0 < \mu \leq 1 \leq \lambda$,

$$\int_0^x (x^\mu - t^\mu)^\lambda \cos t\ dt \geq 0, \qquad x>0. \tag{2.3}$$

Note that (2.3) also follows from Pólya's criterion, since the function $\phi(t) = (1-t^\mu)_+^\lambda$ is convex for $0 < \mu \leq 1 \leq \lambda$. When $0 < \mu < 2$, in which case the function ϕ may no longer be convex, Kuttner [8] proved that there exists a continuous, strictly increasing function $k(\mu)$ such that (2.3) is valid if $\lambda \geq k(\mu)$ and invalid for $\lambda < k(\mu)$. He also proved that $k(\mu) \to \infty$ as $\mu \to 2$; $k(1) = 1$; $k(\mu) > \mu$, $\mu \neq 1$; and $0 < k(0+) < 1$.

Substituting for $\cos t$ in terms of the Bessel function $J_{-1/2}(t)$, (2.3) is clearly equivalent to

$$\int_0^x (x^\mu - t^\mu)^\lambda t^{1/2} J_{-1/2}(t)\ dt \geq 0, \qquad x>0. \tag{2.4}$$

We now extend (1.2) and (2.4) by replacing the Bessel function $J_{-1/2}(t)$ by Bessel functions of a more general index.

PROPOSITION 2.3. *If* ϕ *satisfies the hypotheses of Proposition* 2.1, *and* α, β *are such that*

$$\int_0^x (x-t)t^\alpha J_\beta(t)\ dt \geq 0, \qquad x>0, \tag{2.5}$$

then

$$\int_0^1 \phi(t)t^\alpha J_\beta(xt)\ dt \geq 0, \qquad x>0.$$

Proof. We again apply Williamson's integral representation for ϕ. Then it follows as before that

$$\int_0^1 \phi(t)t^\alpha J_\beta(xt)\ dt = \int_0^1 \int_0^u (u-t)t^\alpha J_\beta(xt)\ dt\ dG(u).$$

Since (2.5) is equivalent to

$$\int_0^u (u-t)t^\alpha J_\beta(xt)\,dt \geq 0, \qquad u>0, x>0,$$

then the result follows. □

There appear to be numerous conditions under which (2.5) is valid. We list all conditions known to us in the following result.

COROLLARY 2.4. *For* $0<\mu\leq 1\leq\lambda$,

$$\int_0^x (x^\mu - t^\mu)^\lambda t^\alpha J_\beta(t)\,dt \geq 0, \qquad x>0,$$

if α *and* β *satisfy any of the following conditions*:

(i) $\frac{3}{2}=\alpha\leq\beta$;
(ii) $-\frac{1}{2}\leq\alpha=\beta\leq 3/2$;
(iii) $\alpha=0$, $\beta>-1$;
(iv) $\frac{1}{2}=\alpha<\beta$;
(v) $-1-\beta<\alpha<\alpha(\beta)$, $-1<\beta<\frac{1}{2}$, *where* $\alpha(\beta)$ *is defined by*

$$\int_0^{j_{\beta,2}} t^{\alpha(\beta)}J_\beta(t)\,dt = 0,$$

and $j_{\beta,2}$ *denotes the second positive zero of* $J_\beta(t)$.

(vi) $\alpha=\alpha_0-\delta$, $\beta=\beta_0+\delta$, $\delta\geq 0$, *where* (α_0,β_0) *satisfy any of* (i)–(v).

Proof. In all six cases, the conclusion is obtained by proving that (2.5) is valid under the corresponding restriction on α and β, and then substituting $\phi(t)=(1-t^\mu)_+^\lambda$.

That (2.5) holds under (i) follows from the result of Moak [10], and (ii) follows from Gasper [5, eq. (1.5)]. Next, (iii)–(v) all follow from the identity

$$\int_0^x (x-t)t^\alpha J_\beta(t)\,dt = \int_0^x\int_0^t u^\alpha J_\beta(u)\,du,$$

and then noting that

$$\int_0^t u^\alpha J_\beta(u)\,du \geq 0, \qquad t\geq 0, \tag{2.6}$$

under the corresponding conditions on α and β. In particular, (iii) implies (2.6) by Cooke's inequality (cf. [5, eq. (1.1)]); (iv) implies (2.6) by [5, eq. (1.5)]; and (v) implies (2.6) by the results of Askey and Steinig [3] and Makai [9]. Finally, (vi) follows from [5, eq. (3.17)]. □

Another way to extend (2.1) is to use the fractional integration methods of Askey [1], [2] and Gasper [5], [6].

PROPOSITION 2.5. *If* ϕ *is as before and* $\gamma>-1$, *then*

$$\int_0^1 \phi(t)t^{-(\gamma+\frac{1}{2})}J_{\gamma+\frac{1}{2}}(xt)\,dt \geq 0, \qquad x>0. \tag{2.7}$$

In particular, for $0<\mu\leq 1\leq\lambda$,

$$\int_0^x (x^\mu-t^\mu)^\lambda t^{-(\gamma+\frac{1}{2})}J_{\gamma+\frac{1}{2}}(t)\,dt \geq 0, \qquad x>0. \tag{2.8}$$

Proof. By Poisson's integral [11, §12.11(1)],

$$t^{-(\gamma+\frac{1}{2})}J_{\gamma+\frac{1}{2}}(t) = c_\gamma \int_0^1 (1-u^2)^\gamma \cos tu \, du, \tag{2.9}$$

where the positive constant c_γ depends only on γ. In (2.1), replace x by ux, multiply both sides by $(1-u^2)^\gamma$, and integrate with respect to u over $(0,1)$; then we obtain (2.7) after interchanging the order of integration and applying (2.9). Finally, (2.8) follows from (2.7) in the usual way. □

From the results of [6], we know that the assumption $\mu \leq \lambda$ is necessary for positivity of integrals of the form considered here. In particular, the proof that $k(\mu) > \mu$ ($\mu \neq 1$) follows from the asymptotic behavior of the integral (2.3) as $x \to \infty$. In the following result we show that, even for small x, the assumption $\mu \leq \lambda$ cannot be dispensed with completely.

PROPOSITION 2.6. *Suppose that μ is fixed, $0 < \mu < 1$. Then there exists λ_0, $0 < \lambda_0 < 1$, such that*

$$\int_0^1 (1-t^\mu)^\lambda \cos(3\pi t/2)\, dt \quad \begin{cases} < 0, & 0 < \lambda < \lambda_0, \\ > 0, & \lambda_0 < \lambda < 1. \end{cases} \tag{2.10}$$

Proof. Denote the integral (2.10) by $\psi(\lambda)$. It is simple to check that the kernel $(\lambda, t) \mapsto (1-t^\mu)^\lambda$, $0 \leq \lambda, t \leq 1$, is strictly reverse rule of order 2 (Karlin [7, p. 12]). Further, the function $t \mapsto \cos(3\pi t/2)$, $0 \leq t \leq 1$, has one sign change. By the fundamental theorem of variation diminishing transformations [7, p. 233], it follows that $\psi(\lambda)$ changes sign at most once on $[0,1]$.

By a calculation we have $\psi(0) < 0$; and by Corollary 2.2, $\psi(1) > 0$; hence ψ changes sign at least once. Therefore ψ changes sign exactly once on $[0,1]$. □

The proof also shows that (2.10) remains valid if the function $\cos(3\pi t/2)$ is replaced by $\cos xt$, where $\pi/2 < x \leq 3\pi/2$. More generally, we can get similar results for the situation when $\cos xt$ is replaced by $t^\alpha J_\beta(xt)$ for suitably chosen α and β.

We conjecture that, in (2.10), λ_0 is a strictly increasing function of μ. Evidence supporting this conjecture is based on extensive calculation of λ_0, the results of which are presented partially in Table 1. The values of λ_0 were generated iteratively through Romberg integration of the integral in (2.10).

TABLE 1

μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
λ_0	0.409	0.434	0.461	0.489	0.519	0.551	0.583	0.618	0.654	0.692

Acknowledgments. We wish to thank Richard Askey and a referee for their comments on an earlier version of this manuscript, and especially for noting that results like (2.10) should be valid.

REFERENCES

[1] R. ASKEY, *Orthogonal Polynomials and Special Functions*, Regional Conference Series in Applied Math., Vol. 21, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.

[2] R. ASKEY, *Inequalities via fractional integration*, in Fractional Calculus and its Applications, B. Ross, ed., Lecture Notes in Math. 457, Springer-Verlag, New York, 1975, pp. 106–115.
[3] R. ASKEY AND J. STEINIG, *Some positive trigonometric sums*, Trans. Amer. Math. Soc., 187 (1974), pp. 295–307.
[4] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., John Wiley, New York, 1971.
[5] G. GASPER, *Positive integrals of Bessel functions*, SIAM J. Math. Anal., 6 (1975), pp. 868–881.
[6] ———, *Formulas of the Dirichlet–Mehler type*, in Fractional Calculus and its Applications, B. Ross, ed., Lecture Notes in Math. 457, Springer-Verlag, New York, 1975, pp. 106–115.
[7] S. KARLIN, *Total Positivity*, Vol. 1, Stanford Univ. Press, Stanford, CA, 1968.
[8] B. KUTTNER, *On the Riesz means of a Fourier series* (II), J. London Math. Soc., 19 (1944), pp. 77–84.
[9] E. MAKAI, *An integral inequality satisfied by Bessel functions*, Acta Math. Hungar., 25 (1974), pp. 387–390
[10] D. S. MOAK, *Completely monotonic functions of the form* $s^{-b}(s^2+1)^{-a}$, Rocky Mountain J. Math., 17 (1987), pp. 719–725.
[11] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, New York, 1944.
[12] R. E. WILLIAMSON, *Multiply monotone functions and their Laplace transforms*, Duke Math. J., 23 (1956), pp. 189–207.
[13] A. WINTNER, *Fourier Transforms and the Theory of Probability Distributions*, Edwards Bros., Ann Arbor, MI, 1938.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 602–614, March 1994

024

GENERALIZED JACOBI WEIGHTS, CHRISTOFFEL FUNCTIONS, AND JACOBI POLYNOMIALS*

PAUL NEVAI†, TAMÁS ERDÉLYI‡, AND ALPHONSE P. MAGNUS§

Abstract. The authors obtain upper bounds for Jacobi polynomials which are uniform in all the parameters involved and which contain explicit constants. This is done by a combination of some results on generalized Christoffel functions and some estimates of Jacobi polynomials in terms of Christoffel functions.

Key words. orthogonal polynomials, Christoffel functions, Jacobi weights, Jacobi polynomials

AMS subject classifications. primary 33A65; secondary 26C05, 42C05

1. Introduction.

DEFINITION. *Orthogonal Polynomials.* Given $w(\geq 0) \in L^1(\mathbb{R})$, $p_n(w)$ denotes the corresponding orthonormal polynomial of precise degree n with leading coefficient $\gamma_n(w) > 0$.

DEFINITION. *Jacobi Weights and Jacobi Polynomials.* Given $\alpha > -1$ and $\beta > -1$, w is called a Jacobi weight if $\operatorname{supp}(w) = [-1, 1]$ and $w(x) = (1-x)^\alpha(1+x)^\beta$ for $x \in [-1, 1]$. The corresponding orthogonal polynomials (for historical reasons with various normalizations) are called Jacobi polynomials.

For a wide class of orthogonal polynomials associated with weight functions supported in $[-1, 1]$, the expression

$$\sqrt{\sqrt{1-x^2}w(x)}p_n(w, x)$$

asymptotically equioscillates between $\pm\sqrt{2/\pi}$ for $x \in (-1, 1)$ when n tends to ∞ (cf. [22, Chaps. VIII and X–XII]). Therefore it is natural to seek inequalities for $\sqrt{1-x^2}w(x)p_n^2(w, x)$ for $x \in [-1, 1]$.

Such inequalities for Jacobi polynomials involving optimal constants are truly fascinating. They are easy to prove for the first and second (and third and fourth) kinds of Chebysev polynomials since they are related to simple trigonometric functions. For Legendre polynomials this is somewhat more complicated, and the appropriate inequality was proved by Bernstein (cf. [22, eq. (7.3.8), p. 165] for Bernstein's result and [1] and [14] for a sharper version of it). Bernstein's results can be extended to Jacobi (i.e., ultraspherical or Gegenbauer) polynomials with parameters $-\frac{1}{2} < \alpha = \beta < \frac{1}{2}$ (cf. [22, eq. (7.33.4) and eq. (7.33.5), p. 171] and also [15] for a refinement). In addition, for a wider range of the parameters, similar inequalities have been proved in [13] ($\alpha = \beta > -\frac{1}{2}$) and [7] ($\alpha = \beta > \frac{1}{2}$).

* Received by the editors September 2, 1992; accepted for publication April 13, 1993.

† Department of Mathematics, Ohio State University, 231 West 18th Avenue, Columbus, Ohio 43210-1174 (nevai@math.ohio-state.edu). This material is based upon work supported by National Science Foundation grant DMS–9024901, and by North Atlantic Treaty Orgnaization grant CRG.870806.

‡ Present address, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada. Department of Mathematics, Ohio State University, 231 West 18th Avenue, Columbus, Ohio 43210-1174 (terdelyi@math.ohio-state.edu). This material is based upon work supported by National Science Foundation grant DMS–9024901.

§ Institut de Mathématique Pure et Appliquée, Université Catholique de Louvain, Chemin du Cyclotron, 2, Louvain-la-Neuve, B-1348, Belgium (magnus@anma.ucl.ac.be).

For instance, Lorch [15, form. (10), p. 115] proved[1]

$$\max_{x\in[-1,1]}\left|(1-x^2)^{\frac{\lambda}{2}}P_n^{(\lambda)}(x)\right| \le \frac{2^{1-\lambda}(n+\lambda)^{\lambda-1}}{\Gamma(\lambda)}$$

for $n = 0, 1, \ldots$ and $0 < \lambda < 1$, which, in terms of the orthonormal Jacobi polynomials, can be stated as

$$\max_{x\in[-1,1]}\left|\sqrt{\sqrt{1-x^2}w(x)}\,p_n(w,x)\right| \le \sqrt{\frac{2\Gamma(n+1)}{\pi\Gamma(n+2\alpha+1)}}\left(n+\alpha+\frac{1}{2}\right)^{\alpha}$$

for $n = 0, 1, \ldots$, and $-\frac{1}{2} < \alpha < \frac{1}{2}$, where $w(x) = (1-x^2)^{\alpha}$.

For nonsymmetric Jacobi weights much less is known. In 1988, Gatteschi [10] extended Bernstein's results to Jacobi polynomials with $-\frac{1}{2} < \alpha, \beta < \frac{1}{2}$. For instance, he proved that if $-\frac{1}{2} < \alpha, \beta < \frac{1}{2}$, and $\alpha + \beta > 0$, then[2]

$$\max_{\theta\in[0,\frac{\pi}{2}]}\left|(\sin\theta/2)^{\alpha+\frac{1}{2}}(\cos\theta/2)^{\beta+\frac{1}{2}}P_n^{(\alpha,\beta)}(\cos\theta)\right| \le \frac{\Gamma(\beta+1)}{\Gamma\left(\frac{1}{2}\right)\left(n+\frac{\alpha+\beta+1}{2}\right)^{\beta+\frac{1}{2}}}\binom{n+\beta}{n}$$

for $n = 0, 1, \ldots$.[3] Again, in terms of the orthonormal Jacobi polynomials, this can be stated as

$$\max_{x\in[0,1]}\left|\sqrt{\sqrt{1-x^2}w(x)}\,p_n(w,x)\right| \le \sqrt{\frac{2^{2\beta+1}\Gamma(n+\alpha+\beta+1)\Gamma(n+\beta+1)}{\pi\Gamma(n+1)\Gamma(n+\alpha+1)}}\,(2n+\alpha+\beta+1)^{-\beta}$$

for $n = 0, 1, \ldots$, where $w(x) = (1-x)^{\alpha}(1+x)^{\beta}$ with $-\frac{1}{2} < \alpha, \beta < \frac{1}{2}$, and $\alpha + \beta > 0$.

In a sense our goal is less ambitious than the previously mentioned inequalities in that we do not expect to be able to obtain sharp constants with our techniques. On the other hand, our techniques enable us to extend these Jacobi polynomial inequalities with very explicit constants for all parameters $\alpha \ge -\frac{1}{2}$ and $\beta \ge -\frac{1}{2}$.

DEFINITION. *Generalized Polynomials.* The function f given by

$$f(z) \stackrel{\text{def}}{=} |\omega|\prod_{j=1}^{k}|z-z_j|^{r_j}, \qquad \omega \neq 0,\ z_j \in \mathbb{C},\ z \in \mathbb{C},\ r_j > 0,$$

is called a *generalized* (*nonnegative*) *algebraic polynomial* of (*generalized*) degree

$$N \stackrel{\text{def}}{=} \sum_{j=1}^{k} r_j\,,$$

and we will write $f \in |\,\mathrm{GCAP}\,|_N$.

If $w(x) = (1-x)^{\alpha}(1+x)^{\beta}$ is a Jacobi weight, then $\sqrt{1-x^2}w(x)p_n^2(w,x)$ is a generalized polynomial (of degree $2n+\alpha+\beta+1$), and as such the framework of generalized polynomials is (one of) the perfect setting for studying Jacobi polynomials. As

[1] Here $P_n^{(\lambda)}$ is the standard normalization of the Gegenbauer polynomials, that is, $P_n^{(\lambda)}(1) = \binom{n+2\lambda-1}{n}$ and Γ denotes the gamma function.

[2] Here $P_n^{(\alpha,\beta)}$ is the standard normalization of the Jacobi polynomials, that is, $P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$ and Γ again denotes the gamma function.

[3] For $\theta \in [\frac{\pi}{2}, \pi]$, one can use $P_n^{(\alpha,\beta)}(-x) = (-1)^n P_n^{(\beta,\alpha)}(x)$ to obtain an analogous inequality.

a matter of fact, this was the primary reason for introducing generalized polynomials in the first place (cf. [6] and [5]).

This paper is a modest attempt to demonstrate the applicability of generalized polynomials to problems which have not yet been settled in a satisfactory way despite more than a hundred years of undiminished interest in them.

Our method consists of two steps. First, in §2, we use *generalized polynomials* to estimate the *Christoffel function* $\sum_{k=0}^{n} p_k^2(w)$, and then, in §3, we obtain a Riccati equation which yields estimates for the ratio $p_n^2(w)/\sum_{k=0}^{n} p_k^2(w)$. The reason that we have to limit ourselves to considering $\alpha \geq -\frac{1}{2}$ and $\beta \geq -\frac{1}{2}$ is that the function $\sqrt{1-x^2}w(x)$ for either $\alpha < -\frac{1}{2}$ or $\beta < -\frac{1}{2}$ is no longer a generalized polynomial.

Our main result is the following.

THEOREM 1. *For all Jacobi weight functions* $w(x) = (1-x)^\alpha(1+x)^\beta$ *with* $\alpha \geq -\frac{1}{2}$ *and* $\beta \geq -\frac{1}{2}$, *the inequalities*

$$\max_{x\in[-1,1]} \frac{p_n^2(w,x)}{\sum_{k=0}^{n} p_k^2(w,x)} \leq \frac{4\left(2+\sqrt{\alpha^2+\beta^2}\right)}{2n+\alpha+\beta+2} \tag{1}$$

and

$$\max_{x\in[-1,1]} \sqrt{1-x^2}w(x)p_n^2(w,x) \leq \frac{2e\left(2+\sqrt{\alpha^2+\beta^2}\right)}{\pi} \tag{2}$$

hold for $n = 0, 1, \ldots$.

Our method is therefore able to give $O((\alpha^2+\beta^2)^{1/2})$ estimates for large $\alpha^2+\beta^2$. It is natural to ask how a sharp bound should behave. Numerically computed examples of the actual maximum of $\sqrt{1-x^2}w(x)p_n^2(w,x)$ suggest that small values of n give relevant information. For instance, with $\alpha = 10$ and $\beta = 2$,

n	0	1	2	3	4	5	6	7	8	9	10
max	1.478	1.251	1.191	1.161	0.845	0.747	1.123	0.727	1.112	0.703	0.685.

For $n = 0$, explicit calculation yields

$$\max_{x\in[-1,1]} \sqrt{1-x^2}w(x)p_0^2(w,x) = \frac{\Gamma(\alpha+\beta+2)}{2^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)} \max_{x\in[-1,1]} (1-x)^{\alpha+1/2}(1+x)^{\beta+1/2},$$

that is,[4]

$$\max_{x\in[-1,1]} \sqrt{1-x^2}w(x)p_0^2(w,x) = \frac{(\alpha+1/2)^{\alpha+1/2}(\beta+1/2)^{\beta+1/2}\Gamma(\alpha+\beta+2)}{(\alpha+\beta+1)^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)},$$

which behaves like $[(\alpha+\beta)/(2\pi)]^{1/2}$ for α and β large. We expect $O((\alpha^2+\beta^2)^{1/4})$ bound to be valid for all $n \geq 1$.

2. Generalized Christoffel functions and generalized polynomials.

DEFINITION. *Generalized Christoffel Functions.* Given $w(\geq 0) \in L^1(\mathbb{R})$ and $p \in (0,\infty)$,

$$\lambda_n^*(w,p,z) \stackrel{\text{def}}{=} \inf_{f\in|\,\mathrm{GCAP}\,|_{n-1}} \int_{\mathbb{R}} \frac{f^p(t)}{f^p(z)}\, w(t)\, dt, \qquad z \in \mathbb{C}, \tag{3}$$

[4] The maximum is taken at $x = (\beta-\alpha)/(\alpha+\beta+1)$.

where $n \geq 1$ is *real*, that is, n is not necessarily an integer.

Remark 2. Of course, $\lambda_n^*(w,p) \equiv \lambda_{np-p+1}^*(w,1)$. As a matter of fact, this is one of the underlying reasons for the usefulness of the concept of *generalized polynomials*. The notation $\lambda_n^*(w,p)$ was kept for historical reasons. Eventually, the parameter p may disappear from it.

DEFINITION. *Generalized Jacobi Weights.* Given a nonnegative integer m, the function w satisfying $\operatorname{supp}(w) = [-1,1]$ and

$$w(x) \stackrel{\text{def}}{=} (1-x)^{\tau_0} \prod_{k=1}^{m} |x-a_k|^{\tau_k} (1+x)^{\tau_{m+1}}, \qquad a_k \in \mathbb{R}, \quad \tau_k \in \mathbb{R}, \tag{4}$$

for $x \in [-1,1]$, is called a *generalized Jacobi weight*, and its *degree* is denoted by $\deg(w) \stackrel{\text{def}}{=} \sum_{k=0}^{m+1} \tau_k$.

We start with the following.

THEOREM 3. *Let w be a generalized Jacobi weight of the form* (4) *such that $a_k \neq a_j$ for $k \neq j$, $a_k \neq \pm 1$ for $k = 1,2,\dots,m$, $\tau_0 \geq -\frac{1}{2}$, $\tau_k > 0$ for $k = 1,2,\dots,m$, and $\tau_{m+1} \geq -\frac{1}{2}$. Then, for all $0 < p < \infty$ and $n \geq 1$, the generalized Christoffel functions $\lambda_n^*(w,p)$ satisfy the inequality*

$$\max_{x\in[-1,1]} \sqrt{1-x^2} w(x) \left[\lambda_n^*(w,p,x)\right]^{-1} \leq \frac{(2+pn-p+\deg(w))\, e}{2\pi}.$$

Since the reciprocal of $\sum_{k=0}^{n-1} p_k^2(w,z)$ equals the right-hand side of (3) with the infimum (that is, *minimum*) taken for all ordinary polynomials of degree at most $n-1$,

$$\sum_{k=0}^{n} p_k^2(w,x) \leq \left[\lambda_{n+1}^*(w,2,x)\right]^{-1}, \qquad n = 0,1,\dots,$$

and thus we have the following.

COROLLARY 4. *For all Jacobi weights $w(x) = (1-x)^\alpha (1+x)^\beta$ with $\alpha \geq -\frac{1}{2}$ and $\beta \geq -\frac{1}{2}$,*

$$\max_{x\in[-1,1]} \sqrt{1-x^2} w(x) \sum_{k=0}^{n} p_k^2(w,x) \leq \frac{(2n+\alpha+\beta+2)\, e}{2\pi}, \qquad n = 0,1,\dots, \tag{5}$$

holds.

Remark 5. We point out the uniformity of (5) in all parameters.

Remark 6. The corresponding lower estimates are essentially the same with a proper interpretation of the word "essentially" (cf. [6, Thms. 2.1 and 2.2, p. 113]).

Remark 7. Of course, given $\epsilon > 0$, for all Jacobi weights we have

$$\lim_{n\to\infty} \frac{\sqrt{1-x^2} w(x) \sum_{k=0}^{n} p_k^2(w,x)}{n} = \frac{1}{\pi}$$

uniformly for $-1+\epsilon \leq x \leq 1-\epsilon$ (cf. [19, Thm. 6.2.35, p. 94]).

Question 8. It remains to be seen how to extend (5) for all Jacobi weights, with parameters $\alpha > -1$ and $\beta > -1$.

Proof of Theorem 3. We start out as in the proof of [16, Thm. 6, p. 149], and we follow closely the proof of [6, Thm. 3.2, p. 126]. If h is analytic in the unit disk, then

$$(1-|rz|^2) h(rz) = \frac{1}{2\pi i} \int_{|u|=1} h(u) \frac{1-r\bar{z}u}{u-rz} du, \qquad |z| \leq 1, \quad 0 \leq r < 1.$$

Hence, if P is a polynomial and $0 < p < \infty$, then

$$(1-|r|^2)|P^*(rz)|^p \le \frac{1}{2\pi}\int_{|u|=1}|P^*(u)|^p|du|, \qquad |z|=1, \quad 0\le r\le 1,$$

where P^* is obtained from P by replacing all the zeros z^* of P which are inside the unit disk by $\bar{z^*}^{-1}$.

Since

$$\frac{1+r}{2}|z-\sigma| \le |rz-\sigma|, \qquad |\sigma|\ge 1, \quad |z|=1, \quad 0\le r\le 1,$$

we have

$$(1-|r|^2)\left(\frac{1+r}{2}\right)^{p\deg(P)}|P^*(z)|^p \le \frac{1}{2\pi}\int_{|u|=1}|P^*(u)|^p|du|, \qquad |z|=1, \quad 0\le r\le 1.$$

Maximizing the left-hand side here for $0 \le r \le 1$ and using $|P^*(z)| = |P(z)|$ for $|z| = 1$, the inequality

$$|P(z)|^p \le \frac{(2+p\deg(P))\,e}{8\pi}\int_{-\pi}^{\pi}|P(e^{i\theta})|^p\,d\theta, \qquad |z|=1,$$

follows.

For every real trigonometric polynomial R_n of degree at most n there is an algebraic polynomial $P_{2n} \in \Pi_{2n}$ such that $R_n^2(\theta) = |P_{2n}(e^{i\theta})|^2$. Therefore,

$$\|R_n\|^p_{L^\infty(\mathbb{R})} \le \frac{(1+pn)\,e}{4\pi}\int_{-\pi}^{\pi}|R_n(\theta)|^p\,d\theta \tag{6}$$

for every such trigonometric polynomial R_n.

If the multiplicity of each zero of $g \in |\,\mathrm{GCAP}\,|_N$ is rational, then there is $q > 0$ such that $g^q(\cos\cdot)$ is a nonnegative trigonometric polynomial, so that applying (6) with $R_{Nq} = g^q$ and $p = \frac{1}{q}$ yields

$$\|g(\cos\cdot)\|_{L^\infty(\mathbb{R})} \le \frac{(1+N)\,e}{4\pi}\int_{-\pi}^{\pi}g(\cos\theta)\,d\theta. \tag{7}$$

Once (7) holds for all $g \in |\,\mathrm{GCAP}\,|_N$ such that the multiplicity of each zero of g is rational, by continuity it remains valid for all $g \in |\,\mathrm{GCAP}\,|_N$. Hence,

$$\|G(\cos\cdot)\|_{L^\infty(\mathbb{R})} \le \frac{(1+N)\,e}{4\pi}\int_{-\pi}^{\pi}G(\cos\theta)\,d\theta \quad \forall\, G \in |\,\mathrm{GCAP}\,|_N \tag{8}$$

for every $N \ge 0$.

Thus,

$$\left\|\sqrt{1-(\cdot)^2}F\right\|_{L^\infty([-1,1])} \le \frac{(2+N)\,e}{2\pi}\int_{-1}^{1}F(t)\,dt \quad \forall\, \sqrt{1-(\cdot)^2}F \in |\,\mathrm{GCAP}\,|_{N+1},$$

$N \ge 0$. Applying this inequality with $F = f^p w$, Theorem 3 follows immediately. □

3. Christoffel Functions and Jacobi Polynomials.

"I've tried A! I've tried B! I've tried C!"

Tom Wolfe, *The Right Stuff.*

If we want to find upper bounds for p_n^2 from upper bounds for $\sum_0^n p_k^2$, then we must have upper bounds for $p_n^2/\sum_0^n p_k^2$, and that is precisely what is attempted here.

THEOREM 9. *Given $n = 1, 2, \ldots$, and a Jacobi weight $w(x) = (1-x)^\alpha(1+x)^\beta$ with $\alpha > -1$ and $\beta > -1$, let $x_{nn}(w) < x_{1n}(w)$ be the extreme zeros of the corresponding nth-degree Jacobi polynomial. Then the inequalities*

$$(9)\qquad \frac{p_n^2(w,x)}{\sum_{k=0}^n p_k^2(w,x)} \le \begin{cases} \frac{(2n+\alpha+\beta+1)(\beta+1)}{(n+\alpha+\beta+1)(n+\beta+1)}, & -1 \le x \le 2x_{nn}(w)+1, \\ \frac{4(2n+\alpha+\beta+1)}{(2n+\alpha+\beta+2)^2 - \frac{2\alpha^2}{1-x} - \frac{2\beta^2}{1+x}}, & \xi_1 \le x \le \xi_2, \\ \frac{(2n+\alpha+\beta+1)(\alpha+1)}{(n+\alpha+\beta+1)(n+\alpha+1)}, & 2x_{1n}(w)-1 \le x \le 1, \end{cases}$$

hold provided $-1 < \xi_1 \le \xi_2 < 1$ are such that $(2n+\alpha+\beta+2)^2 - 2\alpha^2/(1-x) - 2\beta^2/(1+x)$ on the right-hand side of the second inequality of (9) is positive in $[\xi_1, \xi_2]$.

Remark 10. The inequality

$$\frac{p_n^2(w,x)}{\sum_{k=0}^n p_k^2(w,x)} \le \frac{\text{const}}{n}, \qquad -1 \le x \le 1,$$

is well known [19, Lem. 6.2.17, p. 82] (see [18, Lem. 2.1, p. 336] for the necessary Christoffel function estimates) but its proof is rather cumbersome. Theorem 1 of the present paper yields a new proof with an explicit formula for the constant which depends on α and β.

Proof of Theorem 9. By Christoffel–Darboux's formula

$$\begin{aligned}\sum_{k=0}^n p_k^2(w,x) &= \frac{\gamma_n(w)}{\gamma_{n+1}(w)}\left[p'_{n+1}(w,x)p_n(w,x) - p'_n(w,x)p_{n+1}(w,x)\right] \\ &= \frac{\gamma_n(w)}{\gamma_{n+1}(w)} p_n^2(w,x)\frac{d}{dx}\left[\frac{p_{n+1}(w,x)}{p_n(w,x)}\right].\end{aligned}$$

Hence, we need to find an appropriate lower bound for r' in $[-1,1]$, where

$$r(x) = \frac{\gamma_n(w)}{\gamma_{n+1}(w)}\frac{p_{n+1}(w,x)}{p_n(w,x)}.$$

Here r is a rational function with simple poles at the zeros $\{x_{kn}(w)\}_{k=1}^n$ of $p_n(w)$. It has an asymptotic behavior $x+c$ for $x \to \infty$, where c is a constant. Since r' is positive everywhere, r must have negative residues at its poles, so that we obtain

$$r(x) = x + c - \sum_{k=1}^n \frac{A_k}{x - x_{kn}(w)} \quad \text{and} \quad r'(x) = 1 + \sum_{k=1}^n \frac{A_k}{(x - x_{kn}(w))^2}$$

with $A_k > 0$, $k = 1, 2, \ldots, n$. See the graphs of r and r' (solid thick line) on upper and lower parts of Fig. 1 as an example. When $-1 \le x < x_{nn}(w)$, r' is the sum of increasing functions of x, and therefore it is greater than $r'(-1)$. When x is slightly greater than $x_{1n}(w)$, r' is decreasing but it is still greater or equal than $r'(-1)$ as long as each term $A_k/(x - x_{kn}(w))^2$ is greater or equal than the corresponding term at -1, that is, $(x - x_{kn}(w))^2 \le (-1 - x_{kn}(w))^2$ for $k = 1, \ldots, n$ or $(x+1)(x - 2x_{kn}(w) - 1) \le 0$, which holds if $-1 \le x \le 2\min x_{kn}(w) + 1 = 2x_{nn}(w) + 1$. A similar argument shows that $r'(x) \ge r'(1)$ if $x \ge 2\max x_{kn}(w) - 1 = 2x_{1n}(w) - 1$. This will prove the first and third inequalities of (9) as soon as we get the actual values of $r'(-1)$ and $r'(1)$.

The bottom part of Fig. 1 shows a graph of r' for $-1 \le x \le 1$ when $\alpha = 3/2$, $\beta = -3/10$, and $n = 3$. A short-dashed horizontal line has been drawn between -1 and $2x_{nn}(w)+1$ at the ordinate $r'(-1)$. One can see that this horizontal line segment

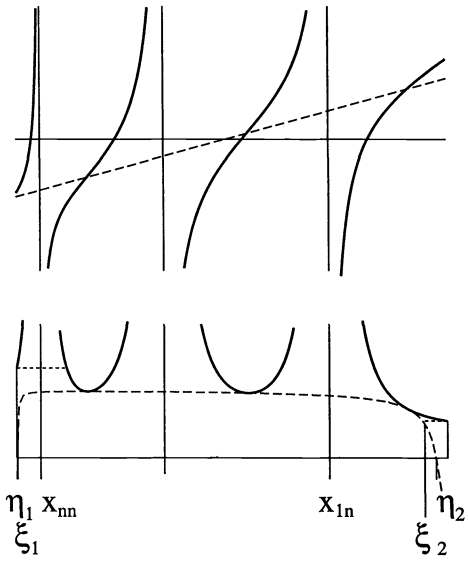


FIG. 1

lies indeed under the graph of r'. As $r'(1)$ is a lower bound of r' on the whole interval $[-1,1]$ in the present case, only a part of the horizontal line of the ordinate $r'(1)$ has been drawn. Other features of this figure will be explained later.

In order to establish the second inequality of (9) and to compute $r'(\pm 1)$, we need the following formulas for the orthonormal Jacobi polynomials.

(i) The recurrence relation:

$$\frac{\gamma_n(w)}{\gamma_{n+1}(w)}p_{n+1}(w,x) = (x - b_n(w))p_n(w,x) - \frac{\gamma_{n-1}(w)}{\gamma_n(w)}p_{n-1}(w,x), \tag{10}$$

where

$$b_n(w) = \frac{\beta^2 - \alpha^2}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)},$$

and

$$\frac{\gamma_n(w)}{\gamma_{n+1}(w)} = 2 \times \sqrt{\frac{(n+1)(n+\alpha+\beta+1)(n+\alpha+1)(n+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)^2(2n+\alpha+\beta+3)}},$$

$n = 1, 2, \ldots$ (cf. [22, form. (4.5.1), p. 71] or [3, Table III.11, p. 220]).

(ii) The differential relation:

$$\begin{aligned}(1-x^2)p_n'(w,x) &= n\left(\frac{\alpha-\beta}{2n+\alpha+\beta} - x\right)p_n(w,x) + (2n+\alpha+\beta+1)\frac{\gamma_{n-1}(w)}{\gamma_n(w)}p_{n-1}(w,x) \\ &= (n+\alpha+\beta+1)\left(\frac{\alpha-\beta}{2n+\alpha+\beta+2} + x\right)p_n(w,x) \\ &\quad - (2n+\alpha+\beta+1)\frac{\gamma_n(w)}{\gamma_{n+1}(w)}p_{n+1}(w,x),\end{aligned} \tag{11}$$

$n = 1, 2, \ldots$ (cf. [22, form. (4.5.7), p. 72]). Of course, each one of these formulas can be deduced from the other one by the three-term recurrence formula.

The combination of (i) and (ii) yields (iii).

(iii) The differential equation:

$$(12)\quad (1-x^2)p_n''(w,x) + [\beta - \alpha - (\alpha+\beta+2)x]p_n'(w,x) + n(n+\alpha+\beta+1)p_n(w,x) = 0$$

(cf. [22, form. (4.2.1), p. 60] or [3, form. (2.20), p. 149]).

In order to compute $r'(\pm 1)$, we proceed as follows. From (11),

$$\frac{\gamma_n(w)}{\gamma_{n+1}(w)}\frac{p_{n+1}(w,\pm 1)}{p_n(w,\pm 1)} = r(\pm 1) = \frac{(n+\alpha+\beta+1)[\alpha-\beta\pm(2n+\alpha+\beta+2)]}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)},$$

and from (12),

$$\frac{p_n'(w,\pm 1)}{p_n(w,\pm 1)} = \frac{n(n+\alpha+\beta+1)}{\pm(\alpha+\beta+2)+\alpha-\beta},$$

so that we have

$$r'(\pm 1) = \frac{\gamma_n}{\gamma_{n+1}}\frac{p_{n+1}(w,\pm 1)}{p_n(w,\pm 1)}\left(\frac{p_{n+1}'(w,\pm 1)}{p_{n+1}(w,\pm 1)} - \frac{p_n'(w,\pm 1)}{p_n(w,\pm 1)}\right),$$

which allows the computation of the requested special values as given in the following table.

(13)

$\boxed{f}$	$\boxed{f(-1)}$	$\boxed{f(1)}$
r	$-\frac{2(n+\alpha+\beta+1)(n+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}$	$\frac{2(n+\alpha+\beta+1)(n+\alpha+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}$
p_n'/p_n	$-\frac{n(n+\alpha+\beta+1)}{2(\beta+1)}$	$\frac{n(n+\alpha+\beta+1)}{2(\alpha+1)}$
r'	$\frac{(n+\alpha+\beta+1)(n+\beta+1)}{(2n+\alpha+\beta+1)(\beta+1)}$	$\frac{(n+\alpha+\beta+1)(n+\alpha+1)}{(2n+\alpha+\beta+1)(\alpha+1)}$

The values of $1/r'(-1)$ and $1/r'(1)$ are used in the right-hand sides of the first and third inequalities of (9).

Now, we come to the second inequality. This one will be established through a *Riccati* equation for r. Use the differential relations (11) for eliminating p_n' and p_{n+1}' in $p_{n+1}'p_n - p_n'p_{n+1}$. This gives an equation in terms of p_{n+1}^2, p_np_{n+1}, and p_n^2, so that after some rather tedious calculations,

$$r'(x) = \frac{A + B(x)r(x) + Cr(x)^2}{1-x^2}$$

with

$$A = (2n+\alpha+\beta+3)\left(\frac{\gamma_n(w)}{\gamma_{n+1}(w)}\right)^2 = 4\frac{(n+1)(n+\alpha+\beta+1)(n+\alpha+1)(n+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)^2},$$

$$B(x) = -(2n+\alpha+\beta+2)x - \frac{\alpha^2-\beta^2}{2n+\alpha+\beta+2}, \qquad C = 2n+\alpha+\beta+1.$$

The idea is that whatever the actual value of r is, $A+Br+Cr^2$ will always be greater than the absolute minimum of this trinomial, that is,

$$r'(x) \geq \frac{4AC - B(x)^2}{4C(1-x^2)} .$$

Equality will occur whenever $r(x)$ is equal to $-B(x)/(2C)$, which happens once between any pair of consecutive zeros of p_n, as can be seen in Fig. 1, where the graphs of r and $-B/(2C)$ are shown in the upper part, and the graphs of r' and its lower bound (dashed line) in the lower part. Working out the numerator yields

$$r'(x) \geq \frac{(2n+\alpha+\beta+2)^2 - \frac{2\alpha^2}{1-x} - \frac{2\beta^2}{1+x}}{4(2n+\alpha+\beta+1)} , \qquad -1 < x < 1,$$

and, thus the theorem follows, as long as x is restricted to an interval $[\xi_1, \xi_2]$ where the above lower bound is positive. □

Combining Theorem 3 (that is, Corollary 4) and Theorem 9, we obtain the following pointwise estimate for the Jacobi polynomials.

THEOREM 11. *For all Jacobi weight functions* $w(x) = (1-x)^\alpha(1+x)^\beta$ *with* $\alpha \geq -\frac{1}{2}$ *and* $\beta \geq -\frac{1}{2}$, *we have*

$$p_n^2(w,x) \leq \frac{2}{\pi\sqrt{1-x^2}w(x)} \frac{e\,(2n+\alpha+\beta+2)(2n+\alpha+\beta+1)}{(2n+\alpha+\beta+2)^2 - \frac{2\alpha^2}{1-x} - \frac{2\beta^2}{1+x}} , \qquad n = 1, 2, \ldots,$$

for $-1 < x < 1$, *as long as the denominator* $(2n+\alpha+\beta+2)^2 - 2\alpha^2/(1-x) - 2\beta^2/(1+x)$ *on the right-hand side is positive. In particular, given* $0 < \epsilon < 1$,

$$p_n^2(w,x) \leq \frac{2}{\pi\sqrt{1-x^2}w(x)} \frac{e}{1 - \frac{2(\alpha^2+\beta^2)}{(2n+\alpha+\beta+2)^2\,\epsilon}}$$

for $-1+\epsilon \leq x \leq 1-\epsilon$ *and* $n > \sqrt{2(\alpha^2+\beta^2)/\epsilon} - (\alpha+\beta)/2 - 1$.

For fixed $x \in (-1,1)$ and $n \to \infty$, this is no more than $e \times (1+o(1))$ times worse than an optimal inequality could be. However, when x is close to ± 1, the parameter n needs to be sufficiently large so that the estimate would become useful. The quest for estimates valid for every $n > 0$ is the subject of the following investigations.

First, we deal with the first and third inequalities in (9).

LEMMA 12. *When* $\alpha \geq -\frac{1}{2}$, $\beta \geq -\frac{1}{2}$, *and* $n > 0$, *we have*

$$\begin{aligned} &\max\left[\frac{(2n+\alpha+\beta+1)(\beta+1)}{(n+\alpha+\beta+1)(n+\beta+1)} , \frac{(2n+\alpha+\beta+1)(\alpha+1)}{(n+\alpha+\beta+1)(n+\alpha+1)} \right] \\ &\qquad \leq \frac{4(1+\max(\alpha,\beta))}{2n+\alpha+\beta+2} . \end{aligned} \tag{14}$$

This is the first instance showing how the right-hand sides of (9) behaves. The factor $2n+\alpha+\beta+2$ has been chosen because it will reappear when (5) is used.

Proof of Lemma 12. First of all, since $t/(n+t)$ is an increasing function for $t > -n$, we only have to consider

$$\frac{(2n+\alpha+\beta+1)(\gamma+1)}{(n+\alpha+\beta+1)(n+\gamma+1)} ,$$

where $\gamma = \max(\alpha,\beta)$. Let $\delta = \min(\alpha,\beta)$. Then, we have to show

$$(2n+\gamma+\delta+2)(2n+\gamma+\delta+1) \leq 4(n+\gamma+1)(n+\gamma+\delta+1),$$

when $\gamma \geq \delta \geq -\frac{1}{2}$ and $n > 0$. This is quite elementary and amounts to

$$2n(2\gamma+1)+(\gamma+\delta+1)(3\gamma-\delta+2) \geq 0,$$

which holds since $\gamma \geq \delta$ and $2\gamma+1 \geq 0$. □

In order to use the second inequality of (9), we must find a valid interval $[\xi_1, \xi_2]$ containing $[2x_{nn}(w)+1, 2x_{1n}(w)-1]$, so that then we would have upper bounds of $p_n^2(w)/\sum_0^n p_k^2(w)$ in the whole interval $[-1,1]$. Since no simple formulas for $x_{nn}(w)$ and $x_{1n}(w)$ are known, we will now find a lower bound η_1 for $x_{nn}(w)$ large enough for allowing $\xi_1 = 2\eta_1 + 1$ to be a valid choice in (9), and, similarly, a sufficiently small upper bound η_2 for $x_{1n}(w)$.

There is much literature on bounds for the zeros of Jacobi polynomials (see, e.g., [22, §§6.2 and 6.21, p. 116–123]), but most are useful only when α and β are between $-\frac{1}{2}$ and $\frac{1}{2}$. For large n, the extreme zeros behave like $-1+j_\beta^2/(2n^2)$ and $1-j_\alpha^2/(2n^2)$, where j_κ denotes the smallest positive zero of the Bessel function J_κ [22, § 8.1, p. 192].

The next theorem gives reasonably satisfactory lower and upper estimates for the zeros of the Jacobi polynomials.

THEOREM 13. *Given $n = 1, 2, \ldots$, the zeros $\{x_{kn}(w)\}_{k=1}^n$ of the nth-degree Jacobi polynomial corresponding to a Jacobi weight $w(x) = (1-x)^\alpha(1+x)^\beta$ with $\alpha \geq -\frac{1}{2}$ and $\beta \geq -\frac{1}{2}$ satisfy*

$$\eta_1 = -1 + \frac{2\beta^2}{N^2} \leq x_{kn}(w) \leq \eta_2 = 1 - \frac{2\alpha^2}{N^2}, \tag{15}$$

$k = 1, 2, \ldots, n$, where $N = 2n+\alpha+\beta+1$.

The proof of this theorem requires the following lemma on oscillations of solutions of differential equations.

LEMMA 14. *Let Y, Z, Y', Z', Y'', Z'', K, and L be continuous functions in the open interval (a,b), with $Y \not\equiv 0$, such that*

$$Y''(x)+K(x)Y(x)=0, \qquad Z''(x)+L(x)Z(x)=0, \qquad x \in (a,b).$$

If

$$\text{(i)} \qquad K(x) \leq L(x), \qquad x \in (a,b),$$

$$\text{(ii)} \qquad Y'(x)Z(x)-Y(x)Z'(x) \to 0, \qquad x \to x_0,$$

where x_0 is one of the endpoints of (a,b), and if $Z(x)$ has no zero in (a,b), then Y has no zero in (a,b) either.

Proof of Lemma 14. The lemma is a variant of the Sturmian comparison theorems for solutions of second-order linear differential equations. It is almost the same as *Szegő's comparison theorem* in #16.626 of the 1980 edition of Gradshteyn and Ryzhik's book [11], coming from Theorem 1.82.1 of Szegő [22 §1.82 p. 19], known as "Sturm's theorem for open intervals"; see also the introduction of [9].[5] Here is a self-contained proof.

Suppose that $Y(x_1) = 0$ for some $x_1 \in (a,b)$. Since the equations are homogeneous, we may assume that $Z(x) > 0$ on (a,b), and $(x_0 - x_1)Y'(x_1) > 0$.[6] Therefore, $Y(x) = \int_{x_1}^x Y'(t)dt > 0$ when x is between x_1 and x_0 and it is sufficiently close to x_1,

[5] We thank our dear friend Luigi Gatteschi for drawing our attention to [9].

[6] N.B. $Y'(x_1)$ must be different from zero, otherwise $Y \equiv 0$.

and

$$Y'(x)Z(x) - Y(x)Z'(x) = Y'(x_1)Z(x_1) + \int_{x_1}^{x} [L(t) - K(t)]Y(t)Z(t)\,dt$$

keeps the sign of $x_0 - x_1$ with an increasing absolute value when x varies from x_1 to x_0 (since $Y(x) = Z(x)\int_{x_1}^{x}(Z(t))^{-2}[Y'(t)Z(t) - Y(t)Z'(t)]dt$ keeps its positive sign), and it cannot vanish when $x \to x_0$. □

Proof of Theorem 13. First, one uses the fact that if all the zeros of a polynomial p_n are real and are contained in an interval (a, b), a smaller interval containing all the zeros is $a - p_n(a)/p_n'(a), b - p_n(b)/p_n'(b)$. This is a well known theorem of the numerical analysis of the Newton–Raphson iteration method (see, for instance, [21, Chap. 9, p. 55]). Hence, by (13),

$$-1 + \frac{2(\beta+1)}{n(n+\alpha+\beta+1)} \le x_{kn}(w) \le 1 - \frac{2(\alpha+1)}{n(n+\alpha+\beta+1)}, \tag{16}$$

$k = 1, 2, \ldots, n$, is valid for every $\alpha > -1$, $\beta > -1$, and $n \ge 1$ (and is *exact* if $n = 1$). However, considering only the upper bound, it behaves like $1 - 2(\alpha+1)/n^2$ for large n, instead of $1 - j_\alpha^2/(2n^2)$, and j_α behaves like $\alpha + (1.855757\ldots)\alpha^{1/3}$ for large α (Tricomi's formula; see [4, p. 60]). Thus, we need better estimates when either n, α, or β are large.

Since $Y(x) = (1-x)^{(\alpha+1)/2}(1+x)^{(\beta+1)/2}p_n(w, x)$ is a solution of $Y'' + KY = 0$ with

$$K(x) = \frac{1-\alpha^2}{4(1-x)^2} + \frac{1-\beta^2}{4(1+x)^2} + \frac{2n(n+\alpha+\beta+1) + (\alpha+1)(\beta+1)}{2(1-x^2)},$$

(cf. [22, form. (4.24.1), p. 67]), we take $Z(x) = (1-x)^{(\widetilde{\alpha}+1)/2}(1+x)^{(\widetilde{\beta}+1)/2}$ (cf. Lemma 14), so that

$$L(x) = \frac{1-\widetilde{\alpha}^2}{4(1-x)^2} + \frac{1-\widetilde{\beta}^2}{4(1+x)^2} + \frac{(\widetilde{\alpha}+1)(\widetilde{\beta}+1)}{2(1-x^2)},$$

and

$$L(x) - K(x) = \frac{\alpha^2 - \widetilde{\alpha}^2}{4(1-x)^2} + \frac{\beta^2 - \widetilde{\beta}^2}{4(1+x)^2} + \frac{(\widetilde{\alpha}+1)(\widetilde{\beta}+1) - (\alpha+1)(\beta+1) - 2n(n+\alpha+\beta+1)}{2(1-x^2)}.$$

Now we turn to the upper bound in (15). Since $L - K$ must be positive in a neighborhood of 1, $\widetilde{\alpha}^2 < \alpha^2$, and since $Y'Z - YZ'$ behaves like $(1-x)^{(\alpha+\widetilde{\alpha})/2}$ near 1, one must have $\widetilde{\alpha} > -\alpha$. This implies that the method will work only when $\alpha > 0$. Let us choose $\widetilde{\alpha} = 0$ and $\widetilde{\beta} = \beta$ so that

$$L(x) - K(x) = \frac{1}{4(1-x^2)}\left[\alpha^2\frac{1+x}{1-x} - 2\alpha(\beta+1) - 4n(n+\alpha+\beta+1)\right],$$

which is positive between $1 - (2\alpha^2/(2n+\alpha)(2n+\alpha+2\beta+2)) < 1 - (2\alpha^2/N^2)$ and 1. Finally, considering that the first upper bound in (16) satisfies $1 - (2(\alpha+1)/n(n+\alpha+\beta+1)) \le 1 - 8(\alpha+1)/N^2$ when α and $\beta \ge -\frac{1}{2}$, and that $8(\alpha+1) > 2\alpha^2$ when $-\frac{1}{2} \le \alpha \le 0$, we conclude that $1 - (2\alpha^2/N^2)$ is a valid upper bound.

For the lower bound in (15) one can use the symmetry property of the Jacobi polynomials $P_n^{(\alpha,\beta)}$ given by $P_n^{(\alpha,\beta)}(-x) = (-1)^n P_n^{(\beta,\alpha)}(x)$. □

It is interesting to compare (15) with formulas given in [17, p. 160] (which is also quoted in [2, p. 1448]) stating that, when α, β, and n tend to ∞ in such a way that $\lim_{n\to\infty} \alpha/N = A$ and $\lim_{n\to\infty} \beta/N = B$, then the zeros remain smaller than $B^2 - A^2 + [(1 - A^2 - B^2)^2 - 4A^2B^2]^{1/2}$, which is indeed smaller than $1 - 2A^2$ (for an alternative proof see [12, Thm. 8, p. 137]).

We will also need the following estimates for the second inequality of (9).

LEMMA 15. *For real α, β, and $N \geq [2(\alpha^2 + \beta^2)]^{1/2}$, the inequality*

$$(N+1)^2 - \frac{2\alpha^2}{1-x} - \frac{2\beta^2}{1+x} \geq CN(N+1), \qquad -1 + \frac{4\beta^2}{N^2} \leq x \leq 1 - \frac{4\alpha^2}{N^2},$$

holds with

$$C = \min\left(\frac{1}{2}, \frac{3}{\sqrt{8(\alpha^2+\beta^2)}}\right).$$

Proof of Lemma 15. First, since $(N+1)^2 - 2\alpha^2/(1-x) - 2\beta^2/(1+x)$ is a concave function of x in $[-1,1]$, we only have to check its values at $-1+4\beta^2/N^2$ and $1-4\alpha^2/N^2$, which are $(N+1)^2 - \alpha^2/(1-2\beta^2/N^2) - N^2/2$ and $(N+1)^2 - N^2/2 - \beta^2/(1-2\alpha^2/N^2)$, respectively. Let $\gamma = \max(|\alpha|, |\beta|)$ and $\delta = \min(|\alpha|, |\beta|)$.[7] Since $\gamma^2/(1-2\delta^2/N^2) \geq \delta^2/(1-2\gamma^2/N^2)$, we have to find a lower bound for

$$F(N) \stackrel{\text{def}}{=} \frac{(N+1)^2 - \dfrac{N^2}{2} - \dfrac{\gamma^2 N^2}{N^2 - 2\delta^2}}{N(N+1)}$$

when $N \geq [2(\alpha^2+\beta^2)]^{1/2}$. Note that

$$F(N) \geq \frac{\dfrac{N^2}{2} + 2N - \dfrac{\gamma^2 N^2}{N^2 - 2\delta^2}}{N(N+1)} = \frac{1}{2} + \frac{\dfrac{3}{2} - \dfrac{\gamma^2 N}{N^2 - 2\delta^2}}{N+1},$$

so that $F(N) \geq \frac{1}{2}$ when $N \to \infty$. $F(N)$ is greater than $\frac{1}{2}$ for all $N \geq [2(\gamma^2+\delta^2)]^{1/2}$ if

$$G(N) \stackrel{\text{def}}{=} \frac{3}{2} - \frac{\gamma^2 N}{N^2 - 2\delta^2} \geq 0$$

for all these values of N, that is, if γ^2 is smaller than the values of the increasing function $3N/2 - 3\delta^2/N$, so that the least value is taken at $N = [2(\gamma^2+\delta^2)]^{1/2}$. This happens when $\alpha^2 + \beta^2 = \gamma^2 + \delta^2 \leq 9/2$, and, hence, the minimum of $F(N)$ is $\frac{1}{2}$ in this case.

When $\alpha^2 + \beta^2 = \gamma^2 + \delta^2 > 9/2$, we only have to search for the negative values of $G(N)$ in $F(N) \geq \frac{1}{2} + G(N)/(N+1)$. We will show that $G(N)/(N+1)$ is an increasing function of N, that is, $-G(N)/(N+1)$ is a positive decreasing function of N. Indeed, $-G(N) = \gamma^2/(N - 2\delta^2/N) - \frac{3}{2}$ is a decreasing function itself, as $N - 2\delta^2/N$ is increasing. Thus, the minimum of $F(N)$ is not smaller than $\frac{1}{2} + G(N)/(N+1) \geq \frac{1}{2} + G(N)/N$ at $N = [2(\gamma^2+\delta^2)]^{1/2}$. This gives $G(N) = \frac{3}{2} - \frac{N}{2}$ and $F(N) \geq 3/\{2[2(\gamma^2+\delta^2)]^{1/2}\} = 3/[8(\alpha^2+\beta^2)]^{1/2}$. □

Now we are ready for the following.

Proof of Theorem 1. First we prove formula (1). Since it is obvious for $n = 0$, we can assume that $n \geq 1$. Let $N = 2n + \alpha + \beta + 1$. If $x \in [-1, -1 + 4\beta^2/N^2] \cup [1 - 4\alpha^2/N^2, 1]$, then (1) follows from Theorem 9, Lemma 12, and Theorem 13. If

[7] N.B. these values γ and δ are not the same as in Lemma 12 if α or β is negative.

$N < [2(\alpha^2+\beta^2)]^{1/2}$, then $[-1+4\beta^2/N^2, 1-4\alpha^2/N^2]$ is empty. If $N \geq [2(\alpha^2+\beta^2)]^{1/2}$ and $x \in [-1+4\beta^2/N^2, 1-4\alpha^2/N^2]$, then (1) follows from Theorem 9 and Lemma 15. Finally, Corollary 4 and (1) yield (2). □

REFERENCES

[1] V. A. ANTONOV AND K. V. HOLŠEVNIKOV, *An estimate of the remainder in the expansion of the generating function for the Legendre polynomials (Generalization and improvement of Bernstein's inequality)*, Vestnik Leningrad Univ. Math., 13 (1981), pp. 163–166.

[2] L. CHEN AND M. E. H. ISMAIL, *On asymptotics of Jacobi polynomials*, SIAM J. Math. Anal., 22 (1991), pp. 1442–1449.

[3] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, London, Paris, 1978.

[4] A. ERDÉLYI WITH W. MAGNUS, F, OBERHETTINGER, AND F. G. TRICOMI, *Bateman Manuscript Project: Higher Transcendental Functions*, McGraw-Hill, New York, 1953.

[5] T. ERDÉLYI, A. MÁTÉ, AND P. NEVAI, *Inequalities for generalized non-negative polynomials*, Constr. Approx., 8 (1992), pp. 241–255.

[6] T. ERDÉLYI AND P. NEVAI, *Generalized polynomial weights, Christoffel functions and zeros of orthogonal polynomials*, J. Approx. Theory, 69 (1992), pp. 111–132.

[7] K.–J. FÖRSTER, *Inequalities for ultraspherical polynomials and application to quadrature*, manuscript, 1991.

[8] G. FREUD, *Orthogonal Polynomials*, Pergamon Press, Oxford, 1971.

[9] L. GATTESCHI, *New inequalities for the zeros of Jacobi polynomials*, SIAM J. Math. Anal., 18 (1987), pp. 1549–1562.

[10] ———, *Bernstein type inequalities for Jacobi polynomials*, written communication, September 14, 1988.

[11] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series, and Products*, Academic Press, New York, 1980.

[12] M. E. H. ISMAIL AND X. LI, *Bounds on the extreme zeros of orthogonal polynomials*, Proc. Amer. Math. Soc., 115 (1992), pp. 131–140.

[13] G. LOHÖFER, *Inequalities for Legendre functions and Gegenbauer functions*, J. Approx. Theory, 64 (1991), pp. 226–234.

[14] L. LORCH, *Alternative proof of a sharpened form of Bernstein's inequality for Legendre polynomials*, Appl. Anal., 14 (1983), pp. 237–240.

[15] ———, *Inequalities for ultraspherical polynomials and the gamma function*, J. Approx. Theory, 40 (1984), pp. 115–120.

[16] A. MÁTÉ AND P. NEVAI, *Bernstein's inequality in L^p for $0 < p < 1$ and $(C,1)$ bounds for orthogonal polynomials*, Ann. Math., 111 (1980), pp. 145–154.

[17] D. S. MOAK, E. B. SAFF, AND R. S. VARGA, *On the zeros of Jacobi polynomials $P_n^{(\alpha_n,\beta_n)}(x)$*, Trans. Amer. Math. Soc., 249 (1979), pp. 159–163.

[18] P. NEVAI, *Orthogonal polynomials on the real line associated with the weight $|x|^\alpha exp(-|x|^\beta)$*, I, Acta Math. Acad. Sci. Hungar., 24 (1973), pp. 335–342.

[19] ———, *Orthogonal Polynomials*, Mem. Amer. Math. Soc., 213, Providence, RI, 1979.

[20] ———, *Géza Freud, orthogonal polynomials and Christoffel functions. A case study*, J. Approx. Theory, 48 (1986), pp. 3–167.

[21] A. M. OSTROWSKI, *Solution of Equations and Systems of Equations*, Academic Press, New York, 1960.

[22] G. SZEGŐ, *Orthogonal Polynomials*, Colloquium Publications, Vol. 23, American Mathematical Society, Providence, RI, 1967.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 615–629, March 1994

025

REGULARIZATION OF NONLINEAR DIFFERENTIAL-ALGEBRAIC EQUATIONS*

ROBERT E. O'MALLEY, JR.† AND LEONID V. KALACHEV‡

Abstract. This paper illustrates how initial value problems for nonlinear differential-algebraic equations can be regularized, i.e., converted to tractable singularly perturbed problems, by appropriate introduction of a small positive parameter ϵ. The corresponding outer limit provides the desired solution, including consistent initial conditions. Examples are given for problems with indices one, two, and three.

1. Introduction: The index-one problem.

Consider the initial value problem

$$\begin{cases} \dot{u} = f(u,v,t), \quad u(0) \text{ prescribed}, \\ 0 = g(u,v,t) \end{cases} \tag{1.1}$$

on some interval $t \geq 0$ consisting of a differential equation for the m-vector u, a corresponding initial value $u(0)$, and an n-dimensional algebraic constraint $g = 0$ corresponding in dimension to the unknown n-vector v. Such semiexplicit differential-algebraic equations (DAEs) naturally arise in describing constrained mechanical systems, electrical circuits, and, indeed, a wide variety of other significant applications. The initial value $u(0)$ may be restricted to be consistent with the constraint, though users and computations typically do not enforce this. Such problems have received considerable attention recently in the numerical literature (cf. [4] and [10]), because they cannot generally be readily integrated using standard codes for ordinary differential equations.

The simplest situation occurs when the Jacobian matrix g_v is nonsingular for *all* arguments. Then, the implicit function theorem implies that $g = 0$ can be locally solved for, say,

$$v = h(u,t). \tag{1.2}$$

Presuming appropriate smoothness and stability hypotheses, a solution of the initial value problem (1.1) then becomes specified through the unique solution of the initial value problem

$$\dot{u} = f(u,h(u,t),t), \qquad u(0) \text{ prescribed}. \tag{1.3}$$

An alternative solution procedure results if we first differentiate $g = 0$ to obtain

$$g_u\dot{u} + g_v\dot{v} + g_t = 0.$$

Solving for $\dot{v}$ then shows that (1.1) is equivalent to integrating the initial value problem for the $m+n$-dimensional system

$$\begin{cases} \dot{u} = f(u,v,t), \\ \dot{v} = -g_v^{-1}(u,v,t)[g_u(u,v,t)f(u,v,t) + g_t(u,v,t)], \end{cases} \tag{1.4}$$

*Received by the editors February 18, 1992; accepted for publication (in revised form) February 12, 1993. This research was supported in part by National Science Foundation grant DMS 9107197.

†Department of Applied Mathematics, University of Washington, Seattle, Washington 98195.

‡Department of Mathematical Sciences, University of Montana, Missoula, Montana 59812.

with any $u(0)$ prescribed and with $v(0)$ being obtained as a solution of the nonlinear equation

$$g(u(0), v(0), 0) = 0 \tag{1.5}$$

at $t = 0$. We observe that the constraint manifold $g = 0$ is invariant during an exact integration of (1.4)–(1.5), although numerical errors may in practice result in a drift away from the manifold. Since $\dot{u}$ and $\dot{v}$ are obtained as functions of u, v, and t after one differentiation of the constraint equation (because g_v remains nonsingular), we say the problem then has index one (cf. [8]).

Another simple way of solving (1.1) when g_v remains nonsingular is to consider the "nearby" singularly perturbed system

$$\begin{cases} \dot{u} = f(u, v, t), \\ -\epsilon g_v(u(0), v(0), 0)\dot{v} = g(u, v, t) \end{cases} \tag{1.6}$$

for small positive values of ϵ. Note that this generalizes the pencil regularization that was classically used for linear problems (cf. [8]). The Tikhonov–Levinson theory (cf. [17]) for singularly perturbed initial value problems states when the unique solution of (1.6) with the prescribed initial vector $u(0)$ and *any* $v(0)$ will have the asymptotic form

$$\begin{cases} u(t, \epsilon) = U_0(t) + O(\epsilon), \\ v(t, \epsilon) = V_0(t) + \beta_0(\tau) + O(\epsilon) \end{cases} \tag{1.7}$$

on bounded subintervals of $t \geq 0$, where $\binom{U_0}{V_0}$ satisfies the reduced problem

$$\begin{cases} \dot{U}_0 = f(U_0, V_0, t), \qquad U_0(0) = u(0), \\ 0 = g(U_0, V_0, t) \end{cases} \tag{1.8}$$

(i.e., the DAE (1.1)), and where the initial layer corrector $\beta_0(\tau)$ will satisfy the layer problem

$$\begin{aligned} -g_v(u(0), v(0), 0)\frac{d\beta_0}{d\tau} &= g(u(0), V_0(0) + \beta_0(\tau), 0), \\ \beta_0(0) &= v(0) - V_0(0) \end{aligned} \tag{1.9}$$

on the stretched interval $\tau = t/\epsilon \geq 0$. A sufficient hypothesis is that the Jacobian

$$-g_v^{-1}(u(0), v(0), 0)g_v(u(0), v, 0) \tag{1.10}$$

will remain a stable matrix for *any* v. Then the solution β_0 of (1.9) will decay to zero as $\tau \to \infty$. If $v(0)$ is consistent with the constraint, i.e., $g(u(0), v(0), 0) = 0$, we will take $V_0(0) = v(0)$, so $\beta_0(\tau) \equiv 0$. Otherwise, the solution (1.7) of (1.6) features an $O(\epsilon)$ thick initial t-layer of nonuniform convergence within which the solution converges to a solution of the DAE (1.1) as $\epsilon \to 0$. We note that the numerical integration of (1.6) for any such $v(0)$ is tractable (cf. [16], [20], and [1]). Our artificial introduction of the parameter ϵ is an example of a Tikhonov regularization (cf. [22]). We note that Rubin and Ungar [18] introduced an analogous approach for constrained mechanical

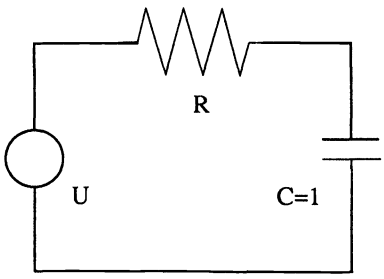


FIG. 1

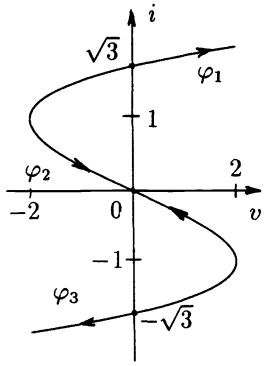


FIG. 2

systems with tangential initial velocities and that Boggs [2] and Boggs and Tolle [3] develop similar approaches for optimization problems. In many physical contexts, such ε parameters can be given natural concrete interpretations (cf. the linear examples in [12] and the circuit theory problem immediately following). Readers are referred to [15] for a survey of related approximate methods to find consistent initial conditions for DAEs. Kopell [14] also presents a related analytic method.

Example. Consider the RC circuit pictured in Fig. 1, with a DC voltage source and a nonlinear resistor in series with a capacitor (see [6] for a discussion of this and more general circuit models). Let i be the current flowing, U be the constant voltage provided, and $3i - i^3$ be the voltage drop across the resistor. Then $U = 3i - i^3 + V$ and $i = \dot{V}$ for a one unit capacitance. Setting $v = V - U$ yields the DAE

$$\begin{cases} \dot{v} = i, \\ 0 = v + 3i - i^3. \end{cases} \tag{1.11}$$

To have a positive voltage drop across the resistor, we'll have to restrict the current so that $0 < i < \sqrt{3}$.

A natural way to solve the DAE is to first solve the constraint for i as a function of v. Graphically, or by use of the implicit function theorem, we'll have one of three possibilities (as shown in Fig. 2):

$$i = \begin{cases} \varphi_1(v) > 1 & \text{if } v \geq -2, \\ \varphi_2(v) & \text{if } -2 \leq v \leq 2, \\ \varphi_3(v) < -1 & \text{if } v \leq 2. \end{cases} \tag{1.12}$$

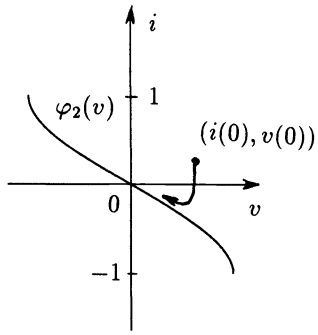


FIG. 3

The sign of $\dot{v}$ in the remaining differential equation

$$\dot{v} = \varphi_j(v) \tag{1.13}$$

implies that $|v(t)|$ will increase monotonically according to the implicit solution

$$t = \int_{v(0)}^{v(t)} \frac{dr}{\varphi_j(r)} \tag{1.14}$$

for $j = 1$ or 3 if $|v(0)| > 2$. Otherwise, for $|v(0)| < 2$, $v \to 0$ monotonically as $t \to \infty$.

Let us, instead, introduce the singular perturbation

$$\begin{cases} \dot{v} = i, \\ -3\epsilon(1 - i^2(0))\dfrac{di}{dt} = v + 3i - i^3 . \end{cases} \tag{1.15}$$

We note that the parameter ϵ corresponds to the introduction of a small inductance of magnitude $3\epsilon(1 - i^2(0))$ into the circuit (cf. [6] or [19]). If $i^2(0) < 1$, the Tikhonov–Levinson result will apply as long as $i^2(t) < 1$, guaranteeing the existence of a limiting solution $\binom{V_0}{I_0}$ for finite t, which satisfies the DAE (1.11). We naturally take

$$I_0 = \varphi_2(V_0), \tag{1.16}$$

for $|V_0| < 2$, so the limiting solution will be determined through the initial value problem

$$\dot{V}_0 = \varphi_2(V_0), \qquad V_0(0) = v(0). \tag{1.17}$$

Because of asymptotic stability, $V_0(t)$ will actually be appropriate for all $t > 0$ (cf. [11]) and the solution will have the asymptotic form

$$\begin{cases} v(t, \epsilon) = V_0(t) + O(\epsilon), \\ i(t, \epsilon) = I_0(t) + \beta_0(\tau) + O(\epsilon), \end{cases} \tag{1.18}$$

where the initial layer correction β_0 is a decaying solution of

$$\frac{d\beta_0}{d\tau} = -\frac{(1 - I_0^2(0))}{1 - i^2(0)}\beta_0 + \frac{\beta_0^2}{3(1 - i^2(0))}(3I_0(0) + \beta_0), \tag{1.19}$$

on $\tau \geq 0$ (cf. Fig. 3). Note that working with the limiting inner solution $\alpha_0(\tau) = I_0(0) + \beta_0(\tau)$ would actually be more convenient to verify the continued existence of $\alpha_0(\tau)$ and $\beta_0(\tau)$.

If instead we had $i(0) > 1$, the Tikhonov–Levinson theory would apply as long as $i^2 > 1$, yielding a limiting solution with current $I_0 = \varphi_1(V_0(t))$ and voltage $V_0(t)$ defined by $\dot{V}_0 = \varphi_1(V_0) > 1$, $V_0(0) = v(0) > -2$. Then, we will have to limit our approximation to finite t intervals to keep solutions bounded. Likewise, a solution can be obtained when $i(0) < -1$ and $v(0) < 2$.

2. The index-two problem. Consider the DAE

$$\begin{cases} \dot{u} = f(u, v, t), \\ 0 = g(u, v, t) \end{cases} \tag{2.1}$$

when the $n \times n$ matrix g_v is singular with constant *positive* rank $r < n$. By possibly reordering the constraints $g = 0$ as

$$g_1(u, v, t) = 0 \quad \text{and} \quad g_2(u, v, t) = 0, \tag{2.2}$$

we can assume that the $r \times n$ matrix g_{1v} has rank r. If we then order the components of v appropriately as

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{2.3}$$

we can even assume that the $r \times r$ matrix

$$\frac{\partial g_1}{\partial v_1} \quad \text{is nonsingular.} \tag{2.4}$$

Then, we will be able to locally solve $g_1 = 0$ for

$$v_1 = h_1(u, v_2, t), \tag{2.5}$$

leaving us the constrained differential system

$$\begin{cases} \dot{u} = \mathcal{F}(u, v_2, t) \equiv f\left(u, \begin{pmatrix} h_1(u, v_2, t) \\ v_2 \end{pmatrix}, t\right), \\ 0 = G(u, t) \equiv g_2\left(u, \begin{pmatrix} h_1(u, v_2, t) \\ v_2 \end{pmatrix}, t\right), \end{cases} \tag{2.6}$$

with the remaining $n - r$ constraints $g_2 = 0$ expressed as $G = 0$. It is important to realize that G is indeed *independent* of v_2, i.e., $\partial G / \partial v_2 \equiv 0$, since otherwise the rank of g_v would exceed its constant rank r. Of course, now we cannot solve the constraint $G = 0$ for v_2, nor can we let $u(0)$ be inconsistent with $G(u(0), 0) = 0$ without expecting an initial discontinuity.

If we now differentiate the constraint in (2.6) and use the differential equation for u, we get the constrained differential system

$$\begin{cases} \dot{u} = \mathcal{F}(u, v_2, t), \\ 0 = G_u(u, t)\mathcal{F}(u, v_2, t) + G_t(u, t). \end{cases} \tag{2.7}$$

If we further assume that the $(n - r) \times (n - r)$ matrix

$$G_u(u, t)\mathcal{F}_{v_2}(u, v_2, t) \quad \text{remains } \textit{nonsingular} \tag{2.8}$$

everywhere (cf., e.g., [7]), (2.7) is an index-one problem since its $n-r$ constraints can be locally uniquely solved for

$$v_2 = h_2(u,t), \tag{2.9}$$

and (2.7) reduces simply to the constrained differential equation

$$\begin{cases} \dot{u} = F(u,t) \equiv \mathcal{F}(u, h_2(u,t), t), \\ 0 = G(u,t), \end{cases} \tag{2.10}$$

consisting of m differential equations and the $n-r$ constraints defined in (2.6). Since we differentiated the constraint $G = 0$ once to get the index-one problem (2.7), we naturally say that the DAE (2.1) has index-two whenever (2.8) holds. We will now further assume that $m > n - r > 0$.

We will now regularize the constrained differential system (2.10) by introducing a small positive parameter ϵ to scale a Lagrange multiplier λ as a slack variable to obtain

$$\begin{cases} \dot{u} = F(u,t) + G_u^T(u,t)\lambda, \\ -\epsilon\lambda = G(u,t)\,. \end{cases} \tag{2.11}$$

Note that Gear [7] effectively uses the Lagrange multiplier λ for $\epsilon = 0$. Introducing ϵ, however, allows us to use inconsistent initial values $u(0)$. If we eliminate λ in (2.11), we obtain

$$\epsilon\dot{u} = -G_u^T(u,t)G(u,t) + \epsilon F(u,t), \tag{2.12}$$

which is a singular singular-perturbation problem (cf. [9] and [23]). Nonetheless we naturally seek an asymptotic solution to the initial value problem for (2.12) in the form

$$u(t,\epsilon) = U(t,\epsilon) + \alpha(\tau,\epsilon), \tag{2.13}$$

where the initial layer correction $\alpha(\tau,\epsilon) \to 0$ as $\tau = t/\epsilon \to \infty$ (cf. [23] or [17]).

Alternatively, we could seek an outer solution

$$U(t,\epsilon) \sim \sum_{j=0}^{\infty} U_j(t)\epsilon^j \tag{2.14}$$

of (2.12) for $t > 0$ and an inner solution

$$\beta(\tau,\epsilon) \equiv U(\epsilon\tau,\epsilon) + \alpha(\tau,\epsilon) \sim \sum_{j=0}^{\infty} \beta_j(\tau)\epsilon^j \tag{2.15}$$

that satisfies (2.12) for all finite τ and coincides asymptotically with $U(\epsilon\tau,\epsilon)$ as $\tau \to \infty$.

To obtain the outer solution, we successively equate coefficients of like powers of ϵ in (2.12). Thus, U_0 must satisfy the limiting equation

$$G_u^T(U_0,t)G(U_0,t) = 0.$$

Because the $m \times (n-r)$ matrix G_u^T has rank $n-r$, multiplication by $G_u(U_0,t)$ implies that U_0 must satisfy the expected constraint

$$G(U_0,t) = 0. \tag{2.16}$$

This determines a locally unique solution $U_0(t)$ when $m = n-r$. For $m > n-r$, it only restricts U_0 to a manifold. The coefficient of ϵ implies that U_1 must satisfy the linear equation

$$-G_u^T(U_0,t)G_u(U_0,t)U_1 = \dot{U}_0 - F(U_0,t) \tag{2.17}$$

(since U_0 satisfies (2.16)). Multiplying (2.17) from the left by $G_u(U_0,t)$ provides $G_u(U_0,t)U_1$ in terms of U_0. Since differentiation of (2.16) implies that $G_u(U_0,t)\dot{U}_0 = -G_t(U_0,t)$, (2.17) finally provides a differential equation

$$\begin{aligned}\dot{U}_0 =& [I - G_u^T(U_0,t)(G_u(U_0,t)G_u^T(U_0,t))^{-1}G_u(U_0,t)]F(U_0,t) \\ & - G_u^T(U_0,t)(G_u(U_0,t)G_u^T(U_0,t))^{-1}G_t(U_0,t)\end{aligned} \tag{2.18}$$

for the desired solution $U_0(t)$ of the DAE (2.1). We note that a differential equation for U_1 will follow analogously using the $O(\epsilon^2)$ terms in (2.12). Equation (2.18) must be solved subject to an initial condition that satisfies the constraint $G = 0$ at $t = 0$. We will get the initial value $U_0(0)$ by matching the limiting inner and outer solutions, i.e., by taking

$$U_0(0) = \lim_{\tau \to \infty} \beta_0(\tau) \tag{2.19}$$

presuming this limit exists.

The leading term of the inner solution $\beta_0(\tau)$ must naturally satisfy the m-dimensional nonlinear initial value problem

$$\frac{d\beta_0}{d\tau} = -G_u^T(\beta_0,0)G(\beta_0,0), \qquad \beta_0(0) = u(0). \tag{2.20}$$

If $G(u(0),0) = 0$, we naturally take $\beta_0(\tau) \equiv u(0)$, so we obtain the consistent initial value $U_0(0) = u(0)$. More generally, any rest point $\beta_0(\infty) = U_0(0)$ will lie on the constraint $G(\beta_0(\infty),0) = 0$, presuming β_0 exists for all $\tau \geq 0$. The equation for β_1 will be a linearized version of (2.20).

Observe that the limiting Jacobian matrix $G_u^T(\beta_0(\infty),0)G_u(\beta_0(\infty),0)$ for (2.20) has $n-r$ stable eigenvalues and $m-(n-r)$ trivial eigenvalues. Reordering the components of β_0, if necessary, let us partition β_0 after its first $n-r$ rows as

$$\beta_0 = \begin{pmatrix} \beta_{01} \\ \beta_{02} \end{pmatrix} \tag{2.21}$$

and *assume* that the $(n-r)$-dimensional stable manifold for (2.20) can be described in the form

$$\beta_{02} = \gamma(\beta_{01}) \tag{2.22}$$

(cf. the analogous use of a *dynamic* manifold in [13] and of invariant manifolds in [5]). Obtaining an explicit representation (2.22) is, admittedly, a difficult problem in

general. If we let $H_1(\beta_{01})$ represent the first $n-r$ components of $-G_u^T G$ evaluated along the stable manifold, (2.20) and (2.22) imply that β_{01} will necessarily have to satisfy the initial value problem

$$\frac{d\beta_{01}}{d\tau} = H_1(\beta_{01}), \qquad \beta_{01}(0) = u_1(0). \tag{2.23}$$

Here, the representation (2.22) restricts the initial value $u(0)$ to lie on the corresponding $(n-r)$-dimensional stable manifold. Since the motion of β_0 is restricted to this stable manifold during integration, the limit at infinity, $\beta_0(\infty) = U_0(0)$ will necessarily satisfy both the $m-(n-r)$-dimensional restriction

$$\beta_{02}(\infty) = \gamma(\beta_{01}(\infty))$$

and the $(n-r)$-dimensional outer constraint

$$G\begin{pmatrix} \beta_{01}(\infty) \\ \gamma(\beta_{01}(\infty)) \end{pmatrix} = 0.$$

The specific limit $\beta_{01}(\infty)$ might be obtained through numerical integration of (2.23) on $\tau \geq 0$ or by solving the last $n-r$ equations. This matching procedure thereby defines a consistent initial m-vector $U_0(0)$.

Example. Consider the DAE

$$\begin{cases} \dot{u} = f_1(u,v,w) + uz, \\ \dot{v} = f_2(u,v,w) + vz, \\ \dot{w} = f_3(u,v,w) + wz, \\ u^2 + v^2 + w^2 = 1, \end{cases} \tag{2.24}$$

consisting of three scalar differential equations and a scalar algebraic constraint. Though this particular problem is unmotivated, similar examples result from constrained mechanics. If we differentiate the constraint with respect to t, we obtain

$$2u(f_1 + uz) + 2v(f_2 + vz) + 2w(f_3 + wz) = 0.$$

This allows us to eliminate

$$z = -uf_1 - vf_2 - wf_3 \tag{2.25}$$

and obtain the vector system

$$\begin{cases} \dot{p} = A(p)f(p), \\ 0 = p^T p - 1 \end{cases} \tag{2.26}$$

for

$$p = \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \qquad f = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix},$$

$$A(p) = \begin{pmatrix} 1-u^2 & -uv & -uw \\ -uv & 1-v^2 & -vw \\ -uw & -vw & 1-w^2 \end{pmatrix}.$$

To solve the constrained system (2.26), we might use the Lagrange multiplier $\lambda = -\frac{1}{\epsilon}(p^T p - 1)$ as in (2.11) or directly consider the regularized problem

$$\epsilon \dot{p} = \epsilon A(p) f(p) - 2p(p^T p - 1). \tag{2.27}$$

We naturally seek an asymptotic solution of the corresponding initial value problem in the form

$$p(t, \epsilon) = P(t, \epsilon) + \alpha(\tau, \epsilon), \tag{2.28}$$

where $\alpha \to 0$ as $\tau = t/\epsilon \to \infty$. Necessarily, the outer solution $P(t, \epsilon)$ will satisfy the equation

$$2P(P^T P - 1) = \epsilon(A(P) f(P) - \dot{P}) \tag{2.29}$$

as a power series in ϵ. Thus, the leading term must satisfy the equation $P_0(P_0^T P_0 - 1) = 0$. We reject the trivial solution since it does not satisfy the expected constraint $P_0^T P_0 = 1$. Note that the latter implies that $\dot{P}_0^T P_0 = 0$. The coefficient of ϵ in (2.29) then implies that $4P_0 P_0^T P_1 = A(P_0) f(P_0) - \dot{P}_0$. Multiplying by P_0^T allows us to eliminate P_1 and to finally obtain the differential equation

$$\dot{P}_0 = (I - P_0 P_0^T) A(P_0) f(P_0) \tag{2.30}$$

for the solution P_0 of the DAE (2.26). The constraint will be an invariant manifold for solutions of (2.30).

The corresponding inner solution

$$\Pi(\tau, \epsilon) = P(\epsilon\tau, \epsilon) + \alpha(\tau, \epsilon) \tag{2.31}$$

will necessarily satisfy the initial value problem

$$\frac{d\Pi}{d\tau} = -2\Pi(\Pi^T \Pi - 1) + \epsilon A(\Pi) f(\Pi), \qquad \Pi(0, \epsilon) = p(0) \tag{2.32}$$

as a power series in ϵ. Its leading term Π_0 will satisfy the vector equation

$$\frac{d\Pi_0}{d\tau} = -2\Pi_0(\Pi_0^T \Pi_0 - 1) \tag{2.33}$$

so the squared norm $D_0 = \Pi_0^T \Pi_0$ will satisfy the scalar equation

$$\frac{dD_0}{d\tau} = -4D_0(D_0 - 1).$$

An explicit integration yields

$$D_0(\tau) = 1 + \frac{e^{-4\tau}(\|p(0)\|^2 - 1)}{\|p(0)\|^2 + e^{-4\tau}(1 - \|p(0)\|^2)}$$

and, through the resulting linear equation for Π_0,

$$\Pi_0(\tau) = \sqrt{D_0(\tau)}\ \frac{p(0)}{\|p(0)\|}. \tag{2.34}$$

Thus, we obtain the consistent initial vector

$$P_0(0) = \Pi_0(\infty) = \frac{p(0)}{\|p(0)\|}. \tag{2.35}$$

This allows the DAE (2.24) to be solved by integrating the vector system (2.30) for P_0 using a unit initial vector in the direction of the arbitrary $p(0)$. The resulting initial impulse corresponds to an initial boundary layer.

3. Some index-three examples.

(i) Consider the initial value problem

$$\begin{cases} \dot{u} = u(w^2-1) + v, u(0) = 1, \\ \dot{v} = v(w^2-1) - u, v(0) = 2, \\ \dot{w} = u + v - z, w(0) = 2, \\ \text{with} \quad u^2 + v^2 = 1. \end{cases} \tag{3.1}$$

Since the initial conditions are inconsistent with the constraint, the solution is expected to exhibit an initial impulse. Differentiating the constraint provides a *hidden* constraint

$$w^2 = 1, \tag{3.2}$$

since $u\dot{u} + v\dot{v} = (u^2+v^2)(w^2-1) = w^2 - 1 = 0$. (Such hidden constraints are common in higher index problems; cf. [12].) Then $\dot{w} = 0$ implies that we also have the second hidden constraint

$$u + v - z = 0\,. \tag{3.3}$$

Eliminating z and using (3.2), we shall now consider the constrained differential system

$$\begin{cases} \dot{u} = v, \\ \dot{v} = -u, \\ \dot{w} = 0, \\ \text{with } u^2 + v^2 = 1, \quad w^2 = 1, \end{cases} \tag{3.4}$$

together with the prescribed initial conditions, which are inconsistent with both constraints.

We'll now introduce scalars λ and μ as Lagrange multipliers and a small positive parameter ϵ to scale them as slack variables to provide us with the regularized problem

$$\begin{cases} \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} v \\ -u \end{pmatrix} + \begin{pmatrix} 2u \\ 2v \end{pmatrix} \lambda + 0\mu, \\ \dot{w} = 0\lambda + 2w\mu, \\ -\epsilon\lambda = u^2 + v^2 - 1, \\ -\epsilon\mu = w^2 - 1. \end{cases} \tag{3.5}$$

Alternatively, eliminating the multipliers implies the singularly perturbed initial value problem

$$\begin{cases} \epsilon\dot{u} = -2u(u^2+v^2-1)+\epsilon v, & u(0)=1, \\ \epsilon\dot{v} = -2v(u^2+v^2-1)-\epsilon u, & v(0)=2, \\ \epsilon\dot{w} = -2w(w^2-1), & w(0)=2. \end{cases} \tag{3.6}$$

It is elementary to solve this system directly by first solving the equations for w and for $r=\sqrt{u^2+v^2}$. Since $\dot{w}=O(1/\epsilon)$ is negative for $w>1$, it follows that

$$w = 1+\gamma_0(\tau), \tag{3.7}$$

where γ_0 is the unique monotonically decreasing solution of the initial value problem

$$\frac{d\gamma_0}{d\tau} = -2\gamma_0(1+\gamma_0)(2+\gamma_0), \qquad \gamma_0(0)=1.$$

(An implicit solution is easily obtained by separating variables.) Clearly, γ_0 decays exponentially to zero as $\tau\to\infty$. Likewise, the initial value problem

$$\epsilon\frac{dr}{dt} = -2r(r^2-1), \qquad r(0)=\sqrt{5} \tag{3.8}$$

has a unique solution of the form

$$r(t,\epsilon) = 1+\beta_0(\tau), \tag{3.9}$$

where β_0 also decays exponentially to zero as $\tau\to\infty$. Since $-2(u^2+v^2-1)=\frac{\epsilon}{r}\dot{r}$, the equations for u and v can be rewritten as $\left(\frac{u}{r}\right)^{\cdot}=\frac{v}{r}$ and $\left(\frac{v}{r}\right)^{\cdot}=-\frac{u}{r}$. Thus,

$$\frac{u}{r} = \frac{u(0)}{r(0)}\cos t + \frac{v(0)}{r(0)}\sin t \tag{3.10}$$

and (3.9) yields

(3.11)
$$u(t,\epsilon)=\frac{1}{\sqrt{5}}(1+\beta_0(\tau))(\cos t+2\sin t) \quad \text{and} \quad v(t,\epsilon)=\frac{1}{\sqrt{5}}(1+\beta_0(\tau))(2\cos t-\sin t).$$

Thus, u and v have the asymptotic form

$$u(t,\epsilon)=U_0(t)+\alpha(\tau,\epsilon), \qquad v(t,\epsilon)=V_0(t)+\delta(\tau,\epsilon) \tag{3.12}$$

with the outer solution

$$U_0(t)=\frac{1}{\sqrt{5}}(\cos t+2\sin t), \qquad V_0(t)=\frac{1}{\sqrt{5}}(2\cos t-\sin t) \tag{3.13}$$

and an initial layer correction

$$\begin{pmatrix}\alpha(\tau,\epsilon)\\ \delta(\tau,\epsilon)\end{pmatrix}$$

that decays exponentially to zero like $\beta_0(\tau)$. Without the regularization (3.5), it would be quite difficult to predict the appropriate initial vector $(U_0(0), V_0(0), W_0(0)) = \frac{1}{\sqrt{5}}(1, 2, \sqrt{5})$ with which to begin a numerical integration of the original DAE.

(ii) The motion of a spherical pendulum with unit mass and length is described by the DAE

$$\begin{cases} \ddot{x} = \lambda x, \\ \ddot{y} = \lambda y, \\ \ddot{z} = \lambda z - g, \\ x^2 + y^2 + z^2 = 1, \end{cases} \tag{3.14}$$

where λ represents the tension and g is acceleration of gravity (cf. [21]). We obtain the additional hidden constraints

$$\begin{gathered} x\dot{x} + y\dot{y} + z\dot{z} = 0, \\ \dot{x}^2 + \dot{y}^2 + \dot{z}^2 + x\ddot{x} + y\ddot{y} + z\ddot{z} = v^2 + \lambda(x^2 + y^2 + z^2) - gz = 0, \end{gathered} \tag{3.15}$$

by differentiating the constraint twice. Introducing v^2 as the square of the speed, we use the latter equation to eliminate

$$\lambda = gz - v^2. \tag{3.16}$$

Thus, there remains the constrained differential system

$$\begin{cases} \ddot{x} = (gz - v^2)x, \\ \ddot{y} = (gz - v^2)y, \\ \ddot{z} = -g(1 - z^2) - v^2 z, \\ x^2 + y^2 + z^2 = 1, \\ x\dot{x} + y\dot{y} + z\dot{z} = 0. \end{cases} \tag{3.17}$$

Note that one might attempt to solve this problem subject to arbitrary initial position and velocity vectors that could fail to satisfy either constraint.

The geometry suggests that it would be natural to introduce spherical coordinates using

$$x = r\cos\varphi\sin\theta, \quad y = r\sin\varphi\sin\theta, \quad \text{and} \quad z = r\cos\theta. \tag{3.18}$$

After some nontrivial manipulations, we are able to express (3.17) as the equivalent problem

$$\begin{cases} \ddot{r} = (r^2 - 1)(g\cos\theta - r\dot{\theta}^2 - r\dot{\varphi}^2\sin^2\theta) - \dot{r}^2 r, \\ r\sin\theta\ddot{\varphi} = 2\dot{\varphi}(\dot{r}\sin\theta + r\dot{\theta}\cos\theta), \\ r\ddot{\theta} = r\dot{\varphi}^2\sin\theta\cos\theta - 2\dot{r}\dot{\theta} + g\sin\theta, \\ r(t) = 1, \\ \dot{r}(t) = 0. \end{cases} \tag{3.19}$$

To allow arbitrary initial values for r and $\dot{r}$, we could naturally write the second-order equation for r as a first-order system for r and $u = \dot{r}$. If we then introduced scaled slack variables/Lagrange multipliers as $-\varepsilon\xi = r - 1$ and $-\varepsilon\eta = u$, we'd get the regularized problem

$$\text{(3.20)} \qquad \begin{cases} \varepsilon\dot{r} = -(r-1) + \varepsilon u, \\ \varepsilon\dot{u} = -u - \varepsilon(r^2-1)(g\cos\theta - r\dot{\theta}^2 - r\dot{\varphi}^2\sin^2\theta) - \varepsilon r u^2, \\ r\sin\theta\ddot{\varphi} = 2\dot{\varphi}(u\sin\theta + r\dot{\theta}^2\cos\theta), \\ r\ddot{\theta} = r\dot{\varphi}^2\sin\theta\cos\theta - 2u\dot{\theta} + g\sin\theta. \end{cases}$$

We'll solve this system as a singularly perturbed initial value problem with $r(0)$, $u(0) = \dot{r}(0)$, $\varphi(0)$, $\dot{\varphi}(0)$, $\theta(0)$, and $\dot{\theta}(0)$ *arbitrarily* prescribed except that $r(0)\sin\theta(0) \neq 0$. We naturally seek an asymptotic solution in the form

$$\text{(3.21)} \qquad \begin{cases} r(t,\varepsilon) = 1 + \alpha(\tau,\varepsilon), \\ u(t,\varepsilon) = \beta(\tau,\varepsilon), \\ \varphi(t,\varepsilon) = \Phi(t,\varepsilon) + \varepsilon^2\gamma(\tau,\varepsilon), \\ \theta(t,\varepsilon) = \Theta(t,\varepsilon) + \varepsilon^2\delta(\tau,\varepsilon), \end{cases}$$

where we ask that the functions of $\tau = t/\varepsilon$ decay to zero as $\tau \to \infty$ in order to provide an initial layer within which the solution of the regularized problem moves rapidly toward the constraint manifold where $r \equiv 1$. This ansatz follows from the usual Tikhonov–Levinson theory (cf. [17]) once one notes that $R(t,\varepsilon) \equiv 1$ in the outer solution for r, to which $U(t,\varepsilon) \equiv 0$ corresponds.

The consistent outer solution

$$\text{(3.22)} \qquad (R(t,\varepsilon), U(t,\varepsilon), \Phi(t,\varepsilon), \Theta(t,\varepsilon)) \equiv (1, 0, \Phi(t,\varepsilon), \Theta(t,\varepsilon))$$

must satisfy the system (3.20) as a power series in ε. Thus, $\Theta(t,0) = \Theta_0(t)$ and $\Phi(t,\varepsilon) = \Phi_0(t)$ must satisfy the coupled system

$$\text{(3.23)} \qquad \begin{cases} \sin\Theta_0\ddot{\Phi}_0 = 2\cos\Theta_0\dot{\Theta}_0\dot{\Phi}_0, \\ \ddot{\Theta}_0 = \left[\dot{\Phi}_0^2\cos\Theta_0 + g\right]\sin\Theta_0, \end{cases}$$

with $\Phi_0(0) = \varphi(0)$, $\dot{\Phi}_0(0) = \dot{\varphi}(0)$, $\Theta_0(0) = \theta(0)$, and $\dot{\Theta}_0(0) = \dot{\theta}(0)$. Multiplying the first equation by $\sin\Theta_0$ and integrating implies that

$$\text{(3.24)} \qquad \dot{\Phi}_0\sin^2\Theta_0 = \dot{\phi}(0)\sin^2\theta(0) \equiv C.$$

Physically, this means that the angular momentum about the z axis is conserved. Substituting for $\dot{\Phi}_0$ in the second equation and integrating the resulting equation for Θ_0 implies the conservation of energy statement

$$\text{(3.25)} \qquad \dot{\Theta}_0^2 = -\frac{C^2}{\sin^2\Theta_0} - 2g\cos\Theta_0 + E_0,$$

where the constant E_0 is specified by the initial conditions. More explicit results follow, as usual, in terms of elliptic integrals.

The stretched system satisfied by the initial layer correction $(\alpha, \beta, \varepsilon^2\gamma, \varepsilon^2\delta)$ follows in a straightforward fashion. In particular, its initial terms must be decaying solutions of the limiting problem

$$\begin{cases} \dfrac{d\alpha_0}{d\tau} = -\alpha_0, \qquad \alpha_0(0) = r(0) - 1, \\ \dfrac{d\beta_0}{d\tau} = -\beta_0, \qquad \beta_0(0) = \dot{r}(0), \\ (1 + \alpha_0(\tau)) \sin\theta(0) \dfrac{d^2\gamma_0}{d\tau^2} = 2\dot{\varphi}(0) \left[\beta_0 \sin\theta(0) + \alpha_0 \dot{\theta}^2(0) \cos\theta(0)\right], \\ (1 + \alpha_0(\tau)) \dfrac{d^2\delta_0}{d\tau^2} = \alpha_0 \dot{\varphi}^2(0) \sin\theta(0) \cos\theta(0) - 2\beta_0 \dot{\theta}(0). \end{cases} \tag{3.26}$$

Thus

$$\alpha_0(\tau) = e^{-\tau}(r(0) - 1) \tag{3.27}$$

will provide an initial impulse for $r(t)$ if $r(0) \neq 1$;

$$\beta_0(\tau) = e^{-\tau}\dot{r}(0) \tag{3.28}$$

will provide an initial impulse for $\dot{r}(t)$ if $\dot{r}(0) \neq 0$; and they together determine the less important terms

$$\begin{cases} \gamma_0(\tau) = 2\dot{\varphi}(0) \displaystyle\int_\tau^\infty \int_r^\infty e^{-s} \left[(r(0) - 1)\dot{\theta}^2(0) \cos\theta(0) + \dot{r}(0) \sin\theta(0)\right] \\ \qquad\qquad \Big/ (1 + e^{-s}(r(0) - 1)) \, ds \, dr \text{ and} \\ \delta_0(\tau) = \displaystyle\int_\tau^\infty \int_r^\infty e^{-s} \left[(r(0) - 1)\dot{\varphi}^2(0) \sin\theta(0) \cos\theta(0) - 2\dot{r}(0)\dot{\theta}(0)\right] \\ \qquad\qquad \Big/ (1 + e^{-s}(r(0) - 1)) \, ds \, dr. \end{cases} \tag{3.29}$$

We note that the asymptotic structure (3.21) of the solution in spherical coordinates directly determines that in the original Cartesian coordinates as well.

Acknowledgment. This work is written in appreciation of Professors Frank W. J. Olver and Richard A. Askey, especially for their valuable and ongoing contributions to this journal.

REFERENCES

[1] R. C. AIKEN, ED., *Stiff Computation*, Oxford University Press, Oxford, 1985.

[2] P. T. BOGGS, *An algorithm, based on singular perturbation theory, for ill-conditioned minimization problems*, SIAM J. Numer. Anal., 14 (1977), pp. 830–843.

[3] P. T. BOGGS AND J. W. TOLLE, *Singular perturbation techniques for nonlinear optimization problems*, in Ottimizzazione non lineare e applicazioni, S. Incerti and G. Treccani, Pitagora Editrice, eds., Bologna, Italy, 1980, pp. 1–32.

[4] K. E. BRENAN, S. L. CAMPBELL, AND L. R. PETZOLD, *Numerical Solution of Initial Value Problems in Differential-Algebraic Equations*, North-Holland, Amsterdam, 1989.

[5] K. W. CHEUNG AND J. H. CHOW, *Stability analysis of singularly perturbed systems via integral manifolds*, 1991.

[6] L. O. CHUA AND A.-C. DENG, *Impasse Points. Part* I: *Numerical Aspects*, Internat. J. Circuit Theory Appl., 17 (1989), pp. 213–235.

[7] C. W. GEAR, *Differential-algebraic equation index transformations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 39–47.

[8] C. W. GEAR AND L. R. PETZOLD, *Differential-algebraic systems and matrix pencils*, Springer-Verlag, Berlin, Lecture Notes in Math. 973, 1983, pp. 75–89.

[9] Z.-M. GU, N. N. NEFEDOV, AND R. E. O'MALLEY, JR., *On singular singularly perturbed initial value problems*, SIAM J. Appl. Math., 49 (1989), pp. 1–25.

[10] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations* II, *Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.

[11] F. C. HOPPENSTEADT, *Singular perturbations on the infinite interval*, Trans. Amer. Math. Soc., 123 (1966), pp. 521–535.

[12] L. V. KALACHEV AND R. E. O'MALLEY, JR., *The regularization of linear differential-algebraic equations*, Tech. Rep., Dept. of Applied Mathematics, University of Washington, Seattle, WA, 1991.

[13] P. V. KOKOTOVIC, H. K. KHALIL, AND J. O'REILLY, *Singular Perturbation Methods in Control*, Academic Press, London, 1986.

[14] N. KOPELL, *Invariant manifolds and the initialization problem for some atmospheric equations*, Physica, D14 (1985), pp. 203–215.

[15] B. LEIMKUHLER, L. R. PETZOLD, AND C. W. GEAR, *Approximation methods for the consistent initialization of differential-algebraic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 205–226.

[16] W. MIRANKER, *Numerical Methods for Stiff Equations*, Reidel, Dordrecht, The Netherlands, 1981.

[17] R. E. O'MALLEY, JR., *Singular Perturbation Methods for Ordinary Differential Equations*, Springer-Verlag, New York, 1991.

[18] H. RUBIN AND P. UNGAR, *Motion under a strong constraining force*, Comm. Pure Appl. Math., 10 (1957), pp. 65–78.

[19] S. SASTRY AND C. A. DESOER, *Jump behavior of circuits and systems*, IEEE Trans. Circuits and Systems, 28 (1981), pp. 1109–1124.

[20] G. SÖDERLIND, *Theoretical and Computational Aspects on Partitioning in the Numerical Integration of Stiff Differential Systems*, Tech. Rep., Dept. of Numerical Analysis and Computing Science, The Royal Institute of Technology, Stockholm, 1981.

[21] A. SOMMERFELD, *Mechanics*, Academic Press, New York, 1950.

[22] A. N. TIKHONOV AND V. Y. ARSENIN, *Methods of Solving Ill-posed Problems*, John Wiley, New York, 1977.

[23] A. B. VASIL'EVA AND V. F. BUTUZOV, *Singularly Perturbed Equations in the Critical Case*, Moscow State University, Moscow, 1978. (Translated as Report No. 2039, Mathematics Research Center, University of Wisconsin, Madison, 1980.)

[24] ———, *Asymptotic Methods in Singular Perturbation Theory*, Vysshaya Shkola, Moscow, 1990. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 630–645, March 1994

026

A GENERALIZATION OF PEARCEY'S INTEGRAL*

R. B. PARIS†

Abstract. This paper considers the special case of a cuspoid canonical diffraction integral consisting of only two parameters X, Y in the form

$$\int_{-\infty}^{\infty} \exp[i(u^{2m} + Xu^m + Yu)]du,$$

where m is a positive integer with $m \geq 2$. This integral is a reduced form of the general cuspoid canonical diffraction integral involving $2m-1$ stationary phase points, and represents a generalization of the familiar Pearcey integral with $m = 2$. The analytic continuation of this integral to arbitrary complex values of X, Y is considered, and its asymptotic behavior when either $|X|$ or $|Y| \to \infty$ is determined.

Key words. asymptotic expansions, Pearcey integral, caustics

AMS subject classifications. 41A60, 30E15, 33A70

1. Introduction.

We consider a generalization of the Pearcey integral defined by

$$P'_m(X,Y) = \int_{-\infty}^{\infty} \exp[i(u^{2m} + Xu^m + Yu)]du, \tag{1.1}$$

where X and Y are real variables and m is an integer with $m \geq 2$. The particular case $m = 2$, corresponding to the familiar Pearcey integral, arises in many physical problems involving short wavelength phenomena, such as wave propagation and optical diffraction; we refer to [4], [9], [12], and the references therein for a summary of the recent literature on Pearcey's integral.

The Pearcey integral belongs to the family of canonical oscillatory integrals that are classified according to the hierarchy introduced to describe the types of singularities arising in catastrophe theory. For one-dimensional integrals, the exponents are the polynomial transformations associated with the so-called cuspoid catastrophes [3]. The general cuspoid canonical diffraction integral has the form

(1.2)

$$I_n(X_1, X_2, \ldots, X_{n-2}) = \int_{-\infty}^{\infty} \exp[if_n(u; \underline{X})]du, \qquad f_n(u; \underline{X}) = u^n + \sum_{k=1}^{n-2} X_k u^k,$$

where $\underline{X} = (X_1, X_2, \ldots, X_{n-2})$ with the parameters X_k being real, and is associated with $n-1$ stationary phase or saddle points. The simplest case $n = 3$ corresponds to the fold catastrophe and involves two stationary points whose positions depend on the single parameter X_1. The canonical integral in this case is the well-known Airy function. The next integral in the hierarchy has $n = 4$ and corresponds to the cusp catastrophe, which involves three coalescing stationary points and two real parameters, X_1 and X_2. The canonical form of this integral is Pearcey's integral. The cases $n = 5$ and $n = 6$ correspond to the so-called swallowtail and butterfly integrals and involve, respectively, four and five stationary points.

* Received by the editors May 6, 1992; accepted for publication (in revised form) June 3, 1993.
† Department of Mathematical and Computer Sciences, Dundee Institute of Technology, Dundee DD1 1HG, United Kingdom.

Comparison with (1.2) shows that $P'_m(X,Y)$ is a reduced case of $I_{2m}(X_1, X_2, \ldots, X_{2m-2})$, with $X_1 = Y$, $X_m = X$ and all other parameters equal to zero. In particular, when $m = 3$, we have the reduced butterfly integral $P'_3(X,Y) \equiv I_6(Y,0,X,0)$. The utility of integrals of the type (1.2) is hampered by the lack of detailed information concerning the numerical values of these functions and the complexity of their large parameter behavior. A considerable effort has been made over the last decade in an attempt to rectify these deficiencies. Extensive work concerning the numerical evaluation of the Pearcey and the swallowtail integrals has been described by Connor and Curtis [4], [5] and Connor, Curtis, and Farelly [6]. Recent asymptotic investigations have been undertaken by Stamnes and Spjelkavik [15], Kaminski [9], Paris [12] for the Pearcey integral, and Kaminski [10] for the swallowtail integral (see also [17, p. 389]).

In this paper we shall be concerned with the analytic continuation of $P'_m(X,Y)$ to complex values of X and Y. This is achieved by rotation of the path of integration in (1.1) through an angle of $\pi/4m$ and appeal to Jordan's lemma to obtain

$$P'_m(X,Y) \equiv P_m(x,y) = e^{\pi i/4m} \int_{-\infty}^{\infty} \exp[-t^{2m} - xt^m + iyt]dt, \tag{1.3}$$
$$x = Xe^{-\pi i/4}, \qquad y = Ye^{\pi i/4m},$$

where we have put $u = t\exp(\pi i/4m)$. We consider the case of even and odd m separately and define the integrals

$$\left\{ \begin{array}{c} I_m(x,y) \\ J_m(x,y) \end{array} \right\} = \int_0^{\infty} e^{-t^{2m}-xt^m} \left\{ \begin{array}{c} \cos yt \\ \sin yt \end{array} \right\} dt \tag{1.4}$$

so that

$$P_m(x,y) = 2e^{\frac{\pi i}{4m}} I_m(x,y) \qquad (m \text{ even}), \tag{1.5}$$

$$P_m(x,y) = e^{\frac{\pi i}{4m}} [\{I_m(x,y) + iJ_m(x,y)\} + \{I_m(-x,y) - iJ_m(-x,y)\}] \qquad (m \text{ odd}). \tag{1.6}$$

It is readily established from (1.3) that $P_m(x,y)$ satisfies the symmetry and conjugacy relations

$$P_m(x,y) = P_m((-)^m x, -y),$$
$$e^{\pi i/4m}\overline{P_m(x,y)} = e^{-\pi i/4m} P_m((-)^m \overline{x}, \overline{y}), \tag{1.7}$$

where the bar denotes the complex conjugate.

Following the method described in [12] for the Pearcey integral, we shall obtain the asymptotic behavior of $P_m(x,y)$ for complex variables when $|x|$ or $|y| \to \infty$ by means of an integral representation involving a Weber parabolic cylinder function. This approach permits the determination of the asymptotics without reference to the above-mentioned stationary points. We remark that this approach has also recently been employed by Janssen [8], who has considered a different generalization of Pearcey's integral in the form

$$\int_0^{\infty} \exp[i(u^4 + Xu^2)] J_v(uY) u^{v+1} du \qquad (-1 < v < 5/2),$$

where J_v denotes the Bessel function of order v. This integral, which equals a multiple of Pearcey's integral when $v = -\frac{1}{2}$, occurs in the problem of image formation in high resolution electron microscopes when $v = 0$.

2. Preliminaries. If we denote the phase function in (1.1) by $f(u;X,Y) \equiv u^{2m} + Xu^m + Yu$, then the $2m-1$ stationary points are given by $\partial f/\partial u = 0$, or

$$u^{2m-1} + \frac{1}{2}Xu^{m-1} + \frac{Y}{2m} = 0. \tag{2.1}$$

For real X, Y, application of Descartes's rule shows that there cannot be more than three real roots of (2.1), the remaining roots being complex conjugate pairs (although when $Y = 0$ one of these roots is a multiple root of order $m-1$ at $u = 0$). For certain values of X and Y two of these real roots can coalesce to form a double root. This occurs when $\partial f/\partial u = \partial^2 f/\partial u^2 = 0$ and corresponds to values of X, Y lying on the caustic

$$Y^m + m\left(\frac{m-1}{2}\right)^{m-1}\left(\frac{mX}{2\mu}\right)^{2\mu} = 0, \qquad \mu = m - \frac{1}{2} \quad (m = 2,3,\ldots). \tag{2.2}$$

In the case $m = 2$, this yields the familiar cusped caustic $Y^2 + (\frac{2}{3}X)^3 = 0$ for the Pearcey integral. For m even the caustic in (2.2) is cusped and symmetrical about the negative X-axis, while for m odd the caustic is asymmetrical (see Fig. 1). Inside the shaded domain in the X,Y-plane bounded by the caustic (and, in the case m odd, by the line $Y = 0$) there are three real, distinct stationary points which will contribute to the asymptotic behavior of $P'_m(X,Y)$. Outside of this domain there is only one real stationary point, but some complex stationary points may also contribute (cf. the case $m = 2$).

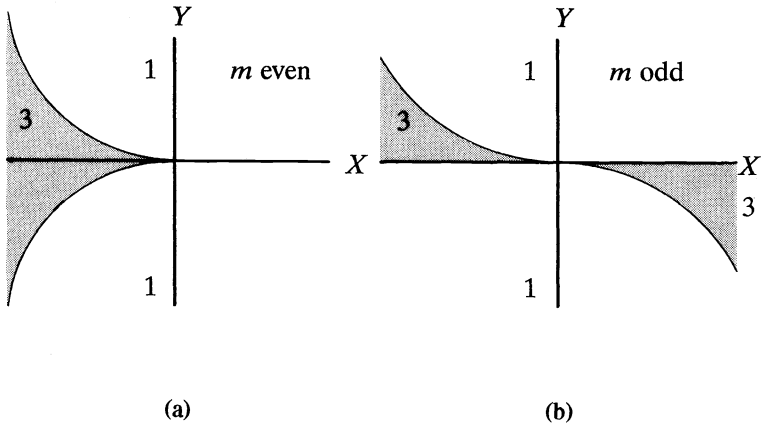


FIG. 1. *The caustic* (2.2) *in the* X,Y*-plane for* (a) m *even and* (b) m *odd. The figures indicate the number of real, distinct stationary points.*

For large values of X, Y inside the caustic the asymptotic behavior of $P'_m(X,Y)$ can be obtained by the method of stationary phase as a sum of three terms. Explicit formulation of these terms, however, is rendered difficult by the fact that the real roots of (2.1) for general m do not seem to be expressible in closed form. The situation for large $|X|$ on the caustic (2.2) is more tractable since the double root can be expressed simply as $u_0 = [(m-1)|X|/4\mu]^{1/m}$. With the new integration variable $\tau = u/u_0$, we then have, on the caustic (when $X < 0$),

$$P'_m(X,Y) = u_0 \int_{-\infty}^{\infty} \exp[iX^{*2}F(\tau)]d\tau, \qquad X^* = \frac{(m-1)}{4\mu}|X|,$$

where

$$F(\tau) = \tau^{2m} - \frac{4\mu}{m-1}\tau^m + \frac{2m^2}{m-1}\tau.$$

The phase function $F(\tau)$ possesses a double positive stationary point at $\tau = 1$ and one negative stationary point, which we define by $\tau = -k$, $k > 0$. Straightforward application of the method of stationary phase (see [11, p. 96] and [17, p. 79]) then shows that for large $|X|$ on the caustic

(2.3)

$$P'_m(X,Y) = \frac{\pi^{1/2}}{m}\alpha(k)(X^*)^{\frac{1}{m}-1}\exp i\left[\beta(k)X^{*2} + \frac{\pi}{4}\right]\left\{1 + O(X^{*-1})\right\}$$
$$+ 3^{-\frac{1}{6}}(2m^2\mu)^{-\frac{1}{3}}(X^*)^{\frac{1}{m}-\frac{2}{3}}\exp\left[2i\mu X^{*2}\right]\left\{\Gamma\left(\frac{1}{3}\right) - \frac{i(3m-5)\Gamma\left(\frac{2}{3}\right)}{(144m^2\mu)^{\frac{1}{3}}}X^{*-\frac{2}{3}} + O\left(X^{*-\frac{4}{3}}\right)\right\},$$

where

$$\alpha(k) \equiv \left[\frac{F''(-k)}{2m^2}\right]^{-\frac{1}{2}} = \frac{k}{\sqrt{k+k^{2m}}}, \qquad \beta(k) \equiv F(-k) = -(2mk + k^{2m}).$$

It does not appear possible to determine the negative stationary point $\tau = -k$ in a simple closed form, except when $m = 2$ and $m = 3$, when $k = 2$ and $k = \frac{2}{3}\left(\frac{5}{4}\right)^{1/3}\left[\left(\frac{5}{4}\right)^{1/3} + \left(\frac{5}{4}\right)^{-1/3} - 1\right] = 0.7221199\ldots$, respectively. Values of k, together with the corresponding values of $\alpha(k)$ and $\beta(k)$, are tabulated in Table 1 for $m = 2$ to $m = 10$. When $m = 2$, equation (2.3) reduces to the asymptotic form obtained in [9] and [12] for the Pearcey integral on the caustic $Y^2 + \left(\frac{2}{3}X\right)^3 = 0$.

TABLE 1
Values of k, $\alpha(k)$, and $\beta(k)$ for different values of m.

m	k	$\alpha(k)$	$-\beta(k)$
2	2.000000	0.471405	24.000000
3	0.722120	0.776916	4.474513
4	1.308203	0.416058	19.043922
5	0.829336	0.836366	8.447287
6	1.181562	0.403243	21.582945
7	0.876927	0.861572	12.436018
8	1.128631	0.397539	24.989666
9	0.903778	0.875506	16.429858
10	1.099584	0.394313	28.668478

In this paper, we shall not be concerned with the asymptotic form of $P'_m(X,Y)$ in the neighborhood of the caustic (2.2), but with the asymptotics of $P_m(x,y)$ for large $|x|$ or $|y|$. We follow the procedure described in [12] for the Pearcey integral to transform the integrals for $I_m(x,y)$ and $J_m(x,y)$ in (1.4) into loop integrals involving a Weber function. To do this, we employ the Mellin–Barnes integral representations, valid for all arg$z$:

$$\left\{ \begin{array}{c} \cos z \\ \sin z \end{array} \right\} = \frac{1}{2\pi i} \int_C \Gamma(s) \left\{ \begin{array}{c} \cos \frac{1}{2}\pi s \\ \\ \sin \frac{1}{2}\pi s \end{array} \right\} z^{-s} ds \qquad (z \neq 0), \tag{2.4}$$

where throughout C will denote a loop that starts and finishes at $-\infty$ and encircles $s = 0$ in the positive sense. Upon reversal of the order of integration, when $\mathrm{Re}(s) < 1$, we then find

$$I_m(x,y) = \frac{1}{2\pi i} \int_C \Gamma(s) \cos \tfrac{1}{2}\pi s \, y^{-s} \left\{ \int_0^\infty e^{-t^{2m} - xt^m} t^{-s} dt \right\} ds \qquad (y \neq 0).$$

The inner integral (with $\tau = t^m$) can be evaluated in terms of the parabolic cylinder function $D_v(z)$, which admits the integral representation [7, p. 119]

$$D_v(z) = \frac{e^{-z^2/4}}{\Gamma(-v)} \int_0^\infty \exp\left(-\tfrac{1}{2}\tau^2 - z\tau\right) \tau^{-v-1} d\tau, \qquad \mathrm{Re}(v) < 0.$$

Similar arguments applied to $J_m(x,y)$ lead to the representations when $\mathrm{Re}(s) < 1$,

$$\left\{ \begin{array}{c} I_m(x,y) \\ J_m(x,y) \end{array} \right\} = \frac{2^{-\frac{1}{2m}}}{m} e^{x^2/8} \frac{1}{2\pi i} \int_C \Gamma(s) \Gamma\left(\frac{1}{m} - \frac{s}{m}\right) \cdot \left\{ \begin{array}{c} \cos \frac{1}{2}\pi s \\ \\ \sin \frac{1}{2}\pi s \end{array} \right\} D_{\frac{s}{m} - \frac{1}{m}}\left(\frac{x}{\sqrt{2}}\right) \left(y 2^{-\frac{1}{2m}}\right)^{-s} ds, \tag{2.5}$$

where the loop C described above separates the poles of $\Gamma(s)$ and $\Gamma(1/m - s/m)$. These representations will form the basis of the present investigation.

From the asymptotic behavior of $D_v(z)$ for fixed z and large $|v|$ (see (4.4)), the convergence of the integrals in (2.5) is seen to be controlled by the quotient of gamma functions $\Gamma(s)\Gamma(1/m - s/m)\Gamma(1/2m + \frac{3}{4} - s/2m)$. Application of Stirling's formula then shows that as $|s| \to \infty$ on C the integrands in (2.5) are dominated by the term $\exp[-(1 - 1/2m)\ln|s|]$, so that the right-hand sides of (2.5) converge for all finite complex values of x and y ($\neq 0$).

We remark that when $m = 2$, use of the properties of the gamma function, followed by replacement of the variable s by $2s$, shows that (1.5) and (2.5) reduce to the alternative representation for Pearcey's integral given in [12]:

$$P_2(x,y) = 2^{-\frac{1}{4}} \pi^{\frac{1}{2}} e^{\frac{x^2}{8} + \frac{\pi i}{8}} \frac{1}{2\pi i} \int_C \Gamma(s) D_{s-\frac{1}{2}}\left(\frac{x}{\sqrt{2}}\right) (y^2/4\sqrt{2})^{-s} ds.$$

3. Asymptotics of $P_m(x,y)$ for large $|x|$. We derive the asymptotic expansions of $I_m(x,y)$ and $J_m(x,y)$ for large $|x|$, y finite. The analysis we present is formal, since no attempt is made here to discuss the asymptotic nature of the various expansions; this has been seen for the Pearcey integral case $m = 2$ in [12]. We substitute the expansion for large $|z|$ of the parabolic cylinder function [16, p. 347], [12]

$$D_v(z) \sim z^v e^{-z^2/4} \sum_{r=0}^\infty \frac{(-v)_{2r}}{r!} (-2z^2)^{-r}, \qquad |\arg z| < \frac{1}{2}\pi, \tag{3.1}$$

where $(a)_r = \Gamma(a+r)/\Gamma(a)$, into (2.5a), to find, when $|\arg x| < \frac{1}{2}\pi$,

$$I_m(x,y) \sim \frac{x^{-\frac{1}{m}}}{m}\frac{1}{2\pi i}\int_C \Gamma(s)\cos\frac{1}{2}\pi s\left(yx^{-\frac{1}{m}}\right)^{-s}\sum_{r=0}^{\infty}\frac{\Gamma\left(\frac{1}{m}-\frac{s}{m}+2r\right)}{r!(-x^2)^r}ds$$

$$= \pi^{\frac{1}{2}}\frac{x^{-\frac{1}{m}}}{m}\sum_{r=0}^{\infty}\frac{(-x^2)^{-r}}{r!}E_r(\chi;m). \tag{3.2}$$

The coefficient functions $E_r(\chi;m)$ $(r=0,1,2,\ldots)$ are defined by

(3.3)

$$E_r(\chi;m) = \frac{\pi^{-1/2}}{2\pi i}\int_C 2^{-s}\Gamma(s)\Gamma\left(\frac{1}{m}-\frac{s}{m}+2r\right)\cos\frac{1}{2}\pi s\,\chi^{-s/m}ds, \qquad \chi \equiv (y/2)^m/x,$$

where, as in (2.5), C is a loop with endpoints at $-\infty$ enclosing the poles of $\Gamma(s)$. Straightforward evaluation of the residues at $s = 0, -2, -4, \ldots$ then leads to the result

$$E_r(\chi;m) = \pi^{-1/2}\sum_{k=0}^{\infty}\frac{(-)^k}{(2k)!}\Gamma\left(\frac{1}{m}+\frac{2k}{m}+2r\right)\left(2\chi^{\frac{1}{m}}\right)^{2k}$$

$$= \sum_{k=0}^{\infty}\frac{(-)^k}{k!}\frac{\Gamma\left(\frac{1}{m}+\frac{2k}{m}+2r\right)}{\Gamma\left(k+\frac{1}{2}\right)}\chi^{2k/m}. \tag{3.4}$$

In a similar manner, we find from (2.5b)

$$J_m(x,y) \sim \frac{\pi^{1/2}}{m}x^{-\frac{1}{m}}\sum_{r=0}^{\infty}\frac{(-x^2)^{-r}}{r!}F_r(\chi;m), \qquad |\arg x| < \tfrac{1}{2}\pi, \tag{3.5}$$

where the coefficient functions $F_r(\chi;m)$ are defined as at (3.3) with the term $\cos\frac{1}{2}\pi s$ replaced by $\sin\frac{1}{2}\pi s$, so that

$$F_r(\chi;m) = \sum_{k=0}^{\infty}\frac{(-)^k}{k!}\frac{\Gamma\left(\frac{2}{m}+\frac{2k}{m}+2r\right)}{\Gamma\left(k+\frac{3}{2}\right)}\chi^{(2k+1)/m}. \tag{3.6}$$

An alternative representation involving derivatives of the coefficient function corresponding to $r=0$ can be found in the form

$$E_r(\chi;m) = \chi^{1-\frac{1}{m}}\frac{d^{2r}}{d\chi^{2r}}\left\{\chi^{2r-1+\frac{1}{m}}E_0(\chi;m)\right\}, \qquad r = 1,2,\ldots$$

with an analogous expression for $F_r(\chi,m)$. The coefficient functions $E_r(\chi;m)$ and $F_r(\chi;m)$ are generalized hypergeometric functions, uniformly and absolutely convergent for all finite values of χ. We note at this point that as $|x| \to \infty$, $\chi \to 0$ and

$$E_r(\chi;m) = \pi^{-\frac{1}{2}}\Gamma\left(2r+\frac{1}{m}\right)\left[1+O\left(x^{-\frac{2}{m}}\right)\right],$$

$$F_r(\chi;m) = \pi^{-\frac{1}{2}}\Gamma\left(2r+\frac{2}{m}\right)yx^{-\frac{1}{m}}\left[1+O\left(x^{-\frac{2}{m}}\right)\right].$$

We remark that in the case of Pearcey's integral when $m = 2$, the coefficients $E_r(\chi; m)$ reduce to a particularly simple form since, from (3.4),

$$E_0(\chi; 2) = e^{-\chi}, \qquad \chi \equiv y^2/4x.$$

The coefficients in this case are then given by

$$E_r(\chi; 2) = \chi^{1/2} \frac{d^{2r}}{d\chi^{2r}} \left[\chi^{2r-1/2} e^{-\chi}\right] \equiv a_r(\chi) e^{-\chi},$$

where $a_r(\chi)$ can be expressed as a Hermite polynomial

$$a_r(\chi) = \pi^{-1/2}\Gamma\left(2r + \frac{1}{2}\right) {}_1F_1\left(-2r; \frac{1}{2}; \chi\right) = 2^{-4r} H_{4r}(\sqrt{\chi}).$$

The domain of validity (in the Poincaré sense) of the expansions (3.2) and (3.5) can be extended to the wider sector $|\arg x| < \frac{3}{4}\pi$ by applying a further transformation to the integrals in (2.5) to make the specific dependence on the parameter group $\chi = (\frac{1}{2}y)^m/x$ apparent; cf. [12, §4a]. We employ the Mellin–Barnes representation of $D_v(z)$ [1, p. 688]

$$D_v(z) = \frac{z^v e^{-z^2/4}}{\Gamma(-v)} \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \Gamma(t)\Gamma(-2t - v)(2z^2)^t dt, \qquad |\arg z| < \frac{3}{4}\pi,$$

where, provided v is not a nonnegative integer, the path of integration is indented at $t = 0$ to separate the poles of $\Gamma(t)$ from those of $\Gamma(-2t - v)$. Substitution of this result in (2.5a), followed by reversal of the order of integration, then yields the alternative representation of $I_m(x, y)$ in the form

$$I_m(x, y) = \frac{\pi^{\frac{1}{2}}}{m} x^{-1/m} \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \Gamma(t) E_{-t}(\chi; m) x^{2t} dt. \tag{3.7}$$

where $E_{-t}(\chi; m)$ is defined as in (3.3) and (3.4) (with r replaced by t) and the path is indented at $t = 0$ in order to lie to the right of the poles of $\Gamma(t)$. It can be established that as $t \to \pm i\infty$, $|E_{-t}(\chi, m)| = e^{-\pi|t|} \exp[O(|\chi t|^{1/m})]$ (we omit the details), so that (3.7) defines $I_m(x, y)$ only in the sector $|\arg x| < \frac{3}{4}\pi$. Application of the standard method for asymptotically evaluating Mellin–Barnes integrals [14, p. 143] then yields the expansion given in (3.2), valid in the wider sector $|\arg x| < \frac{3}{4}\pi$.

By means of similar arguments applied to $J_m(x, y)$, we consequently obtain the results

$$I_m(x, y) \sim \frac{\pi^{\frac{1}{2}}}{m} x^{-\frac{1}{m}} S_1^{(c)}(x, y), \qquad J_m(x, y) \sim \frac{\pi^{\frac{1}{2}}}{m} x^{-\frac{1}{m}} S_1^{(s)}(x, y) \tag{3.8}$$

as $|x| \to \infty$ in $|\arg x| < \frac{3}{4}\pi$, where the formal asymptotic sums $S_1^{(c,s)}(x, y)$ are defined by

$$S_1^{(c)}(x, y) = \sum_{r=0}^{\infty} \frac{(-x^2)^{-r}}{r!} E_r(\chi; m), \qquad S_1^{(s)}(x, y) = \sum_{r=0}^{\infty} \frac{(-x^2)^{-r}}{r!} F_r(\chi; m). \tag{3.9}$$

When $|\arg(-x)| < \frac{1}{2}\pi$, we use the connection formula

$$D_v(-z) = e^{\pm\pi i v} D_v(z) \pm i \frac{(2\pi)^{1/2}}{\Gamma(-v)} e^{\pm\pi i v/2} D_{-v-1}(\pm iz), \tag{3.10}$$

where, in order to apply (3.1), the signs will be chosen such that both arguments of the parabolic cylinder functions lie between $\pm\frac{1}{2}\pi$. This means that we select the upper sign in (3.10) when $\frac{1}{2}\pi < \arg x < \pi$ and the lower sign when $-\pi < \arg x < -\frac{1}{2}\pi$. Substitution of (3.10) in (2.5a) then yields

$$I_m(x,y) = e^{\mp\frac{\pi i}{m}} I_m\left(e^{\mp\pi i}x, e^{\mp\frac{\pi i}{m}}y\right) + \frac{2^{-\frac{1}{2m}}}{m} e^{\frac{x^2}{8}} K, \tag{3.11}$$

where, according to the above ranges of $\arg x$,

$$K = \pm i(2\pi)^{1/2}\frac{e^{\mp\frac{\pi i}{2m}}}{2\pi i}\int_C \Gamma(s)\cos\frac{1}{2}\pi s D_{\frac{1}{m}-\frac{s}{m}-1}\left(e^{\mp\pi i/2}\frac{x}{\sqrt{2}}\right)\left(e^{\mp\frac{\pi i}{2m}}y2^{-\frac{1}{2m}}\right)^{-s} ds.$$

The expansion of the first term on the right-hand side of (3.11) in $|\arg(-x)| < \frac{1}{2}\pi$ follows from (3.8) and (3.9) as

$$I_m\left(e^{\mp\pi i}x, e^{\mp\frac{\pi i}{m}}y\right) \sim \frac{\pi^{1/2}}{m}\left(e^{\mp\pi i}x\right)^{-\frac{1}{m}} S_1^{(c)}(x,y). \tag{3.12}$$

The expansion of the second term can be obtained by substitution of (3.1) into the integral for K, as at (3.2), to find

$$K \sim (2\pi)^{1/2}e^{x^2/8}\left(e^{\mp\pi i}\frac{x}{\sqrt{2}}\right)^{\frac{1}{m}-1}\sum_{r=0}^{\infty}\frac{x^{-2r}}{r!}c_r(\xi),$$

where the coefficients $c_r(\xi)$ are defined by

$$\begin{aligned} c_r(\xi) &= \frac{1}{2\pi i}\int_C \Gamma(s)\cos\frac{1}{2}\pi s\left(1+\frac{s}{m}-\frac{1}{m}\right)_{2r}\xi^{-\frac{s}{m}}ds, \qquad \xi \equiv \frac{1}{2}y^m(e^{\mp\pi i}x) \\ &= \xi^{2r+1-\frac{1}{m}}\frac{d^{2r}}{d\xi^{2r}}\left(\frac{1}{2\pi i}\int_C \Gamma(s)\cos\frac{1}{2}\pi s\,\xi^{-\frac{s}{m}+\frac{1}{m}-1}ds\right) \\ &= \xi^{2r+1-\frac{1}{m}}\frac{d^{2r}}{d\xi^{2r}}\left(\xi^{\frac{1}{m}-1}\cos\xi^{\frac{1}{m}}\right) \qquad (r=0,1,2,\ldots) \end{aligned}$$

upon making use of (2.4). We write

$$C_r(\xi) = (-)^r m^{2r}\xi^{-2r/m}c_r(\xi),$$

so that $C_r(\xi)$ involves a finite series of descending powers of $\xi^{1/m}$ to find

$$C_r(\xi) = P_r(\xi)\cos\xi^{\frac{1}{m}} - Q_r(\xi)\sin\xi^{\frac{1}{m}}, \tag{3.13}$$

where the coefficients $P_r(\xi)$ and $Q_r(\xi)$ are given by

$$\begin{aligned} &P_0(\xi) = 1, \qquad Q_0(\xi) = 0, \\ &P_1(\xi) = 1-(m-1)(2m-1)\xi^{-\frac{2}{m}}, \qquad Q_1(\xi) = 3(m-1)\xi^{-\frac{1}{m}}, \\ &P_2(\xi) = 1-5(m-1)(7m-5)\xi^{-\frac{2}{m}} + (m-1)(2m-1)(3m-1)(4m-1)\xi^{-\frac{4}{m}}, \\ &Q_2(\xi) = 10(m-1)\xi^{-\frac{1}{m}} - 5(m-1)(2m-1)(5m-3)\xi^{-\frac{3}{m}}, \ldots. \end{aligned}$$

Then, as $|x| \to \infty$ in $|\arg(-x)| < \frac{1}{2}\pi$, we have

$$K \sim (2\pi)^{1/2}e^{x^2/8}\left(e^{\mp\pi i}\frac{x}{\sqrt{2}}\right)^{\frac{1}{m}-1} S_2^{(c)}(x,y),$$

where the formal asymptotic sum $S_2^{(c)}(x,y)$ is defined by

$$S_2^{(c)}(x,y) = \sum_{r=0}^{\infty} \frac{(-)^r}{r!} \left(2^{-\frac{1}{m}} \frac{y}{m}\right)^{2r} \left(e^{\mp \pi i} x\right)^{-2r\left(1-\frac{1}{m}\right)} C_r(\xi). \tag{3.14}$$

Collecting together the results in (3.5), (3.12), and (3.14), we then have the expansions for $|x| \to \infty$, y finite:

$$\begin{aligned} I_m(x,y) &\sim \frac{\pi^{1/2}}{m} x^{-\frac{1}{m}} S_1^{(c)}(x,y) \quad \text{in } |\arg x| < \frac{3}{4}\pi \\ &\sim \frac{\pi^{1/2}}{m} x^{-\frac{1}{m}} S_1^{(c)}(x,y) + \frac{\pi^{1/2}}{m} \left(\frac{1}{2} x e^{\mp \pi i}\right)^{\frac{1}{m}-1} e^{x^2/4} S_2^{(c)}(x,y) \\ &\qquad \text{in } |\arg(-x)| < \frac{1}{2}\pi, \end{aligned} \tag{3.15}$$

where the upper or lower sign is chosen according as $\frac{1}{2}\pi < \arg x < \pi$ or $-\pi < \arg x < -\frac{1}{2}\pi$, respectively.[1]

In a similar manner, the expansion of $J_m(x,y)$ for $|x| \to \infty$ is given by

$$\begin{aligned} J_m(x,y) &\sim \frac{\pi^{1/2}}{m} x^{-\frac{1}{m}} S_1^{(s)}(x,y) \quad \text{in } |\arg x| < \frac{3}{4}\pi \\ &\sim \frac{\pi^{1/2}}{m} x^{-\frac{1}{m}} S_1^{(s)}(x,y) + \frac{\pi^{1/2}}{m} \left(\frac{1}{2} x e^{\mp \pi i}\right)^{\frac{1}{m}-1} e^{x^2/4} S_2^{(s)}(x,y) \\ &\qquad \text{in } |\arg(-x)| < \frac{1}{2}\pi, \end{aligned} \tag{3.16}$$

where the formal asymptotic sum $S_2^{(s)}(x,y)$ is defined by

$$S_2^{(s)}(x,y) = \sum_{r=0}^{\infty} \frac{(-)^r}{r!} \left(2^{-\frac{1}{m}} \frac{y}{m}\right)^{2r} \left(e^{\mp \pi i} x\right)^{-2r\left(1-\frac{1}{m}\right)} S_r(\xi) \tag{3.17}$$

with

$$\begin{aligned} S_r(\xi) &= (-)^r m^{2r} \xi^{-2r/m} s_r(\xi), \\ s_r(\xi) &= \xi^{2r+1-\frac{1}{m}} \frac{d^{2r}}{d\xi^{2r}} \left(\xi^{\frac{1}{m}-1} \sin \xi^{\frac{1}{m}}\right) \quad (r = 0, 1, 2, \ldots), \end{aligned}$$

so that

$$S_r(\xi) = P_r(\xi) \sin \xi^{\frac{1}{m}} + Q_r(\xi) \cos \xi^{\frac{1}{m}}. \tag{3.18}$$

The expansion of $P_m(x,y)$ can then be constructed from (1.5), (1.6), (3.15), and (3.16). In the special case $y = 0$, we see from (1.4) that $J_m(x,0) \equiv 0$ and

$$I_m(x,0) = \frac{2^{-1/m}}{m} \Gamma(1/m) e^{x^2/8} D_{-1/m}(x/\sqrt{2}).$$

The expansion of $I_m(x,0)$ in (3.15) can be shown to agree with the large-x asymptotics of the parabolic cylinder function, when the sum $S_2^{(c)}(x,0)$ is evaluated by a limiting process as $y \to 0$.

[1] The statement of this result in the case $m = 2$ in [12, eq. (3.16)] is incorrect; the lower sign should be taken when $\pi < \arg x < \frac{3}{2}\pi$ and not as stated.

In terms of the original variables X and Y, we see that negative real X corresponds to the anti-Stokes line $\arg x = \frac{3}{4}\pi$. Thus, for $X \to -\infty$, $P'_m(X,Y)$ will consist of an algebraic and an exponentially oscillatory expansion, while for $X \to +\infty$, $P'_m(X,Y)$ will either consist of a single algebraic expansion or a mixed algebraic and exponentially oscillatory expansions according as m is even or odd. From (1.5), (1.6), (3.15), and (3.16) we then find (upon taking the upper signs) the leading asymptotic behavior

(3.19)

$$P'_m(X,Y) = \frac{2}{m}X^{-\frac{1}{m}}e^{\frac{\pi i}{2m}}\Gamma\left(\frac{1}{m}\right)\left\{1+O\left(X^{-\frac{2}{m}}\right)\right\}, \qquad X \to +\infty$$

$$= \frac{2}{m}|X|^{-\frac{1}{m}}e^{-\frac{\pi i}{2m}}\Gamma\left(\frac{1}{m}\right)\left\{1+O\left(X^{-\frac{2}{m}}\right)\right\} + \frac{2\pi^{1/2}}{m}\left(\frac{|X|}{2}\right)^{\frac{1}{m}-1}$$

$$\cdot \exp\left[i\left(\frac{\pi}{4}-\frac{X^2}{4}\right)\right]\cos\left[Y\left(\frac{|X|}{2}\right)^{\frac{1}{m}}\right]\left\{1+O\left(X^{-2(m-1)/m}\right)\right\}, \qquad X \to -\infty$$

when m is even, and

$$P'_m(X,Y) = \frac{2}{m}X^{-\frac{1}{m}}\left\{\Gamma\left(\frac{1}{m}\right)\cos\frac{\pi}{2m} - X^{-\frac{1}{m}}Y\Gamma\left(\frac{2}{m}\right)\sin\frac{\pi}{m} + O\left(X^{-\frac{2}{m}}\right)\right\}$$

$$+\frac{\pi^{1/2}}{m}\left(\frac{X}{2}\right)^{\frac{1}{m}-1}\exp\left[i\left\{\frac{\pi}{4}-\frac{X^2}{4}-Y\left(\frac{X}{2}\right)^{\frac{1}{m}}\right\}\right] \tag{3.20}$$

$$\cdot\left\{1+O\left(X^{-2(m-1)/m}\right)\right\}, \qquad X \to +\infty$$

when m is odd, the behavior for $X \to -\infty$ being given by $P'_m(-X,Y)$; cf. (1.7a).

The above leading behavior when m is even can be seen to correspond to a single ($X \to +\infty$) and three ($X \to -\infty$) stationary point contributions discussed in §2. In the case $m = 2$, the asymptotic behavior in (3.19) for $X \to \pm\infty$ agrees with that of the Pearcey integral [12].

4. Asymptotics of $P_m(x,y)$ for $|y| \to \infty$. From the symmetry property in (1.7) it is sufficient to consider only the sector $|\arg y| \le \frac{1}{2}\pi$. We first discuss the case of even $m = 2p$. Using the connection formula (3.10) in the form

$$D_v(z) = (2\pi)^{-1/2}\Gamma(1+v)\left\{e^{\pi v i/2}D_{-v-1}(iz) + e^{-\pi v i/2}D_{-v-1}(-iz)\right\} \tag{4.1}$$

together with the new variable σ given by $s = m\sigma + 1$, we find from (1.5) and (2.5), when m is even, that

$$P_m(x,y) = e^{(x^2/8)+(\pi i/4m)}\left\{e^{-\frac{\pi i}{2m}}K_p\left(x, ye^{-\frac{\pi i}{2m}}\right) + e^{\frac{\pi i}{2m}}K_p\left(-x, ye^{\frac{\pi i}{2m}}\right)\right\}, \tag{4.2}$$

where

$$K_p(x,y) = \frac{(2\pi)^{1/2}}{2\pi i}\int_C 2^{\sigma/2}\Gamma(m\sigma+1)\frac{\sin\pi p\sigma}{\sin\pi\sigma}D_{-\sigma-1}\left(\frac{ix}{\sqrt{2}}\right)y^{-m\sigma-1}d\sigma. \tag{4.3}$$

Here C is a loop in the σ-plane with endpoints at $-\infty$, which encircles $\sigma = -1/m$. As the parabolic cylinder function is an entire function of σ and the integrand has no

poles in $\mathrm{Re}(\sigma) \geq 0$, the restriction $\mathrm{Re}(\sigma) < 0$ imposed in the derivation of (2.5) can be removed by analytic continuation.

We now deform C such that $|\sigma|$ is large everywhere on C. On the expanded loop we may employ the asymptotic expansion of the parabolic cylinder function for large order and finite argument (cf. [12, eq. (A.10)])

$$D_{-\sigma-1}(z) \sim \frac{\pi^{1/2} \exp\left(-z\sqrt{\sigma+\frac{1}{2}}\right)}{2^{\sigma/2+1/2}\Gamma\left(\frac{1}{2}\sigma+1\right)} \sum_{r=0}^{\infty} (-)^r A_r \sigma^{-r/2}, \tag{4.4}$$

valid for $|\sigma| \to \infty$ in $|\arg\sigma| < \pi$, where

$$A_0 = 1, \quad A_1 = \frac{z^3}{24}, \quad A_2 = \frac{z^2}{48}\left(\frac{z^2}{24} - 3\right), \ldots.$$

The integral $K_p(x,y)$ then (formally) becomes

$$K_p(x,y) \sim \frac{2\pi m}{y} \sum_{r=0}^{\infty} (-)^r A_r \frac{1}{2\pi i} \int_C \frac{\Gamma(m\sigma)}{\Gamma(\frac{1}{2}\sigma)} \left(\frac{\sin p\pi\sigma}{\sin \pi\sigma}\right) e^{-ix(\sigma/2+1/4)^{1/2}} y^{-m\sigma} \sigma^{-\frac{1}{2}r} d\sigma.$$

We now express the quotient of gamma functions in terms of a single gamma function [2, p. 260]

$$\frac{\Gamma(m\sigma)}{\Gamma\left(\frac{1}{2}\sigma\right)} = \frac{(\pi m)^{-1/2}}{2} \left(\frac{\mu^\mu}{2^{1/2} m^m}\right)^{-\sigma} \Gamma\left(\mu\sigma + \frac{1}{2}\right) \left\{1 + O(\sigma^{-1})\right\}, \tag{4.5}$$
$$\mu = m - \frac{1}{2},$$

valid for $|\sigma| \to \infty$ in $|\arg\sigma| < \pi$, and employ the result

$$\frac{\sin p\pi\sigma}{\sin \pi\sigma} = \sum_{r=1}^{p} e^{\pi i\sigma(p+1-2r)}$$

to obtain

$$K_p(x,y) = \frac{(\pi m)^{\frac{1}{2}}}{y} \sum_{r=1}^{p} \frac{1}{2\pi i} \int_C \Gamma\left(\mu\sigma + \frac{1}{2}\right) e^{-ix\left(\frac{1}{2}\sigma+\frac{1}{4}\right)^{1/2}} (\omega_r y^*)^{-\mu\sigma} \tag{4.6}$$
$$\cdot \left[1 - A_1\sigma^{-\frac{1}{2}} + O\left(\sigma^{-1}\right)\right] d\sigma$$

where

$$\omega_r = \exp[\pi i(2r - p - 1)/\mu], \qquad y^* = 2\mu(y/2m)^{m/\mu}. \tag{4.7}$$

The integrals appearing in (4.6) can be evaluated asymptotically for large $|y|$ by means of the lemma given in [12], which we write in the following, slightly modified form. For arbitrary complex α, b, c and real a (> 0) and v, the integral

$$\frac{1}{2\pi i} \int_C \Gamma(at+b) e^{\alpha(at+c)^{\frac{1}{2}}} z^{-at} t^v dt$$
$$\sim \left(z^{v+b}/a^{v+1}\right) \exp\left[-z + \alpha z^{\frac{1}{2}} - \alpha^2/8\right] \sum_{r=0}^{\infty} B_r z^{-r/2},$$
$$B_0 = 1, \qquad B_1 = \frac{1}{8}\alpha\left[1 - 4(v+b-c) - \alpha^2/12\right], \ldots,$$

for $|z| \to \infty$ in $|\arg z| < \pi$. Application of this lemma with successively $v = 0, -\frac{1}{2}$ and $\alpha = -ix/(2\mu)^{\frac{1}{2}}$, $c = \mu/2$ to $K_p(x,y)$ then yields the expansion for $|y| \to \infty$ (in the sense of Poincaré)

$$
\begin{aligned}
e^{(x^2/8)+(\pi i/4m)}K_p(x,y) &\sim \sum_{r=1}^{p} E_r^{(p)}(x,y), \quad |\arg y| \le \frac{3\pi}{4m} \\
&\sim E_p^{(p)}(x,y), \quad \frac{3\pi}{4m} < \arg y < \frac{\pi}{2m}(m+1) \\
&\sim E_1^{(p)}(x,y), \quad \frac{-\pi}{2m}(m+1) < \arg y < -\frac{3\pi}{4m},
\end{aligned} \tag{4.8}
$$

where

(4.9)

$$
\begin{aligned}
&E_r^{(p)}(x,y) \\
&\quad = (\pi\omega_r/\mu)^{\frac{1}{2}}\left[y^{1-m}/(2m)^{\frac{1}{2}}\right]^{1/2\mu}\exp\left[\pi i/4m - \omega_r y^* - ix(\omega_r y^*/2\mu)^{\frac{1}{2}} + mx^2/8\mu\right] \\
&\qquad \cdot\left\{1 - \frac{ix(m-1)}{4(2\mu\omega_r y^*)^{\frac{1}{2}}}(1 - mx^2/12\mu) + O(y^{*-1})\right\}, \qquad (r = 1,2,\ldots,p).
\end{aligned}
$$

For odd $m = 2p+1$, we find, from (1.6), (2.5), and (4.1), that

(4.10)

$$
\begin{aligned}
P_m(x,y) &= e^{(x^2/8)+(\pi i/4m)}\frac{i}{2\pi i}\int_C 2^{\sigma/2}\Gamma(m\sigma+1)\Gamma(-\sigma) \\
&\qquad \cdot\left\{e^{(1/2)\pi i m\sigma}D_\sigma\left(\frac{x}{\sqrt{2}}\right) - e^{(-1/2)\pi i m\sigma}D_\sigma\left(\frac{-x}{\sqrt{2}}\right)\right\}y^{-m\sigma-1}d\sigma \\
&= e^{(x^2/8)+(\pi i/4m)}\frac{(2\pi)^{1/2}}{2\pi i}\int_C 2^{\sigma/2}\Gamma(m\sigma+1) \\
&\qquad \cdot\left\{\frac{\sin\pi p\sigma}{\sin\pi\sigma}D_{-\sigma-1}\left(\frac{-ix}{\sqrt{2}}\right) + \frac{\sin\pi(p+1)\sigma}{\sin\pi\sigma}D_{-\sigma-1}\left(\frac{ix}{\sqrt{2}}\right)\right\}y^{-m\sigma-1}d\sigma.
\end{aligned}
$$

It is clear that the only poles of the integrand in (4.10) are those resulting from $\Gamma(m\sigma+1)$, which are enclosed by C. The restriction $\mathrm{Re}(\sigma) < 0$ can again be removed by analytic continuation and the loop C can consequently be expanded as in the case m even. We therefore find

$$
P_m(x,y) = e^{(x^2/8)+(\pi i/4m)}\{K_p(-x,y) + K_{p+1}(x,y)\} \qquad (m \text{ odd}). \tag{4.11}
$$

Then, from (4.2), (4.11), and (4.8), the expansion of $P_m(x,y)$ for $|y| \to \infty$, x finite, when $m = 2p$ is even, is then (in the Poincaré sense) given by

$$
\begin{aligned}
P_m(x,y) &\sim e^{-\pi i/2m}\sum_{r=1}^{p} E_r^{(p)}\left(x, ye^{-\pi i/2m}\right) \\
&\quad + e^{\pi i/2m}\sum_{r=1}^{p} E_r^{(p)}\left(-x, ye^{\pi i/2m}\right) \quad |\arg y| \le \frac{\pi}{4m} \\
&\sim e^{\pi i/2m}E_p^{(p)}\left(-x, ye^{\pi i/2m}\right), \qquad \frac{\pi}{4m} < \arg y < \frac{1}{2}\pi \\
&\sim e^{-\pi i/2m}E_1^{(p)}\left(x, ye^{-\pi i/2m}\right), \qquad -\frac{1}{2}\pi < \arg y < -\frac{\pi}{4m},
\end{aligned} \tag{4.12}
$$

and when $m = 2p+1$ is odd,

$$
\begin{aligned}
P_m(x,y) &\sim \sum_{r=1}^{p} E_r^{(p)}(-x,y) + \sum_{r=1}^{p+1} E_r^{(p+1)}(x,y) \qquad |\arg y| \le \frac{\pi}{4m} \\
&\sim E_{p+1}^{(p+1)}(x,y), \qquad \frac{\pi}{4m} < \arg y < \frac{1}{2}\pi \\
&\sim E_1^{(p+1)}(x,y), \qquad -\frac{1}{2}\pi < \arg y < -\frac{\pi}{4m}.
\end{aligned} \tag{4.13}
$$

In the sectors $|\arg(\pm y)| < \pi/4m$, $P_m(x,y)$ is exponentially small as $|y| \to \infty$, while outside of these sectors $P_m(x,y)$ is exponentially large.

In terms of the original variables X and Y, we see that large positive Y corresponds to the upper boundary of the exponentially small sector $|\arg y| < \pi/4m$. Here, $E_p^{(p)}(-x, ye^{\pi i/2m})$ (m even) or $E_{p+1}^{(p+1)}(x,y)$ (m odd) possesses oscillatory behavior, with the other $m-1$ expansions in (4.12) and (4.13) exponentially small and of different degrees of subdominance. Thus, up to exponentially small terms, the behavior of $P_m'(X,Y)$ for $Y \to +\infty$, X finite, is given by

(4.14)

$$
\begin{aligned}
P_m'(X,Y) &= (\pi/\mu)^{\frac{1}{2}}(2m)^{-1/4\mu Y(1-m)/2\mu}\exp i\left[\pi/4 - Y^* + \frac{(-)^m XY^{*\frac{1}{2}}}{(2\mu)^{\frac{1}{2}}} - mX^2/8\mu\right] \\
&\quad \cdot\left\{1 + \frac{(-)^m(m-1)XY^{*-\frac{1}{2}}}{4(2\mu)^{\frac{1}{2}}}\left(1 + \frac{imX^2}{12\mu}\right) + O\left(Y^{*-1}\right)\right\},
\end{aligned}
$$

where

$$
Y^* = 2\mu(Y/2m)^{m/\mu}, \qquad \mu = m - \tfrac{1}{2}.
$$

This behavior can be seen to correspond to the single real stationary point discussed in §2, with the $m-1$ subdominant contributions (not stated in (4.14)) arising from half of the $2(m-1)$ complex stationary points. When $m = 2$, the expansion (4.14) reduces to that of the Pearcey integral.

The approximations in (4.12) and (4.13) describe the asymptotic behavior of $P_m(x,y)$ for $|y| \to \infty$ in the sense of Poincaré. As noted in §5 of [12], the method adopted in this section is not sufficiently precise to determine the domains of validity

of the subdominant exponentially small terms in the above expansions and their associated Stokes lines. This deficiency results from the manner of approximation of the quotient of gamma functions in (4.5), which does not correctly take into account the appearance of exponentially small terms in the large $|z|$ behavior of $\Gamma(z)$ across its Stokes lines $\arg z = \pm\frac{1}{2}\pi$ (see [13]).

To deal with this situation, and the case $\arg y = \pm\frac{1}{2}\pi$, we revert to consideration of the saddle points of the integrand in (1.3). We omit the detailed description of the nature of the paths of steepest descent through the saddle points, which resembles that in the case $m = 2$ [12]. It is found that as $|y| \to \infty$ there is a sequence of Stokes lines given by $\arg(\pm y) = \theta_1, \theta_2, \ldots, \theta_{m-1}$, where $\theta_r = \pi(2r+1)/4m$, across which the coefficients of the exponentially small subdominant expansions in (4.12) and (4.13) change to become zero (the Stokes phenomenon). When m is even, the expansions $E^{(p)}_{p-r}(x, ye^{\pi i/2m})$ and $E^{(p)}_{p-r+1}(x, ye^{-\pi i/2m})$ "switch off" as one crosses (in the sense of increasing $|\arg y|$) $\theta_{2r}(r = 1, 2, \ldots, p-1)$ and $\theta_{2r-1}(r = 1, 2, \ldots, p)$, respectively. When m is odd, a similar pattern of behavior emerges with the expansions $E^{(p+1)}_{p-r+1}(x, y)$ and $E^{(p)}_{p-r+1}(x, y)$ "switching off" across the rays θ_{2r} and $\theta_{2r-1}(r = 1, 2, \ldots, p)$, respectively. A conjugate behavior applies by virtue of (1.7b) in the sector $-\frac{1}{2}\pi \le \arg y < 0$. In this manner the above expansions for $P_m(x, y)$ can be shown to extend to $|\arg y| \le \frac{1}{2}\pi$, with (4.12a) and (4.13a) holding in the wider sector $|\arg y| < 3\pi/4m$ bounded by the first pair of Stokes lines $\arg y = \pm\theta_1$.

Rather than state the complete expansion in the general case, we present, as an illustrative example, only the specific case $m = 3$. The Stokes lines in $\mathrm{Re}(y) > 0$ in this case are $\arg y = \pm\frac{1}{4}\pi$ and $\arg y = \pm 5\pi/12$, and we find that

$$
\begin{aligned}
P_3(x, y) &\sim E^{(1)}_1(-x, y) + \sum_{r=1}^{2} E^{(2)}_r(x, y), \qquad |\arg y| < \frac{1}{4}\pi \\
&\sim \sum_{r=1}^{2} E^{(2)}_r(x, y), \qquad \frac{1}{4}\pi < |\arg y| < \frac{5\pi}{12}, \\
&\sim E^{(2)}_2(x, y), \qquad \frac{5\pi}{12} < \arg y \le \frac{1}{2}\pi, \\
&\sim E^{(2)}_1(x, y), \qquad -\frac{1}{2}\pi \le \arg y < \frac{5}{12}\pi,
\end{aligned}
\tag{4.15}
$$

where, from (4.9),

$$
E^{(p)}_r(x, y) = (2\pi\omega_r/5)^{\frac{1}{2}} \left(6^{\frac{1}{4}} y\right)^{-2/5} \exp\left[-5\omega_r(y/6)^{6/5} - ix\omega_r^{\frac{1}{2}}(y/6)^{3/5} + 3x^2/20 + \pi i/12\right]
$$

$$
\cdot \left\{1 - \frac{ix\omega_r^{-\frac{1}{2}}}{10}(y/6)^{-3/5}(1 - x^2/10) + O\left(y^{-6/5}\right)\right\},
$$

$$
\omega_r = \exp[2\pi i(2r - p - 1)/5].
$$

The different sectorial behavior for $P_3(x, y)$ is summarized in Fig. 2. We note that the above asymptotic approximations satisfy the conjugacy property (1.7b) and, for real x, to correctly predict real values of $\exp(-\pi i/12)P_3(x, y)$ when $\arg y = \pm\frac{1}{2}\pi$.

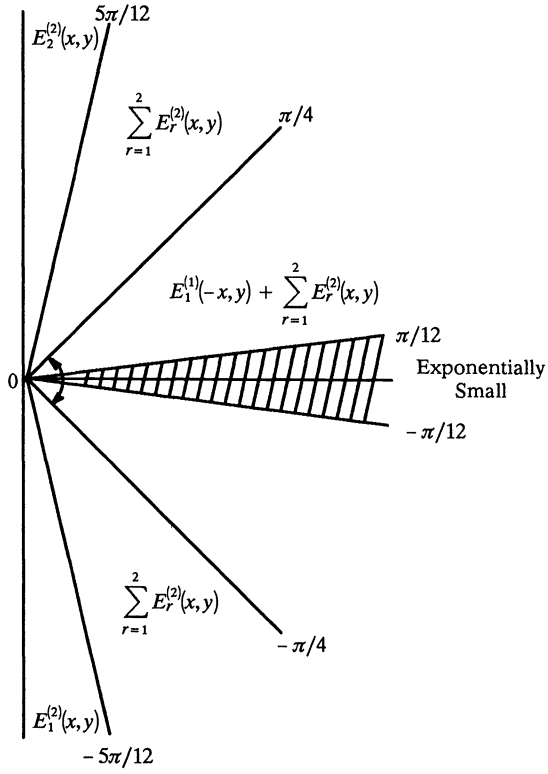


FIG. 2. *The sectorial behavior of* $P_m(x,y)$ *when* $m = 3$ *in the* y*-plane as* $|y| \to \infty$ *in* $|\arg y| \leq \frac{1}{2}\pi$.

REFERENCES

[1] M. ABRAMOWITZ AND I. STEGUN, EDS., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] B. L. J. BRAAKSMA, *Asymptotic expansion and analytic continuation for a class of Barnes integrals*, Compositio Math., 15 (1964), pp. 239–341.

[3] J. N. L. CONNOR, *Semiclassical theory of molecular collisions: three nearly coincident classical trajectories*, Molecular Phys., 26 (1973), pp. 1217–1231.

[4] J. N. L. CONNOR AND P. R. CURTIS, *A method for the numerical evaluation of the oscillatory integrals associated with cuspoid catastrophes: application to Pearcey's integral and its derivatives*, J. Phys. A, 15 (1982), pp. 1179–1190.

[5] ———, *Differential equations for the cuspoid canonical integrals*, J. Math. Phys., 25 (1984), pp. 2895–2902.

[6] J. N. L. CONNOR, P. R. CURTIS, AND D. FARELLY, *A differential equation method for the numerical evaluation of the Airy, Pearcey and swallowtail canonical integrals and their derivatives*, Molecular Phys., 48 (1983), pp. 1305–1330.

[7] A. ERDELYI, ED., *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.

[8] A. J. E. M. JANSSEN, *On the asymptotics of some Pearcey-type integrals*, J. Phys. A, 25 (1992), pp. 823–831.

[9] D. KAMINSKI, *Asymptotic expansion of the Pearcey integral near the caustic*, SIAM J. Math. Anal., 20 (1989), pp. 987–1005.

[10] ———, *Asymptotics of the swallowtail integral near the cusp of the caustic*, SIAM J. Math. Anal., 22 (1992), pp. 262–285.

[11] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[12] R. B. PARIS, *The asymptotic behavior of Pearcey's integral for complex variables*, Proc. Roy. Soc. London, Ser. A 432 (1991), pp. 391–426.

[13] R. B. PARIS AND A. D. WOOD, *Exponentially-improved asymptotics for the gamma function*, J. Comp. Appl. Math., 41 (1992), pp. 135–143.

[14] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.

[15] J. J. STAMNES AND B. SPJELKAVIK, *Evaluation of the field near a cusp of a caustic*, Optica Acta, 30 (1983), pp. 1331–1358.

[16] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, 1965.

[17] R. WONG, *Asymptotic Approximation of Integrals*, Academic Press, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 646–693, March 1994

027

THE PEARSON EQUATION AND THE BETA INTEGRALS*

MIZAN RAHMAN† AND SERGEI K. SUSLOV‡

Abstract. An alternate proof of the classical beta integral is given by using the Pearson equation whose solution converts the hypergeometric equation into a self-adjoint equation. A generalization of this idea to difference equations on linear lattices enables a proof of Barnes's second lemma by using his first lemma and gives extensions of some of Ramanujan's formulas. Similar analysis on q-quadratic lattices gives an extension of Askey's integral on the real line and the corresponding basic bilateral sum of Gosper. Another q-quadratic lattice gives an extension of the Askey–Wilson integral. A quadratic lattice is used to evaluate a principal value integral on the real line, whose q-analogue is shown to be equivalent to the Askey integral extension. A list of various beta integrals is also presented.

Key words. beta integrals of Euler, Barnes and Ramanujan, q-beta integrals of Askey and Wilson, Pearson equation, very well poised sums and integrals

AMS subject classifications. primary 33A15; secondary 33A10

1. Introduction.

One of the most basic formulas in all of classical analysis is Euler's beta integral

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{1.1}$$

Re $(\alpha,\beta)>0$; the integrand $\rho(x)=x^{\alpha-1}(1-x)^{\beta-1}$ is the weight function associated with the differential equation for the Jacobi polynomials (see [54]):

$$x(1-x)y''+[\alpha-(\alpha+\beta)x]y'-n(n+\alpha+\beta-1)y=0, \tag{1.2}$$

which becomes self-adjoint when multiplied by $\rho(x)$. The function $\rho(x)$ is the solution of the Pearson equation (see, e.g., [17]):

$$[\rho(x)\sigma(x)]'=\rho(x)\tau(x), \tag{1.3}$$

with $\sigma(x)=x(1-x)$, $\tau(x)=\alpha-(\alpha+\beta)x$ for (1.2).

There are quite a few proofs of (1.1), the earliest one by Euler dating back to 1772 (see [58]) and the latest ones by Knuth [35] in 1973 and Askey [3] in 1981. We shall give another one in §2, not because there is any reason to believe it to be any more elegant than the others (in fact, our proof is very similar to the one given by Askey), but because we want to make the point that a similar technique can be applied to other extensions of (1.1). The observation we wish to make here, and which will be a central theme throughout the paper, is that (1.3), being a first-order linear equation in

*Received by the editors April 2, 1992; accepted for publication February 8, 1993. This work was supported in part by Natural Sciences and Engineering Research Council of Canada grant A6197, and was completed while the second author was visiting Carleton University from January to April, 1991 and from January to April, 1992.

†Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

‡Kurchatov Institute of Atomic Energy, Moscow 123182, Russia.

$\rho(x)$, not only yields the solution in a very elementary manner but also enables one to compute the integral over $\rho(x)$, like the one in (1.1), provided there is at least one free parameter in $\sigma(x)$ and/or $\tau(x)$. We shall find that with appropriate generalizations of (1.2) and (1.3), the same procedures lead us to alternate proofs of some old extensions of (1.1) as well as to discovering new ones. But first, let us give a short list of some known extensions of the beta integral.

First is Cauchy's extension [19]:

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\frac{dt}{(1+at)^x(1-bt)^y}=\frac{\Gamma(x+y-1)}{\Gamma(x)\Gamma(y)}a^{y-1}b^{x-1}(a+b)^{1-x-y}, \tag{1.4}$$

Re $(a,b)>0$, Re $(x+y)>1$, which does not resemble (1.1) but looks closer to one of its variants:

$$\int_0^\infty\frac{t^{\alpha-1}dt}{(1+t)^{\alpha+\beta}}=B(\alpha,\beta). \tag{1.5}$$

Then there is Barnes's first lemma [15],

$$\begin{aligned}\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}&\Gamma(a+s)\Gamma(b+s)\Gamma(c-s)\Gamma(d-s)ds\\&=\frac{\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)}{\Gamma(a+b+c+d)},\end{aligned} \tag{1.6}$$

and his second lemma [16],

$$\begin{aligned}\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}&\frac{\Gamma(a+s)\Gamma(b+s)\Gamma(c+s)\Gamma(1-d-s)\Gamma(-s)}{\Gamma(e+s)}ds\\&=\frac{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(1+a-d)\Gamma(1+b-d)\Gamma(1+c-d)}{\Gamma(e-a)\Gamma(e-b)\Gamma(e-c)},\end{aligned} \tag{1.7}$$

where $e=a+b+c-d+1$. In (1.6) the assumption is that none of the poles of $\Gamma(a+s)\Gamma(b+s)$ coincide with any of the poles of $\Gamma(c-s)\Gamma(d-s)$. In (1.7) the line of integration is the imaginary axis or a line parallel to it with indentations, if necessary, so that the decreasing sequences of poles lie to the left, and the increasing sequences of poles lie to the right of the contour. Askey and Roy [8] showed, by using Stirling's formula, how to deduce (1.1) from (1.6) in a limiting process, and hence called (1.6) Barnes's beta integral. It has been pointed out in [10] and [7] that the integrand in (1.6) is the weight function for the continuous Hahn polynomials. Barnes's second lemma (1.7), an extension of (1.6), may likewise be considered a further extension of (1.1), and corresponds to the continuous biorthogonality of the ${}_4F_3$ rational functions introduced in [46].

A different kind of extension was given by Ramanujan [50]:

$$\begin{aligned}\int_{-\infty}^{\infty}&\frac{dx}{\Gamma(a+x)\Gamma(b+x)\Gamma(c-x)\Gamma(d-x)}\\&=\frac{\Gamma(a+b+c+d-3)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)},\end{aligned} \tag{1.8}$$

where Re $(a+b+c+d) > 3$. The integral on the left is along the real line in contrast to (1.6) where the integral is along the imaginary axis. This is an important difference whose significance will become clearer in later sections. Also, the integrand in (1.8) is the reciprocal of that in (1.6) as is the expression on the right side except for a shift in the Γ-functions. The system of Hahn polynomials that corresponds to the measure in (1.8), was discussed in [4]. For other integrals of Ramanujan similar to (1.8), see [50] and [22].

Among the formulas that we shall prove in this article is a generalization of (1.8):

$$\int_{-\infty}^{\infty} \frac{\Gamma(1-f-x)}{\Gamma(a+x)\Gamma(b+x)\Gamma(c-x)\Gamma(d-x)\Gamma(e-x)}dx \tag{1.9}$$
$$= \frac{\Gamma(2-c-f)(\Gamma(2-d-f)\Gamma(2-e-f)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(a+e-1)\Gamma(b+c-1)\Gamma(b+d-1)\Gamma(b+e-1)},$$

provided

$$\begin{aligned} &\text{(i)} && a+b+c+d+e+f=5,\\ &\text{(ii)} && \operatorname{Im} f \neq 0,\\ &\text{(iii)} && \operatorname{Re}(2-f) > \operatorname{Re}(c,d,e),\\ &\text{(iv)} && \text{a "regularizing" condition, to be stated in §5, is}\\ & && \text{also satisfied by the parameters.} \end{aligned} \tag{1.10}$$

There is also a generalization of Barnes's formula (1.6) due to de Branges [18] and Wilson [59]:

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \frac{\Gamma(a+s)\Gamma(b+s)\Gamma(c+s)\Gamma(d+s)}{\Gamma(2s)} \cdot \frac{\Gamma(a-s)\Gamma(b-s)\Gamma(c-s)\Gamma(d-s)}{\Gamma(-2s)}ds \tag{1.11}$$
$$= 2\frac{\Gamma(a+b)\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)\Gamma(c+d)}{\Gamma(a+b+c+d)},$$

provided that any pairwise sum of the four parameters a, b, c, d (including $2a, 2b, 2c, 2d$) is not a nonpositive integer. It is understood that the contour of integration is the imaginary axis suitably indented to separate the increasing sequence of poles from the decreasing one.

Askey mused in [4] whether the integral in (1.11) could be modified in the same way as Barnes's integral (1.6) was modified into the Ramanujan integral (1.8), and he wished for a Ramanujan to look at this problem. Askey's instincts were correct, of course, for we shall prove in this paper that the desired analogue is

$$\int_{-\infty}^{\infty} \frac{\Gamma(1+2s)\Gamma(1-2s)ds}{\Gamma(a+s)\Gamma(a-s)\Gamma(b+s)\Gamma(b-s)\Gamma(c+s)\Gamma(c-s)\Gamma(d+s)\Gamma(d-s)} \tag{1.12}$$
$$= \frac{\Gamma(a+b+c+d-3)}{\Gamma(a+b-1)\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)\Gamma(c+d-1)},$$

but Askey was wrong about the need for a Ramanujan. Far lesser individuals like the authors could handle the problem thanks to the ideas buried in [4] and in some of Askey's other works on beta integrals.

There is a third family of beta integral extensions that are not integrals at all, rather sums, finite, infinite, or doubly infinite (commonly called bilateral series). An important member of this family is Dougall's sum [21]:

$$\sum_{n=-\infty}^{\infty} \frac{1}{\Gamma(a+n)\Gamma(b+n)\Gamma(c-n)\Gamma(d-n)} = \frac{\Gamma(a+b+c+d-3)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)}, \tag{1.13}$$

with Re $(a+b+c+d) > 3$. The connection between the formulas (1.8) and (1.13), which are strikingly similar, was made clear by Osler's [45] extension of (1.13):

$$\sum_{n=-\infty}^{\infty} \frac{\alpha}{\Gamma(a+\alpha n)\Gamma(b+\alpha n)\Gamma(c-\alpha n)\Gamma(d-\alpha n)} = \frac{\Gamma(a+b+c+d-3)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)}, \tag{1.14}$$

$0 < \alpha \leq 1$, which approaches Ramanujan's integral formula (1.8) in the limit $\alpha \to 0$, and to (1.13) when $\alpha = 1$. Another Ramanujan-type extension of (1.13) is

$$\sum_{n=-\infty}^{\infty} \frac{\alpha+n}{\Gamma(a+\alpha+n)\Gamma(a-\alpha-n)\Gamma(b+\alpha+n)\Gamma(b-\alpha-n)} \cdot \frac{1}{\Gamma(c+\alpha+n)\Gamma(c-\alpha-n)\Gamma(d+\alpha+n)\Gamma(d-\alpha-n)} = \frac{\sin 2\pi\alpha}{2\pi} \frac{\Gamma(a+b+c+d-3)}{\Gamma(a+b-1)\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)\Gamma(c+d-1)}, \tag{1.15}$$

where Re $(a+b+c+d) > 3$, which is just the ${}_5H_5$ summation formula due to Dougall [21]; see also Slater [52, III.29]. The sum in (1.15) may be considered as a discrete version of the integral in (1.12).

Among the other discrete extensions of the beta integral are the well-known sums due to Vandermonde, Gauss, Pfaff–Saalschütz, Dixon, Dougall, and Bailey, for which we refer the reader to Bailey [14] and Slater [52].

Important as all these formulas are, the sums and integrals of current interest involve a further generalization—the so-called q-extension. The additional parameter q enters all the formulas through the so-called q-shifted factorial

$$(a;q)_n = \begin{cases} 1, & n = 0 \\ (1-a)(1-aq)\dots(1-aq^{n-1}), & n = 1,2,\dots, \end{cases}$$

$$(a;q)_\infty = \prod_{n=0}^{\infty}(1-aq^n), \qquad |q| < 1, \tag{1.16}$$

$$(a_1, a_2, \dots, a_r; q)_n = \prod_{j=1}^{r}(a_j;q)_n;$$

the q-gamma function introduced by Thomae [55] and Jackson [31], [32]:

$$\Gamma_q(x) = \frac{(q;q)_\infty}{(q^x;q)_\infty}(1-q)^{1-x}, \qquad 0<q<1, \ \operatorname{Re} x>0; \tag{1.17}$$

the q-beta function

$$B_q(x,y) = \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)}; \tag{1.18}$$

Thomae [55], [56] and Jackson's [32] q-integrals

$$\int_0^a f(t)d_qt = a(1-q)\sum_{n=0}^{\infty} f(aq^n)q^n, \tag{1.19}$$

$$\int_0^\infty f(t)d_qt = (1-q)\sum_{n=-\infty}^{\infty} f(q^n)q^n; \tag{1.20}$$

and the q-binomial formula

$$\sum_{n=0}^{\infty}\frac{(a;q)_n}{(q;q)_n}z^n = \frac{(az;q)_\infty}{(z;q)_\infty}, \qquad |z|<1, \ |q|<1, \tag{1.21}$$

due to Cauchy [20] and Heine [30]. Using (1.18) and (1.19) one can show that the q-binomial formula leads to

$$B_q(x,y) = \int_0^1 t^{x-1}\frac{(tq;q)_\infty}{(tq^y;q)_\infty}d_qt, \tag{1.22}$$

$\operatorname{Re} x>0, \ y\neq 0, \ -1,-2,\ldots$. It is easy to show that

$$\lim_{q\to 1^-} B_q(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = B(x,y), \tag{1.23}$$

so that the q-beta integral in (1.22) can be treated as a q-analogue of the classical beta integral in (1.1). See Gasper and Rahman [25] for a detailed analysis of the q-analogues of some of the classical functions.

A q-analogue of the beta integral in the form (1.5), however, does not come directly from the q-binomial formula but from an extension of it, due to Ramanujan [29], which, in the present context, can be expressed in the form

$$\begin{aligned}\int_0^\infty t^{x-1}\frac{(-atq^{x+y};q)_\infty}{(-at;q)_\infty}d_qt \\ = a^{-x}q^{-\binom{x}{2}}\theta(aq^x)B_q(x,y)/\theta(a),\end{aligned} \tag{1.24}$$

where $\operatorname{Re}(x,y)>0$, $\operatorname{Re} a>0$, $|q|<1$, and $\theta(aq^x)$ is a periodic function of x of unit period

$$\theta(aq^x) = q^{-\binom{x+1}{2}}(aq^x)^x(-aq^x,-q^{1-x}/a;q)_\infty. \tag{1.25}$$

Closely related to (1.24) is another q-analogue of (1.5), again due to Ramanujan [29]:

$$\int_0^\infty t^{x-1}\frac{(-tq^{x+y};q)_\infty}{(-t;q)_\infty}dt = \frac{\Gamma(x)\Gamma(1-x)}{\Gamma_q(x)\Gamma_q(1-x)}B_q(x,y), \tag{1.26}$$

Re $(x,y)>0$. If x is a positive integer then a limit needs to be taken. As $q\to 1^-$ this approaches the formula (1.5), but it lacks the symmetry that is self-evident in (1.5), so Askey and Roy [8] and, independently Gasper [23], [24], found an extension of it:

$$\int_0^\infty t^{c-1}\frac{(-tq^{x+c},-q^{1+y-c}/t;q)_\infty}{(-t,-q/t;q)_\infty}dt = \frac{\Gamma(c)\Gamma(1-c)}{\Gamma_q(c)\Gamma_q(1-c)}B_q(x,y), \tag{1.27}$$

Re $(x,y)>0$, that restores the symmetry. There are a number of other q-beta integrals that are important but will be of no particular interest for our purposes, so we refer the reader to Andrews and Askey [2], and to Askey [3]–[6]. We shall, however, be interested in Watson's [57] q-analogue of Barnes's first lemma:

$$\begin{aligned}\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}&\Gamma_q(a+s)\Gamma_q(b+s)\Gamma_q(c-s)\Gamma_q(d-s)\phi(s)q^s ds\\ &= \frac{q^c\Gamma(c-d)\Gamma(1+d-c)}{\Gamma_q(d-c)\Gamma_q(1+c-d)}\cdot\frac{\Gamma_q(a+c)\Gamma_q(a+d)\Gamma_q(b+c)\Gamma_q(b+d)}{\Gamma_q(a+b+c+d)},\end{aligned} \tag{1.28}$$

where

$$\phi(s)=\pi^2[\sin\pi(c-s)\sin\pi(d-s)\Gamma_q(c-s)\Gamma_q(1-c+s)\Gamma_q(d-s)\Gamma_q(1-d+s)]^{-1}, \tag{1.29}$$

so that $\phi(s+1)=q^{2(s+1)-c-d}\phi(s)$, and in Agarwal's [1] q-analogue of Barnes's second lemma:

$$\begin{aligned}\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}&\frac{\Gamma_q(a+s)\Gamma_q(b+s)\Gamma_q(c+s)\Gamma_q(1-d-s)\Gamma_q(-s)}{\Gamma_q(e+s)}\phi_1(s)q^s ds\\ &= \frac{\Gamma(d)\Gamma(1-d)}{\Gamma_q(d)\Gamma_q(1-d)}\,\frac{\Gamma_q(a)\Gamma_q(b)\Gamma_q(c)\Gamma_q(1+a-d)\Gamma_q(1+b-d)\Gamma_q(1+c-d)}{\Gamma_q(e-a)\Gamma_q(e-b)\Gamma_q(e-c)},\end{aligned} \tag{1.30}$$

where $d+e=a+b+c+1$, and

$$\phi_1(s)=\pi^2[\sin\pi s\sin\pi(d+s)\Gamma_q(s+1)\Gamma_q(-s)\Gamma_q(d+s)\Gamma_q(1-d-s)]^{-1}, \tag{1.31}$$

so that $\phi_1(s+1)=q^{2s+1+d}\phi_1(s)$. For the conditions that the parameters must satisfy in order that (1.28) and (1.30) are valid, see the original references or [25]. For a recent proof of (1.28) that is very similar to the technique used in this paper see Kalnins and Miller [34].

A q-analogue of (1.11) is the Askey–Wilson integral [9]:

$$\begin{aligned}\frac{1}{2\pi i}\int_K&\frac{(z^2,z^{-2};q)_\infty}{(az,az^{-1},bz,bz^{-1},cz,cz^{-1},dz,dz^{-1};q)_\infty}\frac{dz}{z}\\ &= \frac{2(abcd;q)_\infty}{(q,ab,ac,ad,bc,bd,cd;q)_\infty},\end{aligned} \tag{1.32}$$

where K is a deformation of the positively oriented unit circle so that the zeros of $(az, bz, cz, dz; q)_\infty$ lie outside the contour and those of $(az^{-1}, bz^{-1}, cz^{-1}, dz^{-1}; q)_\infty$ lie inside. Such a contour always exits if the pairwise products of the parameters a, b, c, d (their squares included) are not of the form q^{-n}, $n = 0, 1, 2, \ldots$, $|q| < 1$. In the case $\max(|a|, |b|, |c|, |d|) < 1$, Askey and Wilson [9] reduced it to a real integral formula:

$$\int_{-1}^{1} \frac{h(x; 1, -1, q^{\frac{1}{2}}, -q^{\frac{1}{2}})}{h(x; a, b, c, d)} \frac{dx}{\sqrt{1-x^2}} = \frac{2\pi(abcd; q)_\infty}{(q, ab, ac, ad, bc, bd, cd; q)_\infty}, \tag{1.33}$$

where

$$\begin{aligned} h(x; a_1, a_2, \ldots, a_r) &= h(x; a_1) \ldots h(x; a_r), \\ h(x; a) &= \prod_{n=0}^{\infty} (1 - 2axq^n + a^2 q^{2n}) \\ &= (ae^{i\theta}, ae^{-i\theta}; q)_\infty \quad \text{if } x = \cos\theta. \end{aligned} \tag{1.34}$$

This is a remarkable formula that has sparked a great deal of research in q-extensions of special functions generally, and of orthogonal polynomials, in particular; for references see [25]. By specializing the parameters $a = q^{1/2}$, $b = q^{\alpha-(1/2)}$, $c = -q^{\beta-1/2}$, $d = -q^{1/2}$, it was shown in [47] that (1.33) approaches (1.1) in the limit $q \to 1^-$.

In [38] and [48] formula (1.33) was extended even further:

$$\begin{aligned} &\int_{-1}^{1} \frac{h(x; 1, -1, q^{\frac{1}{2}}, -q^{\frac{1}{2}}, g)}{h(x; a, b, c, d, f)} \frac{dx}{\sqrt{1-x^2}} \\ &\quad = \frac{2\pi(g/a, g/b, g/c, g/d, g/f; q)_\infty}{(q, ab, ac, ad, af, bc, bd, bf, cd, cf, df; q)_\infty}, \end{aligned} \tag{1.35}$$

where $g = abcdf$, and $\max(|a|, |b|, |c|, |d|, |f|, |q|) < 1$. See Askey [5] for an alternate proof of (1.35). We shall give another proof of (1.35) as an illustration of the methods of this paper.

Askey's name seems to be attached to most of the interesting modern extensions of the beta integral, so it is hardly surprising, but remarkable nonetheless, that he came along, once again, to offer the following q-extension [6] of the Cauchy-form of the beta integral:

$$\begin{aligned} &\int_{-\infty}^{\infty} \frac{h(i\sinh u; a, b, c, d)}{h(i\sinh u; q^{\frac{1}{2}}, -q^{\frac{1}{2}}, q, -q)} du \\ &\quad = (\log q^{-1}) \frac{(ab/q, ac/q, ad/q, bc/q, bd/q, cd/q, q; q)_\infty}{(abcd/q^3; q)_\infty}, \end{aligned} \tag{1.36}$$

provided $|abcdq^{-3}| < 1$ and $|q| < 1$. This is a Ramanujan-type integral in contrast to the Barnes-type integral (1.32). It is not difficult to apply the methods of this paper to give an alternate proof of (1.36), but for considerations of length we will instead give an extension of it in §9, in the same spirit as (1.35) is an extension of (1.33). We shall see, however, that the extension is not quite as straightforward as (1.35) would lead us to believe.

In a beautiful piece of work Gustafson [27] gave multidimensional extensions of many of the beta integrals mentioned above, including (1.11), (1.32), and (1.36). Interested readers should work through this paper as well as the references therein.

The plan of the paper is as follows. In §2 we give a proof of the classical beta integral to set the tone of the basic method that we are going to use throughout the paper. Section 3 is a recapitulation of the finite difference schemes described in [12], [40], [42]–[44], and [53], to point out the role that a Pearson equation type difference equation plays in computing sums and integrals. In §4 we give an alternate proof of Barnes's second lemma to give the readers a taste of what our methods can do for a formula that is less familiar than the beta integral (1.1). Section 5 is devoted to extending the same technique to a Ramanujan-type integral on the real line. In §6 we describe some general properties of a q-quadratic lattice and find possible solutions of the Pearson equation for certain choices of the input function. In §7 we give a proof of (1.35) different from the ones that have appeared before. In §8 we deal with the basic bilateral series ${}_8\psi_8$, which is both very well poised and balanced. We give an extension of Askey's integral (1.36) in §9. An evaluation of the integral in (1.12) as well as an extension is given in §10. Finally we give a q-extension of the formulas of §10 in §11 and point out the connection between these results and those in §9. The paper is concluded by giving a summary of beta integrals, old and new, arising out of solutions of Pearson equations on various types of lattices and different kinds of boundary conditions.

2. An evaluation of the beta integral. Let us rewrite the Pearson equation (1.3) as

$$[x^\alpha(1-x)^\beta]' = \alpha x^{\alpha-1}(1-x)^{\beta-1} - (\alpha+\beta)x^\alpha(1-x)^{\beta-1} \tag{2.1}$$

and integrate from 0 to 1. Since $\operatorname{Re}(\alpha,\beta) > 0$, the left side vanishes at both ends and so the right side gives

$$I(\alpha) = \frac{\alpha+\beta}{\alpha} I(\alpha+1), \tag{2.2}$$

where we are denoting the beta integral in (1.1) by $I(\alpha)$, regarded as a function of α. By iterating this formula n times we get

$$\begin{aligned} I(\alpha) &= \frac{(\alpha+\beta)(\alpha+\beta+1)\dots(\alpha+\beta+n-1)}{\alpha(\alpha+1)\dots(\alpha+n-1)} I(\alpha+n) \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha+\beta)} f(\alpha+n,\beta), \end{aligned} \tag{2.3}$$

where

$$f(\alpha+n,\beta) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+n)} \int_0^1 x^{\alpha+n-1}(1-x)^{\beta-1}dx. \tag{2.4}$$

By Stirling's formula, $\Gamma(\alpha+\beta+n)/\Gamma(\alpha+n) \sim n^\beta$ for large n, which suggests a transformation of the integral by $x \to 1-x$ followed by $x \to x/n$. So we have

$$\begin{aligned} &\lim_{n\to\infty} f(\alpha+n,\beta) \\ &= \lim_{n\to\infty} \int_0^n x^{\beta-1}\Big(1-\frac{x}{n}\Big)^{\alpha+n-1} dx = \int_0^\infty x^{\beta-1}e^{-x}dx = \Gamma(\beta). \end{aligned} \tag{2.5}$$

This completes the proof of (1.1), which is almost identical to the proof given in [3]. Observe that this very elementary proof brings out the point quite clearly that the

Pearson equation not only provides the solution $\rho(x)$ but also helps set up a functional equation that eventually leads to an evaluation of the integral of $\rho(x)$ between the zeros of $\sigma(x)$ through a knowledge of an integral of lower order. By lower order in this particular instance means evaluating Euler's beta integral with two parameters from a knowledge of Euler's gamma integral that has one parameter.

3. A finite-difference analogue of the Pearson equation. We shall start with a discrete analogue of a hypergeometric differential equation, of which (1.2) is an example:

$$\tilde{\sigma}(x(s))\frac{\nabla}{\nabla x_1(s)}\left[\frac{\triangle y(s)}{\triangle x(s)}\right]+\frac{\tilde{\tau}(x(s))}{2}\left[\frac{\triangle y(s)}{\triangle x(s)}+\frac{\nabla y(s)}{\nabla x(s)}\right]+\lambda y(s)=0, \tag{3.1}$$

where $x(s)$ is generally a nonuniform lattice, $x_k(s)=x(s+k/2)$, $k=1,2,\dots,\triangle f(s)=f(s+1)-f(s)$, $\nabla f(s)=\triangle f(s-1)$; $\tilde{\sigma}(x(s)),\tilde{\tau}(x(s))$ are polynomials in $x(s)$ of degrees at most 2 and 1, respectively, and λ is a constant. The idea of difference equations on nonuniform lattices seems a very natural one (see, e.g., the standard Russian textbook [51]), but a complete analysis of the solutions, particularly the polynomial solutions, of equations like (3.1) for linear and quadratic lattices as well as their q-analogues, was made by Soviet mathematicians Nikiforov and Uvarov [42]–[44], Nikiforov and Suslov [41], and Nikiforov, Suslov, and Uvarov [39], [40]. See also Suslov [53], and Atakishiyev and Suslov [11] for nonpolynomial solutions of homogeneous and nonhomogeneous forms of (3.1).

Denoting

$$v_1(s)=\frac{\triangle y(s)}{\triangle x(s)},\quad v_2(s)=\frac{\triangle v_1(s)}{\triangle x_1(s)}, \tag{3.2}$$

(3.1) can be written in a more compact form:

$$\sigma(s)v_2(s-1)+\tau(s)v_1(s)+\lambda y(s)=0, \tag{3.3}$$

where

$$\begin{aligned}\tau(s)&=\tilde{\tau}(x(s)),\\ \sigma(s)&=\tilde{\sigma}(x(s))-\tfrac{1}{2}\tau(s)\nabla x_1(s).\end{aligned} \tag{3.4}$$

If $\rho(s)$ satisfies the Pearson-type equation

$$\nabla[\rho_1(s)]=\rho(s)\tau(s)\nabla x_1(s), \tag{3.5}$$

with $\rho_1(s)=\rho(s+1)\sigma(s+1)$, then (3.3) can be expressed in a self-adjoint form:

$$\frac{\nabla}{\nabla x_1(s)}[\rho_1(s)v_1(s)]+\lambda\rho(s)y(s)=0. \tag{3.6}$$

In principle, there is no difficulty in writing down a formal solution of (3.5) in terms of $\tau(s)$ and $\sigma(s)$, but first we have to decide what sort of a lattice we have and what kind of problems we are interested in. If one is concerned about preserving the hypergeometric character of (3.3), namely, that the successive difference-derivatives of $y(s)$, like the ones in (3.2), also satisfy equations of the same type, then, as was shown

in [12], the necessary and sufficient condition for that to happen is that $x(s)$ be of the form

$$x(s) = \begin{cases} C_1 q^{-s} + C_2 q^s & \text{if } q \neq 1, \\ C_1 s^2 + C_2 s & \text{if } q = 1, \end{cases} \tag{3.7}$$

where C_1 and C_2 are arbitrary constants, not both zero. Hence

$$\nabla x_1(s) = \begin{cases} q^{-\frac{1}{2}}(1-q)(C_1 q^{-s} - C_2 q^s), & q \neq 1, \\ 2C_1 s + C_2, & q = 1, \end{cases} \tag{3.8}$$

implying that $\nabla x_1(s)$ is odd in s if $x(s)$ is even, and vice versa. In [12] we used the polynomial types of $\tilde{\sigma}(x(s))$ and $\tilde{\tau}(x(s))$ to define the family of classical orthogonal polynomials and to classify them according to the lattice type. In this paper we shall still restrict ourselves to lattice type (3.7) but relax the requirement of $\tilde{\sigma}(x(s))$ and $\tilde{\tau}(x(s))$ being polynomials of degrees 2 and 1, respectively, by allowing simple poles in addition to the zeros that we already had. This type of Pearson equation plays a crucial role, in the theory of biorthogonal rational functions [49].

There are two possible situations: (i) s is a discrete variable varying in unit steps from $s = a$ to $s = b - 1$ (it is permissible for a to be $-\infty$ and/or b to be $+\infty$); (ii) s is a continuous complex variable in some domain of the complex plane.

In case (i) the Pearson equation (3.5) provides a telescoping situation, so we get the formula

$$\rho(b)\sigma(b) - \rho(a)\sigma(a) = \sum_{s=a}^{b-1} \rho(s)\tau(s) \nabla x_1(s). \tag{3.9}$$

If a and b are finite, then usually they are both zeros of $\rho(s)\sigma(s)$, so the left side vanishes and we get

$$\sum_{s=a}^{b-1} \rho(s)\tau(s) \nabla x_1(s) = 0. \tag{3.10}$$

This can be used to prove most of the finite summation formulas (e.g., Vandermonde, Dixon, and Dougall) and their q-analogues, but that will be of no interest to us here. We shall instead be concerned with the situation when neither a nor b is finite and the expression on the left side of (3.9) is not necessarily zero. We shall deal with this case in later sections.

In the continuous case (ii) we may find a suitable contour C in the complex s-plane that does not go through any singularities of the integrands so that (3.5) gives

$$\int_C \nabla[\rho_1(s)]ds = \int_C \rho(s)\tau(s) \nabla x_1(s)ds. \tag{3.11}$$

If C has the property that

$$\int_C \rho_1(s)ds = \int_{C'} \rho_1(s')ds', \tag{3.12}$$

where C' is the contour obtained from C by the shift $s' = s - 1$, then (3.11) gives

$$\int_C \rho(s)\tau(s) \bigtriangledown x_1(s)ds = 0. \tag{3.13}$$

Note that, by Cauchy's theorem, (3.12) implies that there are no singularities of $\rho_1(s)$ between C and C'.

4. Barnes's second lemma: linear lattice $x(s) = s$. Since $\bigtriangledown x_1(s) = 1$, (3.5) takes the form

$$\frac{\rho(s+1)}{\rho(s)} = \frac{\sigma(s)+\tau(s)}{\sigma(s+1)}, \tag{4.1}$$

so we want $\sigma(s)$ and $\tau(s)$ to be such that both $\sigma(s+1)$ and $\sigma(s)+\tau(s)$ are rational functions of s with simple linear factors. We take

$$\begin{aligned} \sigma(s) &= (s_1 - s)(s_2 - s), \\ \tau(s) &= \frac{(1 - s_3 - s_6)(1 - s_5 - s_6)(s + s_4)}{1 + s - s_6} - (s_1 + s_4)(s_2 + s_4), \end{aligned} \tag{4.2}$$

where s_i, $i = 1, \dots, 6$, are some complex parameters. It is easy to verify that if these parameters satisfy the "balance" condition

$$s_1 + s_2 + s_3 + s_4 + s_5 + s_6 = 1, \tag{4.3}$$

then

$$\sigma(s) + \tau(s) = \frac{(s + s_3)(s + s_4)(s + s_5)}{1 + s - s_6}, \tag{4.4}$$

and hence

$$\frac{\rho(s+1)}{\rho(s)} = \frac{(s + s_3)(s + s_4)(s + s_5)}{(s_1 - 1 - s)(s_2 - 1 - s)(1 + s - s_6)}. \tag{4.5}$$

This has infinitely many solutions, all differing by a periodic factor of period 1. The solution that is appropriate for our purposes is

$$\rho(s) = \frac{\Gamma(s_1 - s)\Gamma(s_2 - s)\Gamma(s_3 + s)\Gamma(s_4 + s)\Gamma(s_5 + s)}{\Gamma(1 + s - s_6)}. \tag{4.6}$$

Since

$$\begin{aligned} \rho_1(s) &= [\sigma(s) + \tau(s)]\rho(s) \\ &= \frac{\Gamma(s_1 - s)\Gamma(s_2 - s)\Gamma(s_3 + 1 + s)\Gamma(s_4 + 1 + s)\Gamma(s_5 + 1 + s)}{\Gamma(2 + s - s_6)}, \end{aligned} \tag{4.7}$$

we find that there are no poles of $\rho_1(s)$ between C, the imaginary axis, and C', the line one unit to the left of C, provided (4.3) holds and

$$\operatorname{Re} s_i > 0, \qquad i = 1, \dots, 5. \tag{4.8}$$

Assuming this to be the case we then have

$$\int_C \nabla[\rho_1(s)]ds = 0. \tag{4.9}$$

So, by (3.13) and (4.2) we get

$$\int_C \rho(s)ds = \frac{(1-s_3-s_6)(1-s_5-s_6)}{(s_1+s_4)(s_2+s_4)} \int_C \frac{s+s_4}{1+s-s_6}\rho(s)ds. \tag{4.10}$$

Let us fix the parameters s_1, s_2, s_3, and s_5, and think of $\rho(s)$ as a function of s and s_4. Note that, by (4.3), s_6 decreases by 1 if s_4 increases by 1. So, denoting

$$\frac{1}{2\pi i}\int_C \rho(s)ds = I(s_4), \tag{4.11}$$

we find that

$$\frac{1}{2\pi i}\int_C \frac{s+s_4}{1+s-s_6}\rho(s)ds = I(s_4+1). \tag{4.12}$$

So, (4.10) becomes a 2-term recurrence relation

$$I(s_4) = \frac{(1-s_3-s_6)(1-s_5-s_6)}{(s_1+s_4)(s_2+s_4)} I(s_4+1), \tag{4.13}$$

which, on iteration, gives

$$I(s_4) = \frac{\Gamma(s_1+s_4)\Gamma(s_2+s_4)}{\Gamma(1-s_3-s_6)\Gamma(1-s_5-s_6)}A, \tag{4.14}$$

where

$$\begin{aligned} A &= \lim_{n\to\infty} \frac{\Gamma(1-s_3-s_6+n)\Gamma(1-s_5-s_6+n)}{\Gamma(s_1+s_4+n)\Gamma(s_2+s_4+n)} I(s_4+n) \\ &= \lim_{n\to\infty} \frac{1}{2\pi i}\int_C \Gamma(s_1-s)\Gamma(s_2-s)\Gamma(s_3+s)\Gamma(s_5+s) \\ &\quad \cdot \frac{\Gamma(1-s_3-s_6+n)\Gamma(1-s_5-s_6+n)\Gamma(s_4+s+n)}{\Gamma(s_1+s_4+n)\Gamma(s_2+s_4+n)\Gamma(1+s-s_6+n)}ds. \end{aligned} \tag{4.15}$$

By Stirling's formula and (4.3),

$$\begin{aligned} &\frac{\Gamma(1-s_3-s_6+n)\Gamma(1-s_5-s_6+n)\Gamma(s_4+s+n)}{\Gamma(s_1+s_4+n)\Gamma(s_2+s_4+n)\Gamma(1+s-s_6+n)} \\ &\quad \sim n^{1-(s_1+s_2+s_3+s_4+s_5+s_6)} = 1 \quad \text{for large } n. \end{aligned} \tag{4.16}$$

Also, by Barnes's first lemma (1.6),

$$\begin{aligned} &\frac{1}{2\pi i}\int_C \Gamma(s_1-s)\Gamma(s_2-s)\Gamma(s_3+s)\Gamma(s_5+s)ds \\ &\quad = \frac{\Gamma(s_1+s_3)\Gamma(s_1+s_5)\Gamma(s_2+s_3)\Gamma(s_2+s_5)}{\Gamma(s_1+s_2+s_3+s_5)}. \end{aligned} \tag{4.17}$$

(Incidentally, an alternate proof of (1.6) based on somewhat similar ideas was given by Miller [37] and Kalnins and Miller [34].) Combining (4.6), (4.11), (4.14)–(4.17), we find that

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\frac{\Gamma(s_1-s)\Gamma(s_2-s)\Gamma(s_3+s)\Gamma(s_4+s)\Gamma(s_5+s)}{\Gamma(s_1+s_2+s_3+s_4+s_5+s)}ds \tag{4.18}$$
$$=\frac{\Gamma(s_1+s_3)\Gamma(s_1+s_4)\Gamma(s_1+s_5)\Gamma(s_2+s_3)\Gamma(s_2+s_4)\Gamma(s_2+s_5)}{\Gamma(s_1+s_2+s_3+s_4)\Gamma(s_1+s_2+s_3+s_5)\Gamma(s_1+s_2+s_4+s_5)},$$

which is essentially the same as (1.7). (It would be the same if one of s_1, s_2 were zero and the contour were indented around the origin.)

5. Extension of Ramanujan's integral: linear lattice $x(s)=s$. Here we start with the same lattice and the same expressions for $\sigma(s)$ and $\tau(s)$ as in §4 so that the formulas (4.1)–(4.5) all hold. But now that the line of integration is the whole real line instead of the imaginary axis, the asymptotics are entirely different, so the appropriate solution of (4.5) is

$$\rho(s)\equiv\rho(s,s_4) \tag{5.1}$$
$$=\frac{\Gamma(s_6-s)}{\Gamma(1+s-s_1)\Gamma(1+s-s_2)\Gamma(1-s_3-s)\Gamma(1-s_4-s)\Gamma(1-s_5-s)}.$$

From (4.2) and (5.1) we have

$$\rho_1(s-1)=\rho(s)\sigma(s) \tag{5.2}$$
$$=\frac{\Gamma(s_6-s)}{\Gamma(s-s_1)\Gamma(s-s_2)\Gamma(1-s_3-s)\Gamma(1-s_4-s)\Gamma(1-s_5-s)}.$$

To investigate the behaviour of $\rho_1(s-1)$ as $s\to-\infty$ let us first split it into two factors:

$$\rho_1(s-1)=\mu_1(s)g_1(s), \tag{5.3}$$

where

$$g_1(s)=\frac{\Gamma(1+s_1-s)\Gamma(1+s_2-s)\Gamma(s_6-s)}{\Gamma(1-s_3-s)\Gamma(1-s_4-s)\Gamma(1-s_5-s)}, \tag{5.4}$$

and

$$\mu_1(s)=\pi^{-2}\sin\pi(s-s_1)\sin\pi(s-s_2). \tag{5.5}$$

Clearly, $\mu_1(s\pm1)=\mu_1(s)$ and $|\mu_1(s)|\le\pi^{-2}$ for all real s. Also, by (4.3) and Stirling's formula,

$$\lim_{s\to-\infty}g_1(s)=1. \tag{5.6}$$

For s near $+\infty$ we take

$$\rho_1(s-1)=\mu_2(s)g_2(s) \tag{5.7}$$

with

$$g_2(s)=\frac{\Gamma(s_3+s)\Gamma(s_4+s)\Gamma(s_5+s)}{\Gamma(s-s_1)\Gamma(s-s_2)\Gamma(1+s-s_6)} \tag{5.8}$$

and

$$\mu_2(s) = \pi^{-2}\frac{\sin\pi(s+s_3)\sin\pi(s+s_4)\sin\pi(s+s_5)}{\sin\pi(s_6-s)}. \tag{5.9}$$

Once again,

$$\lim_{s\to\infty} g_2(s) = 1 \tag{5.10}$$

and $\mu_2(s \pm 1) = \mu_2(s)$. But now $\mu_2(s)$ is not bounded for all real s unless Im $s_6 \neq 0$.

By defining

$$\int_{-\infty}^{\infty} f(s)ds = \lim_{M,N\to\infty}\int_{-N-\epsilon_1}^{M+\epsilon_2} f(s)ds, \tag{5.11}$$

where M, N are positive integers and ϵ_1, ϵ_2 are some real constants with $0 \leq \epsilon_1, \epsilon_2 < 1$, we will now compute the integral $\int_{-\infty}^{\infty} \nabla[\rho_1(s)]ds$. Since

$$\int_a^b \nabla[\rho_1(s)]ds = \int_b^{b+1} \rho_1(s-1)ds - \int_a^{a+1} \rho_1(s-1)ds, \tag{5.12}$$

we have

$$\begin{aligned}
&\int_{-N-\epsilon_1}^{M+\epsilon_2} \nabla[\rho_1(s)]ds \\
&= \int_{M+\epsilon_2}^{M+\epsilon_2+1} \mu_2(s)g_2(s)ds - \int_{-N-\epsilon_1}^{1-N-\epsilon_1} \mu_1(s)g_1(s)ds \\
&= \int_{\epsilon_2}^{\epsilon_2+1} \mu_2(M+s)g_2(M+s)ds - \int_{-\epsilon_1}^{1-\epsilon_1} \mu_1(s-N)g_1(s-N)ds \\
&= \int_{\epsilon_2}^{\epsilon_2+1} \mu_2(s)g_2(M+s)ds - \int_{-\epsilon_1}^{1-\epsilon_1} \mu_1(s)g_1(s-N)ds,
\end{aligned} \tag{5.13}$$

because of the periodicity of μ_1 and μ_2. So, by (5.6) and (5.10) we get

$$\begin{aligned}
\int_{-\infty}^{\infty} \nabla[\rho_1(s)]ds &= \int_{\epsilon_2}^{\epsilon_2+1} \mu_2(s)ds - \int_{-\epsilon_1}^{1-\epsilon_1} \mu_1(s)ds \\
&= \int_0^1 [\mu_2(s) - \mu_1(s)]ds = \mu, \quad \text{say.}
\end{aligned} \tag{5.14}$$

From (5.5) and (5.9) it is clear that the value of μ remains unchanged if we move s_4 one unit up or down as long as (4.3) holds. We now replace $s_1, s_2, s_3, s_4, s_5, s_6$ by $1-a, 1-b, 1-c, 1-d, 1-e$, and $1-f$, respectively, and find that (3.11) leads to the nonhomogeneous recurrence relation

$$I(d) = \frac{(a+d-1)(b+d-1)}{(2-c-f)(2-e-f)}I(d+1) + \frac{\mu}{(2-c-f)(2-e-f)}, \tag{5.15}$$

where the balance condition (4.3) now reads

$$a+b+c+d+e+f = 5, \tag{5.16}$$

and

$$I(d) = \int_{-\infty}^{\infty} \rho(s)ds = \int_{-\infty}^{\infty} \frac{\Gamma(1-f-s)ds}{\Gamma(a+s)\Gamma(b+s)\Gamma(c-s)\Gamma(d-s)\Gamma(e-s)}. \tag{5.17}$$

It can be shown that, if $\operatorname{Im} f \neq 0$, then

$$|\rho(s)| \sim \begin{cases} (-s)^{-2} & \text{as } s \to -\infty, \\ s^{-2} & \text{as } s \to \infty, \end{cases} \tag{5.18}$$

so that the improper integral in (5.17) converges. Iterating (5.15) n times and then proceeding to the limit $n \to \infty$, we find that

$$I(d) = \frac{\Gamma(2-c-f)\Gamma(2-e-f)}{\Gamma(a+d-1)\Gamma(b+d-1)}B + \frac{\mu}{(2-c-f)(2-e-f)}\, {}_3F_2\left[\begin{matrix} a+d-1, b+d-1, 1 \\ 3-c-f, 3-e-f \end{matrix}; 1\right], \tag{5.19}$$

where

$$B = \lim_{n\to\infty} \int_{-\infty}^{\infty} \frac{\Gamma(a+d+n-1)\Gamma(b+d+n-1)\Gamma(1-f-s+n)}{\Gamma(2-c-f+n)\Gamma(2-e-f+n)\Gamma(d-s+n)} \cdot \frac{ds}{\Gamma(a+s)\Gamma(b+s)\Gamma(c-s)\Gamma(e-s)}, \tag{5.20}$$

provided

$$\operatorname{Re}(2-f) > \max\{\operatorname{Re}(c,d,e)\}. \tag{5.21}$$

By Stirling's formula and Ramanujan's integral (1.8), we have

$$B = \frac{\Gamma(2-d-f)}{\Gamma(a+c-1)\Gamma(a+e-1)\Gamma(b+c-1)\Gamma(b+e-1)}. \tag{5.22}$$

We thus have the formula

(5.23)

$$\int_{-\infty}^{\infty} \frac{\Gamma(1-f-x)dx}{\Gamma(a+x)\Gamma(b+x)\Gamma(c-x)\Gamma(d-x)\Gamma(e-x)} = \frac{\Gamma(2-c-f)\Gamma(2-d-f)\Gamma(2-e-f)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(a+e-1)\Gamma(b+c-1)\Gamma(b+d-1)\Gamma(b+e-1)} + \frac{\mu}{(2-c-f)(2-e-f)}\, {}_3F_2\left[\begin{matrix} a+d-1, b+d-1, 1 \\ 3-c-f, 3-e-f \end{matrix}; 1\right],$$

where the parameters a, b, c, d, e, f satisfy the conditions (i)–(iii) in (1.10), and, by (5.5), (5.9), and (5.14),

$$\mu = \frac{1}{\pi^2}\int_0^1 \Big[\frac{\sin\pi(c-x)\sin\pi(d-x)\sin\pi(e-x)}{\sin\pi(f+x)} - \sin\pi(a+x)\sin\pi(b+x)\Big]dx. \tag{5.24}$$

If the parameters are such that $\mu = 0$, then we have the integration formula (1.9) subject to all four conditions stated in (1.10).

By defining

$$\sum_{-\infty}^{\infty} g(s) = \lim_{k,\ell\to\infty} \sum_{s=\alpha-k}^{\alpha+\ell-1} g(s), \tag{5.25}$$

we obtain, by a similar analysis, that

(5.26)

$$\sum_{n=-\infty}^{\infty} \frac{\Gamma(1-f-\alpha-n)}{\Gamma(a+\alpha+n)\Gamma(b+\alpha+n)\Gamma(c-\alpha-n)\Gamma(d-\alpha-n)\Gamma(e-\alpha-n)}$$
$$= \frac{\Gamma(2-c-f)\Gamma(2-d-f)\Gamma(2-e-f)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(a+e-1)\Gamma(b+c-1)\Gamma(b+d-1)\Gamma(b+e-1)}$$
$$+ \frac{\mu_2(\alpha)-\mu_1(\alpha)}{(2-c-f)(2-e-f)}\,{}_3F_2\left[\begin{matrix} a+d-1, b+d-1, 1 \\ 3-c-f, 3-e-f \end{matrix}; 1\right],$$

with

$$\mu_1(\alpha) = \sin\pi(a+\alpha)\sin\pi(b+\alpha)/\pi^2, \tag{5.27}$$
$$\mu_2(\alpha) = \sin\pi(c-\alpha)\sin\pi(d-\alpha)\sin\pi(e-\alpha)/\pi^2\sin\pi(f+\alpha),$$

and the parameters satisfying the same balance condition (5.16).

Note that the sum on the left side of (5.26) and the first term on the right are symmetrical in c, d, e but the ${}_3F_2$ series is apparently not. This can be rectified by use of the identity [14, Ex. 7, p.98] which, along with (5.16), gives

$$\begin{aligned} &{}_3F_2\left[\begin{matrix} a+d-1, b+d-1, 1; \\ 3-c-f, 3-e-f; \end{matrix} 1\right] \\ &= \frac{2-e-f}{2-d-f}\,{}_3F_2\left[\begin{matrix} a+e-1, b+e-1, 1; \\ 3-c-f, 3-d-f; \end{matrix} 1\right] \\ &= \frac{2-c-f}{2-d-f}\,{}_3F_2\left[\begin{matrix} a+c-1, b+c-1, 1; \\ 3-d-f, 3-e-f; \end{matrix} 1\right], \end{aligned} \tag{5.28}$$

all three ${}_3F_2$'s being convergent because of (5.21). One may deduce from (5.26) the summation formula

$$\sum_{n=-\infty}^{\infty} \frac{(1+\alpha-c)_n(1+\alpha-d)_n(1+\alpha-e)_n}{(\alpha+a)_n(\alpha+b)_n(1+\alpha-a-b-d-e)_n} \tag{5.29}$$
$$= \frac{\Gamma(\alpha+a)\Gamma(\alpha+b)\Gamma(c-\alpha)\Gamma(d-\alpha)\Gamma(e-\alpha)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(a+e-1)\Gamma(a+b+c+d+e-4-\alpha)}$$
$$\cdot \frac{\Gamma(a+b+c+d-3)\Gamma(a+b+c+e-3)\Gamma(a+b+d+e-3)}{\Gamma(b+c-1)\Gamma(b+d-1)\Gamma(b+e-1)},$$

provided

$$\sin\pi(c-\alpha)\sin\pi(d-\alpha)\sin\pi(e-\alpha) \tag{5.30}$$
$$= \sin\pi(a+\alpha)\sin\pi(b+\alpha)\sin\pi(a+b+c+d+e-\alpha).$$

This is a bilateral extension of the Saalschütz formula [52, III. 2, p. 243]. A q-extension is given in §12.

It may be pointed out that by appropriate shift of the parameters we may set $\alpha = 0$ in (5.26), (5.27), (5.29), and (5.30), without any loss of generality.

6. The q-quadratic lattice $x(s) = C_1q^{-s} + C_2q^s$. It would appear that the q-quadratic lattice is far more complicated than the simple linear lattices considered in the previous two sections. However, there is a symmetry in the quadratic case that makes the analysis just as simple as in the linear cases. If we denote $\alpha = C_2/C_1$, $C_1 \neq 0$, then it follows from (3.7) and (3.8) that

$$x(s) = \begin{cases} x(-s-\nu) & \text{if } q \neq 1, \\ x(-s-\alpha) & \text{if } q = 1, \end{cases} \tag{6.1}$$

$$\nabla x_1(s) = \begin{cases} -\nabla x_1(t)|_{t=-s-\nu} & \text{if } q \neq 1, \\ -\nabla x_1(t)|_{t=-s-\alpha} & \text{if } q = 1, \end{cases} \tag{6.2}$$

where $\nu = \log\alpha/\log q$, that is, $\alpha = q^\nu$ when $q \neq 1$, it being understood that the constants C_1, C_2 are not the same in the two cases. Because of (3.4) we are then led to the property

$$\sigma(s) + \tau(s)\nabla x_1(s) = \begin{cases} \sigma(-s-\nu), & q \neq 1, \\ \sigma(-s-\alpha), & q = 1, \end{cases} \tag{6.3}$$

so that the Pearson equation (3.5) takes the form

$$\frac{\rho(s+1)}{\rho(s)} = \sigma^{-1}(s+1)\begin{cases} \sigma(-s-\nu), & q \neq 1, \\ \sigma(-s-\alpha), & q = 1, \end{cases} \tag{6.4}$$

which means that we only need $\sigma(s)$, and not $\tau(s)$ as well, to determine $\rho(s)$. Following the usual structure of the Pearson-type equations discussed in [49] we take

$$\sigma(s) = q^{-2s}\prod_{k=1}^{5}(q^s - s_k)/(q^s - q/s_6), \qquad q \neq 1, \tag{6.5}$$

where $s_1, \dots, s_6$ are six complex parameters that satisfy the balance condition

$$\alpha^2 s_1 s_2 s_3 s_4 s_5 s_6 = q, \qquad \alpha = C_2/C_1 = q^\nu. \tag{6.6}$$

In the $q = 1$ case, the most important quadratic lattice is $x(s) = s^2$, in which case $\alpha = 0$, and

$$\sigma(s) = \frac{(s - s_1)(s - s_2)(s - s_3)(s - s_4)(s - s_5)}{(s - s_1 - s_2 - s_3 - s_4 - s_5)}. \tag{6.7}$$

For now we shall restrict our analysis to the $q \neq 1$ case and return to the $q = 1$ case later. Using (6.2), (6.3), (6.5), and (6.6) we find that

$$\tau(s) = A + B\frac{x(s_0) - x(s)}{x(s_\infty) - x(s)}, \tag{6.8}$$

where

$$q^{s_0} = s_1, \qquad q^{s_\infty} = q/s_6 = \alpha^2 s_1 s_2 s_3 s_4 s_5, \tag{6.9}$$

$$A = \frac{q^{\frac{1}{2}}}{C_1 \alpha^2 s_1 (1 - q)(1 - \alpha q s_1/s_6)} \prod_{k=2}^{5} (1 - \alpha s_1 s_k), \tag{6.10}$$

$$B = \frac{q^{\frac{1}{2}} s_6}{C_1 \alpha^2 q (1 - q)(1 - \alpha q s_1/s_6)} \prod_{k=2}^{5} (1 - q/s_k s_6). \tag{6.11}$$

Clearly, we could have chosen any one of the five parameters $s_1, \dots, s_5$ as q^{s_0} (with corresponding changes in A and B), so this symmetry must be reflected in the final results. We justify the notation s_0 and s_∞ by pointing out that $\sigma(s_0) = 0$ and $\sigma(s_\infty) = \infty$.

Equations (6.4) and (6.5) give

$$\frac{\rho(s+1)}{\rho(s)} = q^{4s+2\nu+2} \frac{1 - s_6 q^s}{1 - s_6 q^{-s-\nu-1}} \prod_{k=1}^{5} \frac{(1 - q^{-s-\nu}/s_k)}{(1 - q^{s+1}/s_k)} \tag{6.12}$$

or

$$\frac{\rho(s+1)}{\rho(s)} = q^{-4s-2\nu-2} \frac{1 - q^{-s}/s_6}{1 - q^{s+\nu+1}/s_6} \prod_{k=1}^{5} \frac{(1 - s_k q^{s+\nu})}{(1 - s_k q^{-s-1})}. \tag{6.13}$$

As we shall see later the expression in (6.12) is suitable for summations and integrations along the real line, while that in (6.13) is more suitable for integration along the imaginary axis.

The general solution of (6.12) is

$$\rho(s) = \rho_0(s) \frac{\prod_{k=1}^{5} (q^{s+1}/s_k, q^{1-s-\nu}/s_k; q)_\infty}{(s_6 q^s, s_6 q^{-s-\nu}; q)_\infty}, \tag{6.14}$$

with

$$\frac{\rho_0(s+1)}{\rho_0(s)} = q^{4s+2\nu+2} \quad \forall\, s \in \boldsymbol{C}, \tag{6.15}$$

while the appropriate solution of (6.13) is

$$\rho(s) = p(s)\frac{(q^{s+\nu+1}/s_6, q^{1-s}/s_6; q)_\infty}{\prod_{k=1}^{5}(s_k q^{s+\nu}, s_k q^{-s}; q)_\infty}, \tag{6.16}$$

with

$$\frac{p(s+1)}{p(s)} = q^{-4s-2\nu-2} \quad \forall\, s \in \boldsymbol{C}. \tag{6.17}$$

Since

$$\begin{aligned}\frac{x(s_0)-x(s)}{x(s_\infty)-x(s)} &= \frac{s_1 s_6}{q}\frac{(1-q^s/s_1)(1-q^{-s-\nu}/s_1)}{(1-s_6 q^{s-1})(1-s_6 q^{-s-\nu-1})} \\ &= \frac{q}{s_1 s_6}\frac{(1-s_1 q^{s+\nu})(1-s_1 q^{-s})}{(1-q^{s+\nu+1}/s_6)(1-q^{1-s}/s_6)},\end{aligned} \tag{6.18}$$

we have

$$\begin{aligned}&\frac{x(s_0)-x(s)}{x(s_\infty)-x(s)}\rho(s) \\ &\quad = \begin{cases} \dfrac{s_1 s_6}{q}\rho_0(s)\dfrac{(q^s/s_1, q^{-s-\nu}/s_1; q)_\infty \prod_{k=2}^{\infty}(q^{s+1}/s_k, q^{1-s-\nu}/s_k; q)_\infty}{(s_6 q^{s-1}, s_6 q^{-s-\nu-1}; q)_\infty}, \\ \dfrac{q}{s_1 s_6}p(s)\dfrac{(q^{2-s}/s_6, q^{s+\nu+2}/s_6; q)_\infty}{(s_1 q^{s+\nu+1}, s_1 q^{1-s}; q)_\infty \prod_{k=2}^{5}(s_k q^{s+\nu}, s_k q^{-s}; q)_\infty}, \end{cases}\end{aligned} \tag{6.19}$$

for the solutions (6.14) and (6.16), respectively.

To emphasize the fact that $\rho(s)$ depends on all 5 parameters $s_1, \ldots, s_5$, and that we have chosen s_1 to be q^{s_0} let us denote

$$\rho(s) = \rho(s; s_1). \tag{6.20}$$

Use of (3.5), (3.8), (6.8)–(6.11), and (6.19) then gives

$$\begin{aligned}\nabla[\rho(s+1)\sigma(s+1)] &= \tau(s)\rho(s; s_1) \nabla x_1(s) \\ &= \frac{\prod_{k=2}^{5}(1-q^\nu s_1 s_k)}{s_1(1-q^{\nu+1}s_1/s_6)} q^{-s-2\nu}(1-q^{\nu+2s})\rho(s; s_1) \\ &\quad + D q^{-s-2\nu}(1-q^{\nu+2s})\rho(s; q s_1),\end{aligned} \tag{6.21}$$

where

$$D = \begin{cases} -\dfrac{s_1 s_6^2}{q^2}\dfrac{\prod_{k=2}^{5}(1-q/s_k s_6)}{(1-q^{\nu+1}s_1/s_6)}, \\ \dfrac{q^{\nu+2}}{s_1 s_6^2}\dfrac{\prod_{k=2}^{5}(1-s_k s_6/q)}{(1-s_6 q^{-\nu-1}/s_1)}, \end{cases} \tag{6.22}$$

the first line of (6.22) corresponding to the first two lines of (6.18) and (6.19), and the second line to the second two lines. In either case, the sum or integral of (6.21) sets up a 2-term recurrence relation for the sum or integral over $q^{-s-2\nu}(1-q^{\nu+2s})\rho(s; s_1)$

as a function of s_1. It is homogeneous or nonhomogenous depending on whether or not the sum or integral of the left side of (6.21) produces a zero.

7. Proof of formula (1.35): lattice $x(s) = \frac{1}{2}(q^{-s} + q^{s})$. In this section we shall consider the integral over $\rho(s)$ along a line parallel to the imaginary axis for the symmetric lattice $x(s) = \frac{1}{2}(q^{-s} + q^{s})$ with $\nu = 0$ in (6.16). We take

$$s_1 = a, \quad s_2 = b, \quad s_3 = c, \quad s_4 = d, \quad s_5 = f \tag{7.1}$$

so that (6.6) gives

$$s_6 = \frac{q}{abcdf}, \tag{7.2}$$

and (6.21) gives

$$\begin{aligned} \nabla\,[\rho(s+1)\sigma(s+1)] &= a^{-1}\frac{(1-ab)(1-ac)(1-ad)(1-af)}{(1-a^2bcdf)}q^{-s}(1-q^{2s})\rho(s;a) \\ &\quad - a^{-1}\frac{(1-abcd)(1-abcf)(1-abdf)(1-acdf)}{(1-a^2bcdf)}q^{-s}(1-q^{2s})\rho(s;aq), \end{aligned} \tag{7.3}$$

with

$$\begin{aligned} \rho(s) &= \rho(s;a) \\ &= p(s)\frac{(abcdfq^{s}, abcdfq^{-s};q)_\infty}{(aq^{s}, aq^{-s}, bq^{s}, bq^{-s}, cq^{s}, cq^{-s}, dq^{s}, dq^{-s}, fq^{s}, fq^{-s};q)_\infty}, \end{aligned} \tag{7.4}$$

$$\frac{p(s+1)}{p(s)} = q^{-4s-2}. \tag{7.5}$$

Let

$$p(s) = \frac{\wp(s)}{q^{-s} - q^{s}}, \tag{7.6}$$

so that

$$\begin{aligned} \frac{\wp(s+1)}{\wp(s)} &= -\frac{(1+q^{-s-1})(1-q^{-s-1})}{(1+q^{s})(1-q^{s})}q^{-2s-1} \\ &= \frac{(1+q^{-s-\frac{1}{2}})(1+q^{-s-1})(1-q^{-s-\frac{1}{2}})(1-q^{-s-1})}{(1+q^{s})(1+q^{s+\frac{1}{2}})(1-q^{s})(1-q^{s+\frac{1}{2}})} \end{aligned} \tag{7.7}$$

with solution

$$\begin{aligned} \wp(s) &= (q^{s}, -q^{s}, q^{-s}, -q^{-s}, q^{s+\frac{1}{2}}, -q^{s+\frac{1}{2}}, q^{\frac{1}{2}-s}, -q^{\frac{1}{2}-s};q)_\infty \\ &= (q^{2s}, q^{-2s};q)_\infty, \end{aligned} \tag{7.8}$$

unique, to within a unit-periodic multiplicative factor. Equations (7.4), (7.6), and (7.8) then give

$$\rho(s;a) = (q^{-s} - q^{s})^{-1}\frac{(q^{2s}, q^{-2s}, abcdfq^{s}, abcdfq^{-s};q)_\infty}{(aq^{s}, aq^{-s}, bq^{s}, bq^{-s}, cq^{s}, cq^{-s}, dq^{s}, dq^{-s}, fq^{s}, fq^{-s};q)_\infty}, \tag{7.9}$$

which agrees with the integrand of (1.33) when $f = 0$, so we may set that multiplicative factor equal to 1.

We now assume that

$$\max(|a|, |b|, |c|, |d|, |f|) < 1, \qquad 0 < q < 1 \tag{7.10}$$

and choose the contour C as $-\pi/(\log q^{-1}) \leq \operatorname{Im} s \leq \pi/\log q^{-1}$, and consider the boundary of the parallelogram defined by

$$0 \leq \operatorname{Re} s \leq 1, \qquad -\frac{\pi}{\log q^{-1}} \leq \operatorname{Im} s \leq \frac{\pi}{\log q^{-1}}. \tag{7.11}$$

Since

$$\rho(s)\sigma(s) = q^{3s}\frac{(q^{s+1}, -q^{s+1}, q^{-s}, -q^{-s}, q^{\frac{1}{2}+s}, -q^{\frac{1}{2}+s}; q)_\infty}{(aq^s, aq^{1-s}, bq^s, bq^{1-s}, cq^s, cq^{1-s}; q)_\infty} \cdot \frac{(q^{\frac{1}{2}-s}, -q^{\frac{1}{2}-s}, abcdfq^s, abcdfq^{1-s}; q)_\infty}{(dq^s, dq^{1-s}, fq^s, fq^{1-s}; q)_\infty}, \tag{7.12}$$

there are no poles inside the basic parallelogram in (7.11). So the integral over $\nabla[\rho(s+1)\sigma(s+1)]$ along the boundary of this parallelogram vanishes and hence, due to the $2\pi i/\log q^{-1}$ periodicity of this integrand, we obtain from (7.3) the recurrence relation

$$I(a) = \frac{(1-abcd)(1-abcf)(1-abdf)(1-acdf)}{(1-ab)(1-ac)(1-ad)(1-af)} I(aq), \tag{7.13}$$

where

$$I(a) = \frac{1}{2\pi i}\int_C \frac{(q^s, -q^s, q^{-s}, -q^{-s}, q^{\frac{1}{2}+s}, -q^{\frac{1}{2}+s}, q^{\frac{1}{2}-s}, -q^{\frac{1}{2}-s}; q)_\infty}{(aq^s, aq^{-s}, bq^s, bq^{-s}, cq^s, cq^{-s}, dq^s, dq^{-s}; q)_\infty} \cdot \frac{(abcdfq^s, abcdfq^{-s}; q)_\infty}{(fq^s, fq^{-s}; q)_\infty} ds. \tag{7.14}$$

By the symmetry of the integrand in (7.14) and the transformation

$$q^s = e^{i\theta}, \qquad 0 \leq \theta \leq \pi, \tag{7.15}$$

we may convert (7.13) into a recurrence relation for the real integral

$$J(a) = \int_0^\pi \frac{h(\cos\theta; 1, -1, q^{\frac{1}{2}}, -q^{\frac{1}{2}}, abcdf)}{h(\cos\theta; a, b, c, d, f)} d\theta, \tag{7.16}$$

namely,

$$J(a) = \frac{(1-abcd)(1-abcf)(1-abdf)(1-acdf)}{(1-ab)(1-ac)(1-ad)(1-af)} J(aq). \tag{7.17}$$

Iterating (7.17) n times and then taking the limit $n \to \infty$ we find that

$$J(a) = \frac{(abcd, abcf, abdf, acdf; q)_\infty}{(ab, ac, ad, af; q)_\infty} J(0), \tag{7.18}$$

where

$$J(0) = \int_0^\pi \frac{h(\cos\theta; 1, -1, q^{\frac{1}{2}}, -q^{\frac{1}{2}})}{h(\cos\theta; b, c, d, f)} d\theta \tag{7.19}$$
$$= \frac{2\pi(bcdf; q)_\infty}{(q, bc, bd, bf, cd, cf, df; q)_\infty}, \quad \text{by (1.33).}$$

Combining (7.18) and (7.19) completes the proof of (1.35).

It may be mentioned that a similar method can also be applied to compute the Askey–Wilson integral (1.33). The idea is to reduce the number of parameters down to zero by successive recursions and then compute the simple parameter-free integral $\int_0^\pi h(\cos\theta; 1, -1, q, -q^{1/2})d\theta$, which can be done in a very elementary manner. See [13] and [33] for details. See also [36] for a similar proof.

8. A very well poised bilateral sum: lattice $x(s) = \frac{1}{2}(q^{-s} + q^s)$. We shall now consider the sum over $\rho(s)$ along the real line for the symmetric lattice $x(s) = \frac{1}{2}(q^{-s} + q^s)$ with $\nu = 0$ in (6.14). We replace the parameters $s_1, \ldots, s_6$ by

$$s_1 = q/a, \quad s_2 = q/b, \quad s_3 = q/c, \quad s_4 = q/d, \quad s_5 = q/e, \quad s_6 = q/f, \tag{8.1}$$

so that the balance condition now reads

$$abcdef = q^5. \tag{8.2}$$

Use of (8.1) in (6.14) gives

$$\rho(s) = \rho(s; a)$$
$$= \rho_0(s) \frac{(aq^s, aq^{-s}, bq^s, bq^{-s}, cq^s, cq^{-s}, dq^s, dq^{-s}, eq^s, eq^{-s}; q)_\infty}{(q^{1+s}/f, q^{1-s}/f; q)_\infty}. \tag{8.3}$$

Defining a bilateral sum by the same limit as in (5.25), where the complex parameter α is not the same as the one used in §6, we find, by summing over (6.21) (after replacing a by aq), that

$$\epsilon(\alpha; aq) = \frac{f}{qa^2} \frac{(1 - ab/q)(1 - ac/q)(1 - ad/q)(1 - ae/q)}{1 - f/qa} I(aq)$$
$$- \frac{q^2}{af^2} \frac{(1 - bf/q^2)(1 - cf/q^2)(1 - df/q^2)(1 - ef/q^2)}{1 - f/qa} I(a), \tag{8.4}$$

where $I(a)$ is the bilateral sum

$$I(a) = \lim_{k,\ell\to\infty} \sum_{\alpha-k}^{\alpha+\ell-1} (1 - q^{2s}) q^{-s} (aq^s, aq^{-s}, bq^s, bq^{-s}, cq^s, cq^{-s}; q)_\infty$$
$$\cdot \frac{(dq^s, dq^{-s}, eq^s, eq^{-s}; q)_\infty}{(q^{1+s}/f, q^{1-s}/f; q)_\infty} \rho_0(s) \tag{8.5}$$

and

$$\epsilon(\alpha; aq) = \lim_{\ell\to\infty} \Big\{ \rho_0(\alpha + \ell) q^{-2\alpha-2\ell} (aq^{\alpha+\ell-1}, aq^{-\alpha-\ell}, bq^{\alpha+\ell-1}, bq^{-\alpha-\ell}; q)_\infty$$
$$\cdot \frac{(cq^{\alpha+\ell-1}, cq^{-\alpha-\ell}, dq^{\alpha+\ell-1}, dq^{-\alpha-\ell}, eq^{\alpha+\ell-1}, eq^{-\alpha-\ell}; q)_\infty}{(q^{\alpha+\ell}/f, q^{1-\alpha-\ell}/f; q)_\infty} \Big\} \tag{8.6}$$
$$- \lim_{k\to\infty} \Big\{ \rho_0(\alpha - k) q^{2k-2\alpha} (aq^{\alpha-k-1}, aq^{-\alpha+k}, bq^{\alpha-k-1}, bq^{-\alpha+k}; q)_\infty$$
$$\cdot \frac{(cq^{\alpha-k-1}, cq^{-\alpha+k}, dq^{\alpha-k-1}, dq^{-\alpha+k}, eq^{\alpha-k-1}, eq^{-\alpha+k}; q)_\infty}{(q^{\alpha-k}/f, q^{1-\alpha+k}/f; q)_\infty} \Big\},$$

provided the limits in (8.5) and (8.6) exist. It can be easily verified that

$$\epsilon(\alpha; aq) = \frac{q}{af}\epsilon(\alpha; a), \tag{8.7}$$

so we may write (8.4) as

$$\begin{aligned} I(a) = {} & \frac{(1-ab/q)(1-ac/q)(1-ad/q)(1-ae/q)}{(1-q^2/bf)(1-q^2/cf)(1-q^2/df)(1-q^2/ef)} I(aq) \\ & + \frac{a(1-aq/f)}{(1-q^2/bf)(1-q^2/cf)(1-q^2/df)(1-q^2/ef)} \left(\frac{q}{af}\right) \epsilon(\alpha; a). \end{aligned} \tag{8.8}$$

Before proceeding any further we have to decide what $\rho_0(s)$ must be in order that the limits indicated above exist. Since $\rho_0(s)$ must satisfy (6.15) (with $\nu = 0$), it suffices to take

$$\rho_0(s) = q^{2s^2}. \tag{8.9}$$

We may now split $\rho(s)\sigma(s)$ into two factors, one that converges as $s \to \infty$ and the other that is periodic of unit period in much the same way as in §5, and find that

$$\begin{aligned} \epsilon(\alpha; a) = q^{2\alpha(\alpha-1)} \Big\{ & (aq^{-\alpha}, q^{\alpha+1}/a, bq^{-\alpha}, q^{\alpha+1}/b, cq^{-\alpha}, q^{\alpha+1}/c; q)_\infty \\ & \cdot \frac{(dq^{-\alpha}, q^{\alpha+1}/d, eq^{-\alpha}, q^{\alpha+1}/e; q)_\infty}{(fq^\alpha, q^{1-\alpha}/f; q)_\infty} \\ & - (aq^{\alpha-1}, q^{2-\alpha}/a, bq^{\alpha-1}, q^{2-\alpha}/b, cq^{\alpha-1}, q^{2-\alpha}/c; q)_\infty \\ & \cdot \frac{(dq^{\alpha-1}, q^{2-\alpha}/d, eq^{\alpha-1}, q^{2-\alpha}/e; q)_\infty}{(fq^{1-\alpha}, q^\alpha/f; q)_\infty} \Big\}. \end{aligned} \tag{8.10}$$

Because of (8.2) the limit in (8.5) also exists, giving us a very well poised bilateral balanced ${}_8\psi_8$ series

$$\begin{aligned} I(a) = {} & q^{2\alpha^2-\alpha}(1-q^{2\alpha})(aq^\alpha, aq^{-\alpha}, bq^\alpha, bq^{-\alpha}, cq^\alpha, cq^{-\alpha}; q)_\infty \\ & \cdot \frac{(dq^\alpha, dq^{-\alpha}, eq^\alpha, eq^{-\alpha}; q)_\infty}{(q^{\alpha+1}/f, q^{1-\alpha}/f; q)_\infty} \\ & \cdot {}_8\psi_8 \left[\begin{matrix} q^{\alpha+1}, -q^{\alpha+1}, q^{\alpha+1}/a, q^{\alpha+1}/b, q^{\alpha+1}/c, q^{\alpha+1}/d, q^{\alpha+1}/e, q^{\alpha+1}/f \\ q^\alpha, -q^\alpha, aq^\alpha, bq^\alpha, cq^\alpha, dq^\alpha, eq^\alpha, fq^\alpha \end{matrix} ; q, q \right]. \end{aligned} \tag{8.11}$$

For the definitions, notation, and properties of various types of basic hypergeometric series see [25].

The limit of $I(a)$, as $a \to 0$, exists provided $|bcde| < q^3$. This limit is

$$\begin{aligned} I(0) = {} & q^{2\alpha^2-\alpha}(1-q^{2\alpha})(bq^\alpha, bq^{-\alpha}, cq^\alpha, cq^{-\alpha}, dq^\alpha, dq^{-\alpha}, eq^\alpha, eq^{-\alpha}; q)_\infty \\ & \cdot {}_6\psi_6 \left[\begin{matrix} q^{\alpha+1}, -q^{\alpha+1}, q^{\alpha+1}/b, q^{\alpha+1}/c, q^{\alpha+1}/d, q^{\alpha+1}/e \\ q^\alpha, -q^\alpha, bq^\alpha, cq^\alpha, dq^\alpha, eq^\alpha \end{matrix} ; q, \frac{bcde}{q^3} \right] \\ = {} & q^{2\alpha^2-\alpha}(q^{2\alpha}, q^{1-2\alpha}; q)_\infty \\ & \cdot \frac{(q, bc/q, bd/q, be/q, cd/q, ce/q, de/q; q)_\infty}{(bcde/q^3; q)_\infty}, \end{aligned} \tag{8.12}$$

by the ${}_6\psi_6$ summation formula [25, eq. (5.3.1), p. 128]. (It is also easy to give an alternate proof of this formula by using the method of this paper.)

We now iterate (8.8) n times, take the limit $n \to \infty$, and use (8.12) to obtain the bilateral extension to Jackson's summation formula; see, for example, [25]:

$$
\begin{aligned}
(8.13)\quad {}_8\psi_8 &\left[\begin{matrix} q^{\alpha+1}, -q^{\alpha+1}, q^{\alpha+1}/a, q^{\alpha+1}/b, q^{\alpha+1}/c, q^{\alpha+1}/d, q^{\alpha+1}/e, q^{\alpha+1}/f \\ q^{\alpha}, -q^{\alpha}, aq^{\alpha}, bq^{\alpha}, cq^{\alpha}, dq^{\alpha}, eq^{\alpha}, fq^{\alpha} \end{matrix}; q, q\right] \\
&= \frac{(q^{1+2\alpha}, q^{1-2\alpha}, q^{\alpha+1}/f, q^{1-\alpha}/f; q)_\infty}{(aq^{\alpha}, aq^{-\alpha}, bq^{\alpha}, bq^{-\alpha}, cq^{\alpha}, cq^{-\alpha}, dq^{\alpha}, dq^{-\alpha}, eq^{\alpha}, eq^{-\alpha}; q)_\infty} \\
&\quad \cdot \frac{(q, ab/q, ac/q, ad/q, ae/q, bc/q, bd/q, be/q, cd/q, ce/q, de/q; q)_\infty}{(q^2/af, q^2/bf, a^2/cf, q^2/df, q^2/ef; q)_\infty} \\
&\quad + \frac{q}{f} \frac{(1 - aq/f)(q^{\alpha+1}/f, q^{1-\alpha}/f; q)_\infty}{(1 - q^{2\alpha})(1 - q^2/bf)(1 - q^2/cf)(1 - q^2/df)(1 - q^2/ef)} \\
&\quad \cdot \frac{q^{\alpha - 2\alpha^2} \epsilon(\alpha; a)}{(aq^{\alpha}, aq^{-\alpha}, bq^{\alpha}, bq^{-\alpha}, cq^{\alpha}, cq^{-\alpha}, dq^{\alpha}, dq^{-\alpha}, eq^{\alpha}, eq^{-\alpha}; q)_\infty} \\
&\quad \cdot {}_8\phi_7 \left[\begin{matrix} aq/f, q(aq/f)^{\frac{1}{2}}, -q(aq/f)^{\frac{1}{2}}, ab/q, ac/q, ad/q, ae/q, q \\ (aq/f)^{\frac{1}{2}}, -(aq/f)^{\frac{1}{2}}, q^3/bf, q^3/cf, q^3/df, q^3/ef, aq/f \end{matrix}; q, bcde/q^3\right],
\end{aligned}
$$

provided

$$
(8.14)\qquad \max(|abcd|, |abce|, |abde|, |acde|, |bcde|) < q^3.
$$

If the parameters satisfy the "regularizing" condition $\epsilon(\alpha; a) = 0$, which, from (8.10), means

$$
\begin{aligned}
(8.15)\quad &\frac{(aq^{-\alpha}, q^{\alpha+1}/a, bq^{-\alpha}, q^{\alpha+1}/b, cq^{-\alpha}, q^{\alpha+1}/c, dq^{-\alpha}, q^{\alpha+1}/d; q)_\infty}{(aq^{\alpha-1}, q^{2-\alpha}/a, bq^{\alpha-1}, q^{2-\alpha}/b, cq^{\alpha-1}, q^{2-\alpha}/c, dq^{\alpha-1}, q^{2-\alpha}/d; q)_\infty} \\
&\quad \cdot \frac{(eq^{-\alpha}, q^{\alpha+1}/e, fq^{1-\alpha}, q^{\alpha}/f; q)_\infty}{(eq^{\alpha-1}, q^{2-\alpha}/e, fq^{\alpha}, q^{1-\alpha}/f; q)_\infty} = 1,
\end{aligned}
$$

then the second term on the right side of (8.13) drops out and we obtain the summation formula due to Gosper [26], see also [25, Ex. 5.12, p. 135].

Clearly, (8.13) is a generalization of (5.26).

9. An extension of Askey's integral (1.36): lattice $x(s) = \frac{1}{2}(q^{-s} - q^{s})$. In this section we find an integral analogue of (8.13) in the same sense as (5.23) is an integral analogue of (5.26). Since the integral is from $-\infty$ to ∞ we have to avoid an integrand that is odd in s, so we take $x(s) = \frac{1}{2}(q^{-s} - q^{s})$ with

$$
(9.1)\qquad \nabla x_1(s) = \frac{1 - q}{2q^{\frac{1}{2}}}(q^{-s} + q^{s}).
$$

The parameter α of §6 is then -1 so that $\nu = \pi i / \log q$. In order to preserve the same basic character of the factors in the first term on the right side of (8.13) corresponding

to $\nu = 0$, we must take all of the parameters $a, \dots, f$ purely imaginary. We choose $\sigma(s)$ and $\tau(s)$ exactly the same as in §8, but replace $s_1, \dots, s_6$ by

$$s_1 = q/ia, \quad s_2 = q/ib, \quad s_3 = q/ic, \quad s_4 = q/id, \quad s_5 = q/if, \quad s_6 = q/ig \tag{9.2}$$

so that the balance condition (6.6) now becomes

$$abcdfg = -q^5. \tag{9.3}$$

We have avoided the symbol e for a parameter since we are going to use it for the standard exponential base a little later. Solution of the Pearson equation now has the form

$$\begin{aligned}\rho(s) &= \rho(s;a) \\ &= \rho_0(s)\frac{(iaq^s, -iaq^{-s}, ibq^s, -ibq^{-s}, icq^s, -icq^{-s}; q)_\infty}{(-iq^{s+1}/g, iq^{1-s}/g; q)_\infty} \\ &\quad \cdot (idq^s, -idq^{-s}, ifq^s, -ifq^{-s}; q)_\infty,\end{aligned} \tag{9.4}$$

with

$$\frac{\rho_0(s+1)}{\rho_0(s)} = q^{4s+2}, \tag{9.5}$$

and

$$\begin{aligned}\rho(s)\sigma(s) &= \rho_0(s)q^{-2s}(iaq^{s-1}, -iaq^{-s}, ibq^{s-1}, -ibq^{-s}; q)_\infty \\ &\quad \cdot \frac{(icq^{s-1}, -icq^{-s}, idq^{s-1}, -idq^{-s}, ifq^{s-1}, -ifq^{-s}; q)_\infty}{(-iq^s/g, iq^{1-s}/g; q)_\infty}.\end{aligned} \tag{9.6}$$

As in (5.12),

$$\begin{aligned}\int_{-\infty}^{\infty} \nabla[\rho_1(s)]ds &= \lim_{M,N\to\infty} \int_{-N-\epsilon_1}^{M+\epsilon_2} \nabla[\rho_1(s)]ds \\ &= \lim_{M\to\infty} \int_{M+\epsilon_2}^{M+\epsilon_2+1} \rho(s)\sigma(s)ds - \lim_{N\to\infty} \int_{-N-\epsilon_1}^{-N-\epsilon_1-1} \rho(s)\sigma(s)ds,\end{aligned} \tag{9.7}$$

where M, N are positive integers and $0 \le \epsilon_1, \epsilon_2 < 1$. Note that

$$\rho(s)\sigma(s) = \mu_2(s)g_2(s), \tag{9.8}$$

where

$$g_2(s) = \frac{(iaq^{s-1}, ibq^{s-1}, icq^{s-1}, idq^{s-1}, ifq^{s-1}, -igq^s; q)_\infty}{(iq^{s+1}/a, iq^{s+1}/b, iq^{s+1}/c, iq^{s+1}/d, iq^{s+1}/f, -iq^s/g; q)_\infty}, \tag{9.9}$$

and

$$\begin{aligned}\mu_2(s) &= \rho_0(s)q^{-2s}(-iaq^{-s}, iq^{s+1}/a, -ibq^{-s}, iq^{s+1}/b; q)_\infty \\ &\quad \cdot \frac{(-icq^{-s}, iq^{s+1}/c, -idq^{-s}, iq^{s+1}/d, -ifq^{-s}, iq^{s+1}/f; q)_\infty}{(iq^{1-s}/g, -igq^s; q)_\infty},\end{aligned} \tag{9.10}$$

so that

$$\mu_2(s+1) = \mu_2(s) \quad \text{by (9.5)} \quad \text{for all } s \in \boldsymbol{C}. \tag{9.11}$$

Hence

$$\lim_{M\to\infty} \mu_2(s+M) = \mu_2(s), \qquad M \text{ an integer.} \tag{9.12}$$

From (9.9) it is also clear that $\lim_{M\to\infty} g_2(s+M) = 1$, so we have

$$\lim_{M\to\infty} \int_{M+\epsilon_2}^{M+\epsilon_2+1} \rho(s)\sigma(s)ds = \int_{\epsilon_2}^{\epsilon_2+1} \mu_2(s)ds = \int_0^1 \mu_2(s)ds, \tag{9.13}$$

by (9.11). Similarly we can show that

$$\lim_{N\to\infty} \int_{-N-\epsilon_1}^{1-N-\epsilon_1} \rho(s)\sigma(s)ds = \int_{-\epsilon_1}^{1-\epsilon_1} \mu_1(s)ds = \int_0^1 \mu_1(s)ds, \tag{9.14}$$

where

$$\begin{aligned}\mu_1(s) = \rho_0(s)q^{-2s}&(iaq^{s-1}, -iq^{2-s}/a, ibq^{s-1}, -iq^{2-s}/b; q)_\infty \\ &\cdot \frac{(icq^{s-1}, -iq^{2-s}/c, idq^{s-1}, -iq^{2-s}/d, ifq^{s-1}, -iq^{2-s}/f; q)_\infty}{(-iq^s/g, iq^{1-s}g; q)_\infty},\end{aligned} \tag{9.15}$$

so that

$$\mu_1(s+1) = \mu_1(s). \tag{9.16}$$

Using (6.8)–(6.11) as well as (9.1)–(9.4) we now find that

$$\begin{aligned}\tau(s)\rho(s;a) \bigtriangledown x_1(s) &= \frac{ia}{q}(q^{-s}+q^s)\rho_0(s)\frac{(1-q^2/ab)(1-q^2/ac)(1-q^2/ad)(1-q^2/af)}{(1+gq/a)} \\ &\cdot (iaq^s, -iaq^{-s}, ibq^s, -ibq^{-s}, icq^s, -icq^{-s}; q)_\infty \\ &\cdot \frac{(idq^s, -idq^{-s}, ifq^s, -ifq^{-s}; q)_\infty}{(-iq^{s+1}/g, iq^{1-s}/g; q)_\infty} \\ &- \frac{iq}{ag^2}(q^{-s}+q^s)\rho_0(s)\frac{(1+bg/q)(1+cg/q)(1+dg/q)(1+fg/q)}{(1+gq/a)} \\ &\cdot (iaq^{s-1}, -iaq^{-s-1}, ibq^s, -ibq^{-s}, icq^s, -icq^{-s}; q)_\infty \\ &\cdot \frac{(idq^s, -idq^{-s}, ifq^s, -ifq^{-s}; q)_\infty}{(-iq^s/g, iq^{-s}/g; q)_\infty}.\end{aligned} \tag{9.17}$$

We now have to choose $\rho_0(s)$ so that (9.5) holds, the integrals on the whole real line over the two expressions on the right side of (9.17) converge, and such that the integral over the first term tends to an Askey integral of the type (1.36) when $a \to 0$. This last requirement eliminates the possibility of $\rho_0(s)$ being of the form (8.9), so we

must look for a unit-periodic function in which powers of q^s occur inside the infinite products only. So let us take

$$\rho_0(s) = \frac{\wp(s)}{q^{-s}+q^s}, \tag{9.18}$$

so that, by (9.5),

$$\begin{aligned} q^{4s+2} &= \frac{\wp(s+1)}{\wp(s)}\frac{q^{-s}+q^s}{q^{-s-1}+q^{s+1}} \\ &= \frac{\wp(s+1)}{\wp(s)}q\frac{1+q^{2s}}{1+q^{2s+2}} \\ &= \frac{\wp(s+1)}{\wp(s)}q\frac{(1+iq^s)(1-iq^s)}{(1+iq^{s+1})(1-iq^{s+1})}, \end{aligned} \tag{9.19}$$

i.e.,

$$\begin{aligned} \frac{\wp(s+1)}{\wp(s)} &= \frac{(1+iq^{s+1})(1-iq^{s+1})}{(1+iq^s)(1-iq^s)}q^{4s+1} \\ &= \frac{(1+iq^{s+1})(1-iq^{s+1})(1+iq^{s+\frac{1}{2}})(1-iq^{s+\frac{1}{2}})}{(1+iq^{-s})(1-iq^{-s})(1+iq^{-s-\frac{1}{2}})(1-iq^{-s-\frac{1}{2}})} \end{aligned} \tag{9.20}$$

with solution

$$\begin{aligned} \wp(s) &= \left[(iq^{\frac{1}{2}+s},-iq^{\frac{1}{2}+s},iq^{\frac{1}{2}-s},-iq^{\frac{1}{2}-s},iq^{s+1},-iq^{s+1},iq^{1-s},-iq^{1-s};q)_\infty\right]^{-1} \\ &= (-q^{1+2s},-q^{1-2s};q)_\infty^{-1}, \end{aligned} \tag{9.21}$$

which is unique to within a periodic factor that we may now take to be unity. Note that (9.21) agrees with the denominator of the integrand in (1.36). Denoting

$$\mu(a) = \int_0^1 [\mu_2(s)-\mu_1(s)]ds \tag{9.22}$$

and

$$\begin{aligned} I(a) = \int_{-\infty}^{\infty} &\frac{(iaq^s,-iaq^{-s},ibq^s,-ibq^{-s},icq^s,-icq^{-s};q)_\infty}{(-q^{1+2s},-q^{1-2s};q)_\infty} \\ &\cdot\frac{(idq^s,-idq^{-s},ifq^s,-ifq^{-s};q)_\infty}{(-iq^{s+1}/g,iq^{1-s}/g;q)_\infty}ds, \end{aligned} \tag{9.23}$$

and then replacing a, g by aq and g/q, respectively, we find that (3.11) in this case leads to the nonhomogeneous recurrence relation

$$\begin{aligned} I(a) = &-\frac{qa^3}{g}\frac{(1-q/ab)(1-q/ac)(1-q/ad)(1-q/af)}{(1+q^2/bg)(1+q^2/cg)(1+q^2/dg)(1+q^2/fg)}I(aq) \\ &-\frac{ai(1+aq/g)}{(1+q^2/bg)(1+q^2/cg)(1+q^2/dg)(1+q^2/fg)}\mu(aq). \end{aligned} \tag{9.24}$$

From (9.10), (9.15), (9.18), and (9.21) it can be verified that

$$\mu(aq) = -\frac{q}{ag}\mu(a). \tag{9.25}$$

Iterating (9.24) n times and then taking the limit $n \to \infty$, we obtain

$$\begin{aligned} I(a) = {} & \frac{(ab/q, ac/q, ad/q, af/q; q)_\infty}{(-q^2/bg, -q^2/cg, -q^2/dg, -q^2/fg; q)_\infty} \lim_{n\to\infty} I(aq^n) \\ & - \frac{iq}{g} \frac{(1+aq/g)\mu(a)}{(1+q^2/bg)(1+q^2/cg)(1+q^2/dg)(1+q^2/fg)} \\ & \cdot {}_8\phi_7\left[\begin{matrix} -aq/g, q(-aq/g)^{\frac{1}{2}}, -q(-aq/g)^{\frac{1}{2}}, ab/q, ac/q, ad/q, \\ (-aq/g)^{\frac{1}{2}}, -(-aq/g)^{\frac{1}{2}}, -q^3/bg, -q^3/cg, -q^3/dg, \end{matrix} \right. \\ & \qquad \left.\begin{matrix} af/q, q \\ -q^3/fg, -aq/g \end{matrix}; q, -q^2/ag\right], \end{aligned} \tag{9.26}$$

provided the limit in (9.26) exists and the integral in (9.23) as well as the infinite series in (9.26) converge. By (9.3), this last requirement implies that

$$|bcdf| < q^3. \tag{9.27}$$

It can be shown that if, in addition, Re $(abcdf) \neq 0$ and

$$\max(|abcd|, |abcf|, |abdf|, |acdf|) < q^3, \tag{9.28}$$

then the integral in (9.23) converges and that

$$\begin{aligned} & \lim_{n\to\infty} I(aq^n) \\ & = \int_{-\infty}^{\infty} \frac{(ibq^s, -ibq^{-s}, icq^s, -icq^{-s}, idq^s, -idq^{-s}, ifq^s, -ifq^{-s}; q)_\infty}{(-q^{1+2s}, -q^{1-2s}; q)_\infty} ds \\ & = \frac{(q, bc/q, bd/q, bf/q, cd/q, cf/q, df/q; q)_\infty}{(bcdf/q^3; q)_\infty}, \end{aligned} \tag{9.29}$$

by (1.36), exists. Thus we have the extension of Askey's formula, with $q^s = e^u$,

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{h(i\sinh u; a, b, c, d, f)}{h(i\sinh u; q^{\frac{1}{2}}, -q^{\frac{1}{2}}, q, -q, -q/g)} du \\ & = (\log q^{-1}) \frac{(q, ab/q, ac/q, ad/q, af/q, bc/q, bd/q, bf/q, cd/q, cf/q, df/q; q)_\infty}{(abcd/q^3, abcf/q^3, abdf/q^3, acdf/q^3, bcdf/q^3; q)_\infty} \\ & \quad - \frac{iq}{g} \frac{(\log q^{-1})(1+aq/g)\mu(a)}{(1-abcd/q^3)(1-abcf/q^3)(1-abdf/q^3)(1-acdf/q^3)} \\ & \quad \cdot {}_8\phi_7\left[\begin{matrix} -aq/g, q(-aq/g)^{\frac{1}{2}}, -q(-aq/g)^{\frac{1}{2}}, ab/q, ac/q, ad/q, \\ (-aq/g)^{\frac{1}{2}}, -(-aq/g)^{\frac{1}{2}}, -q^3/bg, -q^3/cg, -q^3/dg, \end{matrix} \right. \\ & \qquad \left.\begin{matrix} af/q, q \\ -q^3/fg, -aq/g \end{matrix}; q, bcdf/q^3\right]. \end{aligned} \tag{9.30}$$

By (9.10) and (9.15) $\mu(a)$ can be expressed as a single integral

$$\mu(a) = \int_0^1 \frac{q^{-s}}{(-q^{2s}, -q^{1-2s}; q)_\infty (igq^{s+1}, igq^{1-s}, -iq^s/g, -iq^{-s}/g; q)_\infty} \cdot \left\{ \begin{array}{l} (-iaq^{-s}, iq^{s+1}/a, -ibq^{-s}, iq^{s+1}/b, -icq^{-s}, iq^{s+1}/c; q)_\infty \\ \cdot(-idq^{-s}, iq^{s+1}/d, -ifq^{-s}, iq^{s+1}/f, igq^{s+1}, -iq^{-s}/g; q)_\infty \\ -(iaq^{s-1}, -iq^{2-s}/a, ibq^{s-1}, -iq^{2-s}/b, icq^{s-1}, -iq^{2-s}/c; q)_\infty \\ \cdot(idq^{s-1}, -iq^{2-s}/d, ifq^{s-1}, -iq^{2-s}/f, igq^{1-s}, -iq^s/g; q)_\infty \end{array} \right\} ds. \tag{9.31}$$

It is clear from (9.30) and (9.31) that if any one of the five parameters vanishes, then the second term on the right side of (9.30) drops out leaving us with Askey's formula (9.29).

In concluding this section we would like to remark that while it would not be too difficult to prove (8.13) by using the transformation theory of basic hypergeometric series (see [25]), we do not know how one could discover (9.30) by the same procedure. Formulas (5.23) and (9.30) are of the same basic character, in fact, it is not too hard to show that (5.23) is a special limiting case of (9.30). The appearance of the second term on the right sides of (5.23) and (9.30) is a characteristic of Ramanujan-type integrals of higher order.

10. Proof of (1.12) and an extension: lattice $x(s) = s^2$. In this section we will not only give a rigorous proof of (1.12) but also prove the following extension:

$$\begin{aligned} &\int_{-\infty}^{\infty} \frac{\Gamma(1-2s)\Gamma(1+2s)}{\Gamma(a-s)\Gamma(a+s)\Gamma(b-s)\Gamma(b+s)\Gamma(c-s)\Gamma(c+s)\Gamma(d-s)\Gamma(d+s)} \\ &\quad \cdot \frac{\Gamma(1-f-s)\Gamma(1-f+s)}{\Gamma(e-s)\Gamma(e+s)} ds \\ &= \frac{\Gamma(2-f-a)\Gamma(2-f-b)\Gamma(2-f-c)}{\Gamma(a+b-1)\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(a+e-1)\Gamma(b+c-1)} \\ &\quad \cdot \frac{\Gamma(2-f-d)\Gamma(2-f-e)}{\Gamma(b+d-1)\Gamma(b+e-1)\Gamma(c+d-1)\Gamma(c+e-1)\Gamma(d+e-1)} \\ &\quad + \lambda \frac{1+a-f}{(2-f-b)(2-f-c)(2-f-d)(2-f-e)} \\ &\quad \cdot {}_7F_6\left[\begin{matrix} 1+a-f, 1+\frac{1+a-f}{2}, a+b-1, a+c-1, a+d-1, a+e-1, 1 \\ \frac{1+a-f}{2}, 3-f-b, 3-f-c, 3-f-d, 3-f-e, 1+a-f \end{matrix} ; 1 \right], \end{aligned} \tag{10.1}$$

where

$$a+b+c+d+e+f = 5, \quad \mathrm{Re}\,(2-f) > \mathrm{Re}\,(a,b,c,d,e), \quad \mathrm{Im}\, f < 0, \tag{10.2}$$

and

$$\lambda = \pi \int_0^1 \frac{\lambda(s) - \lambda(-s)}{\sin 2\pi s} ds, \tag{10.3}$$

with

$$\lambda(s) = \frac{\sin\pi(a-s)\sin\pi(b-s)\sin\pi(c-s)\sin\pi(d-s)\sin\pi(e-s)}{\pi^4\sin\pi(f+s)}. \tag{10.4}$$

The symmetry of the second term on the right side of (10.1) can be demonstrated by using a transformation formula for the ${}_7F_6$ series (see [14]) as was done in §5.

It is true, as Askey mentioned in [4], that there is the "immediate problem" of poles of the gamma functions in the numerator of the integrand of (1.12), but they are simple poles and so the problem is immediately resolved by interpreting the integral as a Cauchy principal value integral. The same is true for the integral in (10.1). All we need to do is to find conditions so that these principal value integrals exist and then try to evaluate them.

For the Pearson equation (6.4) we have a very simple relation with the symmetric lattice $x(s) = s^2$:

$$\frac{\rho(s+1)}{\rho(s)} = \frac{\sigma(-s)}{\sigma(s+1)}. \tag{10.5}$$

For Askey's integral (1.12) we take

$$\sigma(s) = (a-1+s)(b-1+s)(c-1+s)(d-1+s), \tag{10.6}$$

and for (10.1) we take

$$\sigma(s) = \frac{(a-1+s)(b-1+s)(c-1+s)(d-1+s)(e-1+s)}{s-f}, \tag{10.7}$$

with $a+b+c+d+e+f = 5$. Clearly, (10.7) reduces to (10.6) in the limit $e \to \infty$. The solution of (10.5) that is appropriate for our purposes is

$$\rho(s) = \frac{\pi}{\sin 2\pi s}\frac{\Gamma(1-f-s)}{\Gamma(a-s)\Gamma(b-s)\Gamma(c-s)\Gamma(d-s)\Gamma(e-s)} \cdot \frac{\Gamma(1-f+s)}{\Gamma(a+s)\Gamma(b+s)\Gamma(c+s)\Gamma(d+s)\Gamma(e+s)}, \tag{10.8}$$

while the corresponding formula in the Askey case is

$$\rho_A(s) = \frac{\pi}{\sin 2\pi s}[\Gamma(a-s)\Gamma(a+s)\Gamma(b-s)\Gamma(b+s)\Gamma(c-s)\Gamma(c+s)\Gamma(d-s)\Gamma(d+s)]^{-1}, \tag{10.9}$$

where the suffix A is used to indicate the limiting case of (10.8). Since $\nabla x_1(s) = 2s$ and $\Gamma(2s)\Gamma(1-2s) = \pi/\sin 2\pi s$, we find that

$$\rho(s) \nabla x_1(s) = \frac{\Gamma(1-2s)\Gamma(1-f-s)}{\Gamma(a-s)\Gamma(b-s)\Gamma(c-s)\Gamma(d-s)\Gamma(e-s)} \cdot \frac{\Gamma(1+2s)\Gamma(1-f+s)}{\Gamma(a+s)\Gamma(b+s)\Gamma(c+s)\Gamma(d+s)\Gamma(e+s)} \tag{10.10}$$

and

$$\rho_A(s) \bigtriangledown x_1(s) = \frac{\Gamma(1-2s)}{\Gamma(a-s)\Gamma(b-s)\Gamma(c-s)\Gamma(d-s)} \cdot \frac{\Gamma(1+2s)}{\Gamma(a+s)\Gamma(b+s)\Gamma(c+s)\Gamma(d+s)}. \tag{10.11}$$

From (10.7) and (10.8) we also have

$$\begin{aligned}\rho_1(s-1) &= \rho(s)\sigma(s) \\ &= \frac{\pi}{\sin 2\pi s} \frac{\Gamma(1-f-s)}{\Gamma(a-s)\Gamma(b-s)\Gamma(c-s)\Gamma(d-s)\Gamma(e-s)} \\ &\quad \cdot \frac{\Gamma(s-f)}{\Gamma(a-1+s)\Gamma(b-1+s)\Gamma(c-1+s)\Gamma(d-1+s)\Gamma(e-1+s)}.\end{aligned} \tag{10.12}$$

So, in order to investigate the asymptotic behaviour of $\rho_1(s-1)$ near $\pm\infty$ we split $\rho_1(s-1)$ as follows:

$$\rho_1(s-1) = \begin{cases} \dfrac{\pi}{\sin 2\pi s} g(-s)\lambda(-s) & \text{for Re } s < 0, \\ \dfrac{\pi}{\sin 2\pi s} g(s-1)\lambda(s) & \text{for Re } s > 0, \end{cases} \tag{10.13}$$

where $\lambda(s)$ is defined in (10.4) and

$$g(-s) = \frac{\Gamma(1-f-s)\Gamma(2-a-s)\Gamma(2-b-s)\Gamma(2-c-s)\Gamma(2-d-s)\Gamma(2-e-s)}{\Gamma(a-s)\Gamma(b-s)\Gamma(c-s)\Gamma(d-s)\Gamma(e-s)\Gamma(1+f-s)}. \tag{10.14}$$

It is clear that $\lambda(s)$ is a periodic function of period 1 and that, by Stirling's formula and (10.2),

$$\lim_{s\to-\infty} g(-s) = \lim_{s\to\infty} g(s-1) = 1, \tag{10.15}$$

the limits being valid along the real line as well as along any line parallel to it.

The factor $\Gamma(1-2s)\Gamma(1+2s)$ has poles at $s = \pm(n+1)/2$, $n = 0, 1, 2, \ldots$. In order to avoid these poles on the real line we first use the following definition of the doubly infinite integral:

$$\int_{-\infty}^{\infty} f(s)ds = \lim_{\substack{N_1,N_2\to\infty, \\ \operatorname{Im}(\epsilon_1+\epsilon_2)=0}} \int_{-N_1-\epsilon_1}^{N_2+\epsilon_2} f(s)ds, \tag{10.16}$$

where N_1, N_2 are positive integers and ϵ_1, ϵ_2 are complex numbers such that $-\operatorname{Im}\,\epsilon_1 = \operatorname{Im}\,\epsilon_2 < |\operatorname{Im}\, f|$ and $0 < \operatorname{Re}\,\epsilon_1, \operatorname{Re}\,\epsilon_2 < \frac{1}{2}$. It means that we are replacing the integral along the real line by an integral along a parallel line just above the real axis. By using the symmetry of the location of the poles on the real line we will eventually prove that the limit on the right side of (10.16) is indeed equal to the integral on the left side interpreted as a principal value integral for (1.12) and (10.1).

As before, we can show that

$$\begin{aligned}(10.17)\qquad &\int_{-N_1-\epsilon_1}^{N_2+\epsilon_2} \nabla\rho_1(s)ds \\ &= \int_{N_2+\epsilon_2}^{N_2+\epsilon_2+1} \rho_1(s-1)ds - \int_{-N_1-\epsilon_1}^{1-N_1-\epsilon_2} \rho_1(s-1)ds \\ &= \int_{\epsilon_2}^{\epsilon_2+1} \frac{\pi}{\sin 2\pi t}\lambda(t)g(t-1+N_2)dt \\ &\quad - \int_{-\epsilon_1}^{1-\epsilon_1} \frac{\pi}{\sin 2\pi t}\lambda(-t)g(N_1-t)dt,\end{aligned}$$

so that

$$\begin{aligned}(10.18)\qquad \int_{-\infty}^{\infty} \nabla\rho_1(s)ds &= \int_{\epsilon_2}^{\epsilon_2+1} \frac{\pi\lambda(t)}{\sin 2\pi t}dt - \int_{-\epsilon_1}^{1-\epsilon_1} \frac{\pi\lambda(-t)}{\sin 2\pi t}dt \\ &= \pi \int_{i\ Im\ \epsilon_2}^{1+i\ Im\ \epsilon_2} \frac{\lambda(t)-\lambda(-t)}{\sin 2\pi t}dt.\end{aligned}$$

For $0 \le t \le 1$ the integrand on the right side of (10.18) has no singularities since the zeros of $\sin 2\pi t$ at $t = 0, \frac{1}{2}, 1$ cancel out with those of $\lambda(t)-\lambda(-t)$. Since by assumption Im $f < 0$, the poles of $\lambda(s)$ are in the lower half plane, so the unit-periodicity of this integrand allows us to reduce it to an integral along a parallel segment of the real line. Thus we have

$$\int_{-\infty}^{\infty} \nabla\rho_1(s)ds = \lambda,$$

where λ is defined in (10.3).

Let us denote

$$(10.19)\qquad I(a) = \int_{-\infty}^{\infty} \rho(s) \nabla x_1(s)ds,$$

where the integral is defined by (10.16), so that Im s = Im ϵ_2. By the procedure outlined in §§6 and 8 it can be easily shown that $I(a)$ satisfies the recurrence relation

$$\begin{aligned}(10.20)\qquad I(a) &= \frac{(a+b-1)(a+c-1)(a+d-1)(a+e-1)}{(2-f-b)(2-f-c)(2-f-d)(2-f-e)}I(a+1) \\ &\quad + \lambda\frac{1+a-f}{(2-f-b)(2-f-c)(2-f-d)(2-f-e)}.\end{aligned}$$

Iterating this formula n times and then taking the limit $n \to \infty$ we find that

$$\begin{aligned}(10.21)\qquad I(a) &= \frac{\Gamma(2-f-b)\Gamma(2-f-c)\Gamma(2-f-d)\Gamma(2-f-e)}{\Gamma(a+b-1)\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(a+e-1)}I_0(b,c,d,e) \\ &\quad + \lambda\frac{1+a-f}{(2-f-b)(2-f-c)(2-f-d)(2-f-e)} \\ &\quad \cdot {}_7F_6\left[\begin{matrix}1+a-f, 1+\frac{1+a-f}{2}, a+b-1, a+c-1, a+d-1, a+e-1, 1 \\ \frac{1+a-f}{2}, 3-f-b, 3-f-c, 3-f-d, 3-f-e, 1+a-f\end{matrix}; 1\right],\end{aligned}$$

where

(10.22)

$$I_0(b,c,d,e) = \lim_{n\to\infty} \int_{-\infty}^{\infty} \left\{ \frac{\Gamma(a+b-1+n)\Gamma(a+c-1+n)\Gamma(a+d-1+n)\Gamma(a+e-1+n)}{\Gamma(2-f-b+n)\Gamma(2-f-c+n)\Gamma(2-f-d+n)\Gamma(2-f-e+n)} \right.$$
$$\left. \cdot \frac{\Gamma(1-f-s+n)\Gamma(1-f+s+n)}{\Gamma(a-s+n)\Gamma(a+s+n)} \right\} \frac{\Gamma(1-2s)}{\Gamma(b-s)\Gamma(c-s)\Gamma(d-s)\Gamma(e-s)}$$
$$\cdot \frac{\Gamma(1+2s)}{\Gamma(b+s)\Gamma(c+s)\Gamma(d+s)\Gamma(e+s)} ds.$$

By (10.2) and Stirling's formula, the expression within the curly brackets is $1+O(1/n)$ as $n \to \infty$. So

(10.23)

$$I_0(b,c,d,e) = \int_{-\infty}^{\infty} \frac{\Gamma(1-2s)\Gamma(1+2s)ds}{\Gamma(b-s)\Gamma(b+s)\Gamma(c-s)\Gamma(c+s)\Gamma(d-s)\Gamma(d+s)\Gamma(e-s)\Gamma(e+s)},$$

which is the same as the integral in (1.12). Gustafson [28] recently found a multidimensional version of this formula, so (1.12) is essentially a special case of his. Gustafson's method, however, is entirely different from ours.

Let us first give a formal evaluation of this integral. The recurrence formula for this integral is

$$I_0(b,c,d,e) = \frac{(b+c-1)(b+d-1)(b+e-1)}{b+c+d+e-3} I_0(b+1,c,d,e), \tag{10.24}$$

which is obtained in a manner similar to (10.20). Since $I_0(b,c,d,e)$ is invariant with respect to a permutation of b,c,d,e it follows that

(10.25)

$$I_0(b,c,d,e) = \frac{\Gamma(b+c+d+e-3)}{\Gamma(b+c-1)\Gamma(b+d-1)\Gamma(b+e-1)\Gamma(c+d-1)\Gamma(c+e-1)\Gamma(d+e-1)} \cdot M(b,c,d,e),$$

where M is invariant under any permutation of b,c,d,e and is unit-periodic in all four parameters. So, iterating the recurrence relation obtained from (10.25) by taking $b \to b+1$, one can show by Stirling's formula that M must be independent of b, and hence must be a numerical constant. Thus,

$$M = \Gamma(c+d-1)\Gamma(c+e-1)\Gamma(d+e-1) \cdot \int_{-\infty}^{\infty} \frac{\Gamma(1-2s)\Gamma(1+2s)ds}{\Gamma(c-s)\Gamma(c+s)\Gamma(d-s)\Gamma(d+s)\Gamma(e-s)\Gamma(e+s)}. \tag{10.26}$$

Setting $c = d = 1$, $e = \frac{1}{2}$, we find that

$$M = \pi \int_{-\infty}^{\infty} \frac{\sin \pi s}{\pi^2 s} ds = 1. \tag{10.27}$$

This completes a formal proof of (1.12) provided that the integral is not along the real line but along a parallel line in the upper half plane. The evaluation of $M(b, c, d, e)$ based on symmetry considerations is a reflection of the influence of the elegant work of Miller [37]. See also [33].

Let us now return to the harder question of how to justify the statement we made prior to (10.17). Let us go back to (3.11), take $N_1 = N_2$ in (10.16) (which we can do without any loss of generality), and define the contour C to be a partially indented rectangular curve as follows:

$$\begin{aligned}
&C_0: && -N - \operatorname{Re} \epsilon_1 \leq s \leq N + \operatorname{Re} \epsilon_2,\ s \text{ real, indented at } s = \\
&&& \pm\frac{n+1}{2},\ n = 0, 1, \ldots, 2N-1 \text{ with semicircles of radius} \\
&&& \delta < \operatorname{Im} \epsilon_2 \text{ in the upper half plane;} \\
&C_R: && \operatorname{Re} s = N + \operatorname{Re} \epsilon_2,\ 0 \leq \operatorname{Im} s \leq \operatorname{Im} \epsilon_2 < |\operatorname{Im} f|; \\
&C_N: && \text{a line segment parallel to the real line from } N + \epsilon_2 \text{ to} \\
&&& -N - \epsilon_1; \\
&C_L: && \operatorname{Re} s = -N - \operatorname{Re} \epsilon_1,\ \operatorname{Im} s \text{ from } \operatorname{Im}(-\epsilon_1) \text{ to } 0.
\end{aligned} \tag{10.28}$$

There are no singularities of $\rho_A(s) \bigtriangledown x_1(s)$ or $\rho(s) \bigtriangledown x_1(s)$ on or inside the contour C, so the integrals along C over both of them vanish.

Since both integrands are symmetric in s we need only to consider their asymptotic behaviour at $+\infty$. On C_R, C_N, and C_L

$$\begin{aligned}
&\rho_A(s) \bigtriangledown x_1(s) \\
&\quad = \frac{\sin \pi(a-s) \sin \pi(b-s) \sin \pi(c-s) \sin \pi(d-s)}{\pi^3 \sin 2\pi s} \\
&\qquad \cdot \frac{\Gamma(1-a+s)\Gamma(1-b+s)\Gamma(1-c+s)\Gamma(1-d+s)}{\Gamma(a+s)\Gamma(b+s)\Gamma(c+s)\Gamma(d+s)} 2^{2s} \\
&\qquad \sim 2^{2s} s^{5-2(a+b+c+d)},
\end{aligned} \tag{10.29}$$

and

$$\begin{aligned}
&\rho(s) \bigtriangledown x_1(s) \\
&\quad = \frac{\sin \pi(a-s) \sin \pi(b-s) \sin \pi(c-s) \sin \pi(d-s) \sin \pi(e-s)}{\pi^3 \sin 2\pi s \sin \pi(f+s)} \\
&\qquad \cdot \frac{\Gamma(1-a+s)\Gamma(1-b+s)\Gamma(1-c+s)\Gamma(1-d+s)\Gamma(1-e+s)\Gamma(1-f+s)}{\Gamma(a+s)\Gamma(b+s)\Gamma(c+s)\Gamma(d+s)\Gamma(e+s)\Gamma(f+s)} 2^{2s} \\
&\qquad \sim s^{-3},
\end{aligned} \tag{10.30}$$

as $s \to +\infty$. So the three integrals over $\rho_A(s)\nabla x_1(s)$ converge if $2(a+b+c+d)-5 > 1$, i.e., $a+b+c+d > 3$ while the ones over $\rho(s) \nabla x_1(s)$ converge due to the conditions (10.2). Having thus established the convergence of these integrals it is easy to see that

$$\lim_{N\to\infty} \int_{C_R} f(s)ds = \lim_{N\to\infty} \int_{C_L} f(s)ds = 0, \tag{10.31}$$

where $f(s) = \rho_A(s) \nabla x_1(s)$ in one and $\rho(s) \nabla x_1(s)$ in the other. Finally, since $f(s)$ in either case has only simple poles on the real line, the contribution from a typical pole s_k and a semicircle C_δ around it has the form $-i\pi$ Res $f(s)|_{s=s_k}$. However, $f(-s) = f(s)$, so the contributions from s_k and $-s_k$ cancel out. Thus we have the result that

$$\lim_{\delta\to 0} \int_{C_0} f(s)ds = -\lim_{N\to\infty} \int_{C_N} f(s)ds, \tag{10.32}$$

where $f(s) = \rho(s) \nabla x_1(s)$ or $\rho_A(s) \nabla x_1(s)$. This completes the proof of both (1.12) and (10.1).

11. A q-analogue of (10.1): lattice $x(s) = \frac{1}{2}(q^{-s} + q^s)$. We shall now prove that a q-analogue of (10.1) is

$$\int_{-\infty}^{\infty} \frac{(aq^{-s}, aq^s, bq^{-s}, bq^s, cq^{-s}, cq^s, dq^{-s}, dq^s, eq^{-s}, eq^s; q)_\infty}{(q^{1-s}/f, q^{1+s}/f, q^{1-2s}, q^{1+2s}; q)_\infty} ds \tag{11.1}$$

$$= \frac{(q, ab/q, ac/q, ad/q, ae/q, bc/q, bd/q, be/q, cd/q, ce/q, de/q; q)_\infty}{(q^2/af, q^2/bf, q^2/cf, q^2/df, q^2/ef; q)_\infty}$$

$$+ \frac{q\lambda_q}{f} \frac{(1-aq/f)}{(1-q^2/bf)(1-q^2/cf)(1-q^2/df)(1-q^2/ef)}$$

$$\cdot {}_8\phi_7 \left[\begin{matrix} aq/f, q(aq/f)^{\frac{1}{2}}, -q(aq/f)^{\frac{1}{2}}, ab/q, ac/q, ad/q, ae/q, q \\ (aq/f)^{\frac{1}{2}}, -(aq/f)^{\frac{1}{2}}, q^3/bf, q^3/cf, q^3/df, q^3/ef, aq/f \end{matrix} ; q, q^2/af \right],$$

where

$$abcdef = q^5, \quad |q^2/f| < \min(|a|, |b|, |c|, |d|, |e|), \quad \text{Im } f \neq 0 \tag{11.2}$$

and

$$\lambda_q = \int_0^1 \frac{\lambda_q(s) - \lambda_q(-s)}{(q^{2s}, q^{1-2s}; q)_\infty} q^s ds \tag{11.3}$$

with

$$\lambda_q(s) = q^{-2s} \frac{(aq^{-s}, bq^{-s}cq^{-s}, dq^{-s}, eq^{-s}; q)_\infty}{(q^{1-s}/f; q)_\infty} \cdot \frac{(q^{s+1}/a, q^{s+1}/b, q^{s+1}/c, q^{s+1}/d, q^{s+1}/e; q)_\infty}{(fq^s; q)_\infty}. \tag{11.4}$$

The integrand on the left side of (11.1) has simple poles at $s = \pm\frac{n+1}{2}$ on the real line, with n a nonnegative integer, so the integral must be interpreted as a principal

value integral whose existence can be proved in much the same way as in §10. We shall also show, as in §10, that this integral has exactly the same value as the nonsingular integral

$$\int_{-\infty}^{\infty} \frac{(aq^{-s}, aq^{s}, bq^{-s}, bq^{s}, cq^{-s}, cq^{s}, dq^{-s}, dq^{s}, eq^{-s}, eq^{s}; q)_\infty}{(q^{1-s}/f, q^{1+s}/f, q^{1-2s}, q^{1+2s}; q)_\infty} ds, \tag{11.5}$$

where $-\infty < \operatorname{Re} s < \infty, \operatorname{Im} s = s_0 > 0$; s_0 and $|\arg f|$ are so related that there are no poles of the integrand in the strip $0 < \operatorname{Im} s \leq s_0$. In particular, if we take $s_0 = \pi/2 \log q^{-1}$, then $q^{is_0} = e^{-is_0 \log q^{-1}} = e^{-\pi i/2} = -i$. This implies that we must take $|\arg f| > \pi/2$ to ensure that the above integral is nonsingular.

A very interesting thing happens when we transform the integral (11.5) by the substitution $s = i\pi/2 \log q^{-1} - t$, t real. It becomes

$$\int_{-\infty}^{\infty} \frac{(iaq^{t}, -iaq^{-t}, ibq^{t}, -ibq^{-t}, icq^{t}, -icq^{-t}, idq^{t}, -idq^{-t}, ieq^{t}, -ieq^{-t}; q)_\infty}{(iq^{1+t}/f, -iq^{1-t}/f, -q^{1+2t}, -q^{1-2t}; q)_\infty} dt \tag{11.6}$$

which is precisely the same as the integral in (9.23) with e and f replaced by f and $-g$, respectively. So, once it is proved that the integrals in (11.1) and (11.5) are the same we have shown that

$$Pv \int_{-\infty}^{\infty} \frac{(aq^{-s}, aq^{s}, bq^{-s}, bq^{s}, cq^{-s}, cq^{s}, dq^{-s}, dq^{s}, eq^{-s}, eq^{s}; q)_\infty}{(q^{1-s}/f, q^{1+s}/f, q^{1-2s}, q^{1+2s}; q)_\infty} ds \tag{11.7}$$
$$= \text{ the integral (11.6)},$$

provided the parameters satisfy (11.2) and $|\arg f| > \pi/2$. In the limit $e \to 0$ this leads to the formula

$$Pv \int_{-\infty}^{\infty} \frac{(aq^{-s}, aq^{s}, bq^{-s}, bq^{s}, cq^{-s}, cq^{s}, dq^{-s}, dq^{s}; q)_\infty}{(q^{1-2s}, q^{1+2s}; q)_\infty} ds \tag{11.8}$$
$$= \int_{-\infty}^{\infty} \frac{(iaq^{t}, -iaq^{-t}, ibq^{t}, -ibq^{-t}, icq^{t}, -icq^{-t}, idq^{t}, -idq^{-t}; q)_\infty}{(-q^{1-2t}, -q^{1+2t}; q)_\infty} dt.$$

The right side is Askey's integral (1.36) and the left side is a q-analogue of an integral that Askey wrote down in [4] but wished for a Ramanujan to evaluate!

Let us consider the integral

$$I = \int_C \frac{(aq^{-s}, aq^{s}, bq^{-s}, bq^{s}, cq^{-s}, cq^{s}, dq^{-s}, dq^{s}, eq^{-s}, eq^{s}; q)_\infty}{(q^{1-s}/f, q^{1+s}/f, q^{1-2s}, q^{1+2s}; q)_\infty} ds, \tag{11.9}$$

where C is the same contour as described in (10.28) except that we now take $\operatorname{Im} \epsilon_2 = s_0 > 0$ and $|\arg f| > s_0 \log q^{-1}$. If we denote by $h(s)$ the integrand of (11.9) then it can be shown that

$$\lim_{N\to\infty} \int_{C_R} h(s) ds = \lim_{N\to\infty} \int_{C_L} h(s) ds = 0 \tag{11.10}$$

and that $\lim_{N\to\infty}\int_{C_N} h(s)ds$ exists because of (11.2) when $e \neq 0$, and, provided $|abcd/q^3| < 1$ when $e = 0$. Since there are no poles on or inside the contour C, the integral in (11.9) vanishes, by Cauchy's theorem. Therefore, analogous to (10.32) we have the result

$$\lim_{\delta\to 0}\int_{C_0} h(s)ds = -\lim_{N\to\infty}\int_{C_N} h(s)ds. \tag{11.11}$$

Because of the symmetry of $h(s)$, the contributions from the symmetrically located poles $s = \pm(k+1)/2$, $k = 0, 2, \ldots, 2N-1$, to the integral over C_0 cancel out. So we have the formula

$$\begin{aligned}\int_{-\infty}^{\infty} &\frac{(aq^{-t-is_0}, aq^{t+is_0}, bq^{-t-is_0}, bq^{t+is_0}, cq^{-t-is_0}, cq^{t+is_0}; q)_\infty}{(q^{1-is_0-t}/f, q^{1+is_0+t}/f; q)_\infty}\\ &\cdot\frac{(dq^{-t-is_0}, dq^{t+is_0}, eq^{-t-is_0}, eq^{t+is_0}; q)_\infty}{(q^{1-2t-2is_0}, q^{1+2t+2is_0}; q)_\infty}dt\\ &= Pv\int_{-\infty}^{\infty}\frac{(aq^{-t}, aq^t, bq^{-t}, bq^t, cq^{-t}, cq^t, dq^{-t}, dq^t, eq^{-t}, eq^t; q)_\infty}{(q^{1-t}/f, q^{1+t}/f, q^{1-2t}, q^{1+2t}; q)_\infty}dt,\end{aligned} \tag{11.12}$$

provided (11.2) holds and $|\arg f| > s_0 \log q^{-1} > 0$. The integral on the left side can now be calculated in exactly the same way as was done in §9, which corresponds to taking $s_0 = \pi/(2\log q^{-1})$ and replacing e, f by f and $-g$, respectively. Using the same arguments as in §9 it is easy to show that the integral on the left side of (11.12) is independent of s_0 except for the restriction that $0 < s_0 < |\arg f|/\log q^{-1}$. It is also easy to see that $\lambda_q = i\mu(a)$, where $\mu(a)$ is defined in (9.31) (with e, f replaced by $f, -g$). Formula (11.1) then follows by use of (11.12) via (9.26) and (9.29). The special case

$$\begin{aligned}Pv\int_{-\infty}^{\infty}&\frac{(aq^{-t}, aq^t, bq^{-t}, bq^t, cq^{-t}, cq^t, dq^{-t}, dq^t; q)_\infty}{(q^{1-2t}, q^{1+2t}; q)_\infty}\\ &= \frac{(q, ab/q, ac/q, ad/q, bc/q, bd/q, cd/q; q)_\infty}{(abcd/q^3; q)_\infty}, \qquad |abcd/q^3| < 1,\end{aligned} \tag{11.13}$$

then follows by taking $e = 0$, which is clearly a q-analogue of (1.12).

12. Summary of results. In this article we have dealt with only a small subset of the beta integrals. Readers who are interested in knowing all about them must go through some of Askey's papers as well as some of Ramanujan's. The point of view that we have taken here is that all of these formulas originate from the same source—the Pearson equation. They vary according to the underlying lattice types, and whether we are talking about a sum or an integral, along the real line or the imaginary axis, but most importantly it is the asymptotic properties of the summands and integrands near infinity that determine the final structure of the formulas. We have illustrated these ideas with a few examples in the previous sections but there are many more (old and new) formulas that can be proved by the same technique. For the benefit of the readers who would like to see the whole list of formulas in one place we are adding a few more pages to this paper.

We will classify the beta and q-beta integrals into four broad classes according to their lattice types: Table 1: $x(s) = s$ (linear); Table 2: $x(s) = C_1 s^2 + C_2$ (quadratic); Table 3: $x(s) = C_1 q^{-s} + C_2 q^s$ (q-quadratic); Table 4: $x(s) = q^{-s}$ (q-linear). The corresponding formulas will be subdivided into a polynomial-type formula, indicated by a suffix P, and a rational function-type formula, indicated by a suffix R, according to whether $\sigma(s)$ and/or $\tau(s)$ are polynomials or rational functions.

TABLE 1
Lattice $x(s) = s$.

Polynomial type	Rational function type
$\sigma(s) = (s - s_1)(s - s_2)$, $\sigma(s) + \tau(s) = (s + s_3)(s + s_4)$.	$\sigma(s) = (s - s_1)(s - s_2)$, $\sigma(s) + \tau(s) = \frac{(s + s_3)(s + s_4)(s + s_5)}{1 + s - s_6}$, $\sum_1^6 s_k = 1$.
(1_P) Barnes's first lemma (old),	(1_R) Barnes's second lemma (old),
(2_P) Gauss's ${}_2F_1$ summation formula (old),	(2_R) Nonterminating Saalschütz formula (old),
(3_P) Dougall's ${}_2H_2$ summation formula (old),	(3_R) Bilateral Saalschütz formula (old?),
(4_P) Ramanujan's integral (old).	(4_R) Extension of Ramanujan's integral (possibly new).

The corresponding formulas are

$$(1_P) \qquad \frac{1}{2\pi i}\int_C \Gamma(s_1 - s)\Gamma(s_2 - s)\Gamma(s_3 + s)\Gamma(s_4 + s)ds = \frac{\Gamma(s_1 + s_3)\Gamma(s_1 + s_4)\Gamma(s_2 + s_3)\Gamma(s_2 + s_4)}{\Gamma(s_1 + s_2 + s_3 + s_4)};$$

$$(1_R) \qquad \frac{1}{2\pi i}\int_C \frac{\Gamma(s_1 - s)\Gamma(s_2 - s)\Gamma(s_3 + s)\Gamma(s_4 + s)\Gamma(s_5 + s)}{\Gamma(s_1 + s_2 + s_3 + s_4 + s_5 + s)}ds = \frac{\prod_{k=3}^{5}\Gamma(s_1 + s_k)\Gamma(s_2 + s_k)}{\prod_{3\le j<k\le 5}\Gamma(s_1 + s_2 + s_j + s_k)}.$$

$$(2_P) \qquad {}_2F_1\left[\begin{matrix} s_1 + s_3, s_1 + s_4; \\ 1 + s_1 - s_2; \end{matrix} 1\right] = \frac{\Gamma(1 + s_1 - s_2)\Gamma(1 - s_1 - s_2 - s_3 - s_4)}{\Gamma(1 - s_2 - s_3)\Gamma(1 - s_2 - s_4)},$$

$$(2_R) \qquad {}_3F_2\left[\begin{matrix} s_1 + s_3, s_1 + s_4, s_1 + s_5; \\ 1 + s_1 - s_2, 2s_1 + s_2 + s_3 + s_4 + s_5; \end{matrix} 1\right]$$
$$= \frac{\Gamma(1 + s_1 - s_2)\Gamma(1 - s_1 - s_2 - s_3 - s_4)\Gamma(2s_1 + s_2 + s_3 + s_4 + s_5)}{\Gamma(1 - s_2 - s_3)\Gamma(1 - s_2 - s_4)\Gamma(s_1 + s_2 + s_3 + s_5)\Gamma(s_1 + s_2 + s_4 + s_5)}$$
$$- \frac{\Gamma(1 + s_1 - s_2)\Gamma(2s_1 + s_2 + s_3 + s_4 + s_5)}{\Gamma(s_1 + s_3)\Gamma(s_1 + s_4)\Gamma(s_1 + s_5)}$$
$$\cdot \frac{1}{(s_1 + s_5)(s_2 + s_5)}\,{}_3F_2\left[\begin{matrix} s_1 + s_2 + s_3 + s_5, s_1 + s_2 + s_4 + s_5, 1; \\ 1 + s_1 + s_5, 1 + s_2 + s_5; \end{matrix} 1\right],$$
$$\operatorname{Re}(s_1 + s_2 + s_3 + s_4) < 1.$$

(3_P)

$$\sum_{n=-\infty}^{\infty} \frac{1}{\Gamma(a_1+n)\Gamma(a_2+n)\Gamma(a_3-n)\Gamma(a_4-n)}$$
$$= \frac{\Gamma(a_1+a_2+a_3+a_4-3)}{\Gamma(a_1+a_3-1)\Gamma(a_1+a_4-1)\Gamma(a_2+a_3-1)\Gamma(a_2+a_4-1)}, \qquad \operatorname{Re}\left(\sum_1^4 a_k\right) > 3;$$

(3_R)

$$\sum_{n=-\infty}^{\infty} \frac{\Gamma(1-\alpha-a_6-n)}{\Gamma(a_1+\alpha+n)\Gamma(a_2+\alpha+n)\Gamma(a_3-\alpha-n)\Gamma(a_4-\alpha-n)\Gamma(a_5-\alpha-n)}$$
$$= \frac{\Gamma(2-a_6-a_3)\Gamma(2-a_6-a_4)\Gamma(2-a_6-a_5)}{\Gamma(a_1+a_3-1)\Gamma(a_1+a_4-1)\Gamma(a_1+a_5-1)\Gamma(a_2+a_3-1)\Gamma(a_2+a_4-1)\Gamma(a_2+a_5-1)}$$
$$+ \frac{\mu_2(\alpha)-\mu_1(\alpha)}{(2-a_6-a_3)(2-a_6-a_4)}\, {}_3F_2\left[\begin{matrix} a_1+a_5-1, a_2+a_5-1, 1; \\ 3-a_6-a_3, 3-a_6-a_4; \end{matrix} 1\right],$$

$$\mu_1(\alpha) = \pi^{-2}\sin\pi(a_1+\alpha)\sin\pi(a_2+\alpha),$$
$$\mu_2(\alpha) = \pi^{-2}\sin\pi(a_3-\alpha)\sin\pi(a_4-\alpha)\sin\pi(a_5-\alpha)/\sin\pi(a_6+\alpha),$$
$$\sum_{k=1}^{6} a_k = 5, \ \operatorname{Im}(\alpha+a_6) \neq 0, \ \operatorname{Re}(2-a_6) > \max[\operatorname{Re}(a_3,a_4,a_5)].$$

(4_P)

$$\int_{-\infty}^{\infty} \frac{ds}{\Gamma(a_1+s)\Gamma(a_2+s)\Gamma(a_3-s)\Gamma(a_4-s)}$$
$$= \frac{\Gamma(a_1+a_2+a_3+a_4-3)}{\Gamma(a_1+a_3-1)\Gamma(a_1+a_4-1)\Gamma(a_2+a_3-1)\Gamma(a_2+a_4-1)}, \qquad \operatorname{Re}\left(\sum_1^4 a_k\right) > 3;$$

(4_R)

$$\int_{-\infty}^{\infty} \frac{\Gamma(1-a_6-s)ds}{\Gamma(a_1+s)\Gamma(a_2+s)\Gamma(a_3-s)\Gamma(a_4-s)\Gamma(a_5-s)}$$
$$= \frac{\Gamma(2-a_6-a_3)\Gamma(2-a_6-a_4)\Gamma(2-a_6-a_5)}{\Gamma(a_1+a_3-1)\Gamma(a_1+a_4-1)\Gamma(a_1+a_5-1)\Gamma(a_2+a_3-1)\Gamma(a_2+a_4-1)\Gamma(a_2+a_5-1)}$$
$$+ \frac{\mu}{(2-a_6-a_3)(2-a_6-a_4)}\, {}_3F_2\left[\begin{matrix} a_1+a_5-1, a_2+a_5-1, 1 \\ 3-a_6-a_3, 3-a_6-a_4 \end{matrix} ; 1\right],$$

$$\sum_{k=1}^{6} a_k = 5, \quad \operatorname{Im}\ a_6 \neq 0, \quad \operatorname{Re}(2-a_6) > \max[\operatorname{Re}(a_3,a_4,a_5)], \quad \text{and}$$
$$\mu = \frac{1}{\pi^2}\int_0^1 \Big[\frac{\sin\pi(a_3-s)\sin\pi(a_4-s)\sin\pi(a_5-s)}{\sin\pi(a_6+s)} - \sin\pi(a_1+s)\sin\pi(a_2+s)\Big]ds.$$

The readers may notice a lack of symmetry on the right sides of (2_R), (3_R), and (4_R) as well as the fact that the ${}_3F_2$ series on the right side of (2_R) does not appear to be balanced (i.e., Saalschützian). This can be easily rectified by applying the standard transformation formulas for the ${}_3F_2$ series; see [14]. The purpose of leaving the formulas in this apparent asymmetric form is to emphasize the appearance of 1 as a numerator parameter in each of these cases, a feature that will keep repeating itself in the formulas to follow. The inquisitive reader will also observe the striking similarity between the formulas (3_P) and (4_P) as well as between (3_R) and (4_R), and may wonder if there are no integral analogues for the summation formulas (2_P) and (2_R). The answer is yes and we will report on those formulas in a subsequent paper.

TABLE 2
Lattice $x(s) = s^2$.

Polynomial type	Rational function type
$\sigma(s) = \prod_{k=1}^{4}(s - s_k)$.	$\sigma(s) = \prod_{k=1}^{5}(s - s_k)\Big/(s + s_6 - 1)$, $\sum_{k=1}^{6} s_k = 1$.
(5_P) de Branges–Wilson integral (old),	(5_R) Nassrallah–Rahman integral (old),
(6_P) Very well poised ${}_5F_4$ series (old),	(6_R) Nonterminating ${}_7F_6$ summation formula (old),
(7_P) Dougall's ${}_5H_5$ summation formula (old),	(7_R) Bilateral ${}_7H_7$ summation formula (old?),
(8_P) Askey's integral (new).	(8_R) Extension of Askey's integral (new).

$$(5_P)\qquad \frac{1}{2\pi i}\int_C \frac{\prod_{j=1}^{4}\Gamma(s_j+s)\Gamma(s_j-s)}{\Gamma(2s)\Gamma(-2s)}\,ds = \frac{\prod_{1\le j<k\le 4}\Gamma(s_j+s_k)}{\Gamma(s_1+s_2+s_3+s_4)};$$

$$(5_R)\qquad \frac{1}{2\pi i}\int_C \frac{\prod_{j=1}^{5}\Gamma(s_j+s)\Gamma(s_j-s)}{\Gamma(2s)\Gamma(-2s)\Gamma(1+s-s_6)\Gamma(1-s-s_6)}\,ds$$
$$= 2\frac{\prod_{1\le j<k\le 5}\Gamma(s_j+s_k)}{\prod_{j=1}^{5}\Gamma(1-s_6-s_j)}, \qquad \sum_{j=1}^{6} s_j = 1.$$

$$(6_P)$$
$${}_5F_4\left[\begin{matrix} 2s_1, 1+s_1, s_1+s_2, s_1+s_3, s_1+s_4; \\ s_1, 1+s_1-s_2, 1+s_1-s_3, 1+s_1-s_4; \end{matrix}\,1\right]$$
$$= \frac{\Gamma(1-s_1-s_2-s_3-s_4)}{\Gamma(1+2s_1)}\frac{\prod_{j=2}^{4}\Gamma(1+s_1-s_j)}{\prod_{2\le j<k\le 4}\Gamma(1-s_j-s_k)}, \qquad \mathrm{Re}(s_1+s_2+s_3+s_4) < 1;$$

$$(6_R)$$
$${}_7F_6\left[\begin{matrix} 2s_1, 1+s_1, s_1+s_2, s_1+s_3, s_1+s_4, s_1+s_5, s_1+s_6; \\ s_1, 1+s_1-s_2, 1+s_1-s_3, 1+s_1-s_4, 1+s_1-s_5, 1+s_1-s_6; \end{matrix}\,1\right]$$
$$= \frac{\prod_{j=2}^{5}\Gamma(1+s_1-s_j)\Gamma(s_j+s_6)}{\Gamma(2s_1+1)\Gamma(s_6-s_1)\prod_{1\le j<k\le 5}\Gamma(1-s_j-s_k)}$$

$$+\frac{\prod_{j=2}^{6}\Gamma(1+s_1-s_j)(1+s_6-s_1)}{\Gamma(2s_1+1)\prod_{j=2}^{6}\Gamma(s_1+s_j)(s_2+s_6)(s_3+s_6)(s_4+s_6)(s_5+s_6)}$$

$$\cdot\,{}_7F_6\left[\begin{matrix}1+s_6-s_1,\frac{3+s_6-s_1}{2},1-s_1-s_2,1-s_1-s_3,1-s_1-s_4,1-s_1-s_5,1\\ \frac{1+s_6-s_1}{2},1+s_2+s_6,1+s_3+s_6,1+s_4+s_6,1+s_5+s_6,1+s_6-s_1\end{matrix};1\right],$$

where

$$\sum_{j=1}^{6}s_j=1.$$

(7_P)

$$\sum_{n=-\infty}^{\infty}\frac{\alpha+n}{\prod_{j=1}^{4}\Gamma(a_j+\alpha+n)\Gamma(a_j-\alpha-n)}$$

$$=\frac{\sin 2\pi\alpha}{2\pi}\frac{\Gamma(a_1+a_2+a_3+a_4-3)}{\prod_{1\le j<k\le 4}\Gamma(a_j+a_k-1)},\qquad \mathrm{Re}\left(\sum_{1}^{4}a_j\right)>3;$$

(7_R)

$$\sum_{n=-\infty}^{\infty}\frac{(\alpha+n)\Gamma(1+\alpha-a_6+n)\Gamma(1-\alpha-a_6-n)}{\prod_{j=1}^{5}\Gamma(a_j+\alpha+n)\Gamma(a_j-\alpha-n)}$$

$$=\frac{\sin 2\pi\alpha}{2\pi}\frac{\prod_{j=1}^{5}(2-a_6-a_j)}{\prod_{1\le j<k\le 5}\Gamma(a_j+a_j-1)}+\frac{[\mu(\alpha)-\mu(-\alpha)](1+a_1-a_6)}{\prod_{j=2}^{5}(2-a_6-a_j)}$$

$$\cdot\,{}_7F_6\left[\begin{matrix}1+a_1-a_6,\frac{3+a_1-a_6}{2},a_1+a_2-1,a_1+a_3-1,a_1+a_4-1,a_1+a_5-1,1\\ \frac{1+a_1-a_6}{2},3-a_2-a_6,3-a_3-a_6,3-a_4-a_6,3-a_5-a_6,1+a_1-a_6\end{matrix};1\right],$$

with $\sum_{j=1}^{6}a_j=5$, $\mathrm{Re}(2-a_6)>\max[\mathrm{Re}(a_1,\dots,a_5)]$, $\mathrm{Im}\,(a_6+\alpha)\neq 0$, and $\mu(\alpha)=\prod_{j=1}^{5}\sin\pi(a_j-\alpha)/\pi^4\sin\pi(a_6+\alpha)$.

(8_P)

$$Pv\int_{-\infty}^{\infty}\frac{\Gamma(1-2s)\Gamma(1+2s)}{\prod_{j=1}^{4}\Gamma(a_j+s)\Gamma(a_j-s)}ds$$

$$=\frac{\Gamma(a_1+a_2+a_3+a_4-3)}{\prod_{1\le j<k\le 4}\Gamma(a_j+a_k-1)},\qquad \mathrm{Re}\left(\sum_{1}^{4}a_j\right)>3;$$

(8_R)

$$Pv\int_{-\infty}^{\infty}\frac{\Gamma(1-2s)\Gamma(1+2s)\Gamma(1-a_6-s)\Gamma(1-a_6+s)}{\prod_{j=1}^{5}\Gamma(a_j+s)\Gamma(a_j-s)}ds$$

$$=\frac{\prod_{j=1}^{5}\Gamma(2-a_6-a_j)}{\prod_{1\le j<k\le 5}\Gamma(a_j+a_k-1)}+\frac{\mu\,(1+a_1-a_6)}{\prod_{j=2}^{5}(2-a_6-a_j)}$$

$$\cdot\,{}_7F_6\left[\begin{matrix}1+a_1-a_6,\frac{3+a_1-a_6}{2},a_1+a_2-1,a_1+a_3-1,a_1+a_4-1,a_1+a_5-1,1\\ \frac{1+a_1-a_6}{2},3-a_6-a_2,3-a_6-a_3,3-a_6-a_4,\ 3-a_6-a_5,1+a_1-a_6\end{matrix};1\right],$$

with $\sum_{j=1}^{6} a_j = 5$, $\mathrm{Re}(2-a_6) > \max[\mathrm{Re}(a_1,\dots,a_5)]$, $\mathrm{Im}\, a_6 \neq 0$, and $\mu = \pi \int_0^1 \frac{\mu(s)-\mu(-s)}{\sin 2\pi s} ds$, $\mu(s)$ being the same as in 7_R.

TABLE 3
Lattice $x(s) = C_1 q^{-s} + C_2 q^{s}$.

Polynomial type	Rational function type
$\sigma(s) = q^{-2s}\prod_{j=1}^{4}(q^s - s_j)$.	$\sigma(s) = q^{-2s}\prod_{j=1}^{5}(q^s - s_j)/(q^s - q/s_6)$, $\nu^2 s_1 s_2 s_3 s_4 s_5 s_6 = q$, $C_2/C_1 = \nu$.
(9_P) The Askey–Wilson integral (old),	(9_R) The Nassrallah–Rahman integral (old),
(10_P) Very well poised ${}_6\phi_5$ sum (old),	(10_R) Nonterminating ${}_8\phi_7$ sum (old),
(11_P) Very well poised ${}_6\psi_6$ sum (old),	(11_R) Bilateral Jackson's formula (old?),
(12_P) The Askey integral (old).	(12_R) Extensions of Askey integral (new).

$$(9_P)\quad \int_0^{\pi} \frac{(e^{2i\theta}, e^{-2i\theta}; q)_\infty d\theta}{\prod_{k=1}^{4}(s_k e^{i\theta}, s_k e^{-i\theta}; q)_\infty} = \frac{2\pi (s_1 s_2 s_3 s_4; q)_\infty}{(q;q)_\infty \prod_{1\le j<k\le 4}(s_j s_k; q)_\infty},$$

$$|s_k| < 1, \quad k = 1,2,3,4;$$

$$(9_R)\quad \int_0^{\pi} \frac{(e^{2i\theta}, e^{-2i\theta}; q)_\infty (s_1 s_2 s_3 s_4 s_5 e^{i\theta}, s_1 s_2 s_3 s_4 s_5 e^{-i\theta}; q)_\infty}{\prod_{k=1}^{5}(s_k e^{i\theta}, s_k e^{-i\theta}; q)_\infty} d\theta$$

$$= \frac{2\pi \prod_{k=1}^{5}\left(\frac{s_1 s_2 s_3 s_4 s_5}{s_j}; q\right)_\infty}{(q;q)_\infty \prod_{1\le j<k\le 5}(s_j s_k; q)_\infty}, \quad |s_k| < 1, \quad k = 1,\dots,5.$$

$$(10_P)\quad {}_6\phi_5\left[\begin{matrix} s_1^2, qs_1, -qs_1, s_1 s_2, s_1 s_3, s_1 s_4 \\ s_1, -s_1, qs_1/s_2, qs_1/s_3, qs_1/s_4 \end{matrix}; q, \frac{q}{s_1 s_2 s_3 s_4}\right]$$

$$= \frac{(qs_1^2, q/s_2 s_3, q/s_2 s_4, q/s_3 s_4; q)_\infty}{(qs_1/s_2, qs_1/s_3, qs_1/s_4, q/s_1 s_2 s_3 s_4; q)_\infty}, \qquad \left|\frac{q}{s_1 s_2 s_3 s_4}\right| < 1;$$

$$(10_R)$$

$${}_8\phi_7\left[\begin{matrix} s_1^2, qs_1, -qs_1, s_1 s_2, s_1 s_3, s_1 s_4, s_1 s_5, s_1 s_6 \\ s_1, -s_1, qs_1/s_2, qs_1/s_3, qs_1/s_4, qs_1/s_5, qs_1/s_6 \end{matrix}; q, q\right]$$

$$= \frac{(qs_1^2, s_6/s_1, q/s_2 s_3, q/s_2 s_4, q/s_2 s_5, q/s_3 s_4, q/s_3 s_5, q/s_4 s_5; q)_\infty}{(qs_1/s_2, qs_1/s_3, qs_1/s_4, qs_1/s_5, s_2 s_6, s_3 s_6, s_4 s_6, s_5 s_6; q)_\infty}$$

$$+ \frac{(qs_1^2, s_1 s_2, s_1 s_3, s_1 s_4, s_1 s_5, s_1 s_6; q)_\infty}{(q, qs_1/s_2, qs_1/s_3, qs_1/s_4, qs_1/s_5, qs_1/s_6; q)_\infty} \frac{s_6(qs_6/s_1; q)_1}{s_1(s_2 s_6, s_3 s_6, s_4 s_6, s_5 s_6; q)_1}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} qs_6/s_1, q(qs_6/s_1)^{\frac{1}{2}}, -q(qs_6/s_1)^{\frac{1}{2}}, q/s_1 s_2, q/s_1 s_3, q/s_1 s_4, q/s_1 s_5, q \\ (qs_6/s_1)^{\frac{1}{2}}, -(qs_6/s_1)^{\frac{1}{2}}, qs_2 s_6, qs_3 s_6, qs_4 s_6, qs_5 s_6, qs_6/s_1 \end{matrix}; q, s_1 s_6\right]$$

with $s_1 s_2 s_3 s_4 s_5 s_6 = q$.

(11_P)

$$
{}_6\psi_6\left[\begin{matrix}\alpha q, -\alpha q, \alpha s_1, \alpha s_2, \alpha s_3, \alpha s_4 \\ \alpha, -\alpha, q\alpha/s_1, q\alpha/s_2, q\alpha/s_3, q\alpha/s_4\end{matrix}; q, \frac{q}{s_1 s_2 s_3 s_4}\right]
$$

$$
= \frac{(q, \alpha^2 q, q/\alpha^2, q/s_1 s_2, q/s_1 s_3, q/s_1 s_4, q/s_2 s_3, q/s_2 s_4, q/s_3 s_4; q)_\infty}{(q\alpha/s_1, q/\alpha s_1, q\alpha/s_2, q/\alpha s_2, q\alpha/s_3, q/\alpha s_3, q\alpha/s_4, q/\alpha s_4, q/s_1 s_2 s_3 s_4; q)_\infty},
$$

$$
|q/s_1 s_2 s_3 s_4| < 1;
$$

(11_R)

$$
{}_8\psi_8\left[\begin{matrix}\alpha q, -\alpha q, \alpha s_1, \alpha s_2, \alpha s_3, \alpha s_4, \alpha s_5, \alpha s_6 \\ \alpha, -\alpha, q\alpha/s_1, q\alpha/s_2, q\alpha/s_3, q\alpha/s_4, q\alpha/s_5, q\alpha/s_6\end{matrix}; q, q\right]
$$

$$
= \frac{(q, \alpha^2 q, q/\alpha^2, \alpha s_6, s_6/\alpha; q)_\infty}{(s_1 s_6, s_2 s_6, s_3 s_6, s_4 s_6, s_5 s_6; q)_\infty}
$$

$$
\cdot \frac{(q/s_1 s_2, q/s_1 s_3, q/s_1 s_4, q/s_1 s_5, q/s_2 s_3, q/s_2 s_4, q/s_2 s_5, q/s_3 s_4, q/s_3 s_5, q/s_4 s_5; q)_\infty}{(q\alpha/s_1, q/\alpha s_1, q\alpha/s_2, q/\alpha s_2, q\alpha/s_3, q/\alpha s_3, q\alpha/s_4, q/\alpha s_4, q\alpha/s_5, q/\alpha s_5; q)_\infty}
$$

$$
+ \frac{\mu(\alpha) - \mu(\alpha^{-1})}{\alpha - \alpha^{-1}} \frac{s_6(q s_6/s_1; q)_1}{(s_2 s_6, s_3 s_6, s_4 s_6, s_5 s_6; q)_1}
$$

$$
\cdot {}_8\phi_7\left[\begin{matrix}q s_6/s_1, q(q s_6/s_1)^{\frac{1}{2}}, -q(q s_6/s_1)^{\frac{1}{2}}, q/s_1 s_2, q/s_1 s_3, q/s_1 s_4, q/s_1 s_5, q \\ (q s_6/s_1)^{\frac{1}{2}}, -(q s_6/s_1)^{\frac{1}{2}}, q s_2 s_6, q s_3 s_6, q s_4 s_6, q s_5 s_6, q s_6/s_1\end{matrix}; q, s_1 s_6\right],
$$

where $s_1 s_2 s_3 s_4 s_5 s_6 = q$ and

$$
\mu(\alpha) = \alpha^2 \frac{(s_1/\alpha, s_2/\alpha, s_3/\alpha, s_4/\alpha, s_5/\alpha, s_6/\alpha; q)_\infty}{(q/\alpha s_1, q/\alpha s_2, q/\alpha s_3, q/\alpha s_4, q/\alpha s_5, q/\alpha s_6; q)_\infty}.
$$

(12_P)

$$
Pv \int_{-\infty}^{\infty} \frac{(aq^{-s}, aq^s, bq^{-s}, bq^s, cq^{-s}, cq^s, dq^{-s}, dq^s; q)_\infty}{(q^{1-2s}, q^{1+2s}; q)_\infty} ds
$$

$$
= \int_{-\infty}^{\infty} \frac{(iaq^t, -iaq^{-t}, ibq^t, -ibq^{-t}, icq^t, -icq^{-t}, idq^t, -idq^{-t}; q)_\infty}{(-q^{1+2t}, -q^{1-2t}; q)_\infty} dt
$$

$$
= \frac{(q, ab/q, ac/q, ad/q, bc/q, bd/q, cd/q; q)_\infty}{(abcd/q^3; q)_\infty}, \qquad |abcd/q^3| < 1;
$$

(12_R)

$$
Pv \int_{-\infty}^{\infty} \frac{(aq^{-s}, aq^s, bq^{-s}, bq^s, cq^{-s}, cq^s, dq^{-s}, dq^s, eq^{-s}, eq^s; q)_\infty}{(q^{1-2s}, q^{1+2s}, q^{1-s}/f, q^{1+s}/f; q)_\infty} ds
$$

$$= \int_{-\infty}^{\infty} \frac{(iaq^t, -iaq^{-t}, ibq^t, -ibq^{-t}, icq^t, -icq^{-t}, idq^t, -idq^{-t}, ieq^t, -ieq^{-t}; q)_\infty}{(-q^{1+2t}, -q^{1-2t}, iq^{1+t}/f, -iq^{1-t}/f; q)_\infty} dt$$

$$= \frac{(q, ab/q, ac/q, ad/q, ae/q, bc/q, bd/q, be/q, cd/q, ce/q, de/q; q)_\infty}{(q^2/af, q^2/bf, q^2/cf, q^2/df, q^2/ef; q)_\infty}$$

$$+ \nu \frac{q(aq/f; q)_1}{f(q^2/bf, q^2/cf, q^2/df, q^2/ef; q)_1}$$

$$\cdot {}_8\phi_7 \left[\begin{matrix} aq/f, q(aq/f)^{\frac{1}{2}}, -q(aq/f)^{\frac{1}{2}}, ab/q, ac/q, ad/q, ae/q, q \\ (aq/f)^{\frac{1}{2}}, -(aq/f)^{\frac{1}{2}}, q^3/bf, q^3/cf, q^3/df, q^3/ef, aq/f \end{matrix} ; q, q^2/af \right],$$

where $abcdef = q^5$ and

$$\nu = \int_0^1 \frac{\nu(s) - \nu(-s)}{(q^{2s}, q^{1-2s}; q)_\infty} q^s ds,$$

$$\nu(s) = q^{-2s} \frac{(aq^{-s}, q^{s+1}/a, bq^{-s}, q^{s+1}/b, cq^{-s}, q^{s+1}/c, dq^{-s}, q^{s+1}/d, eq^{-s}, q^{s+1}/e; q)_\infty}{(fq^s, q^{1-s}/f; q)_\infty}.$$

If $\mu(\alpha) = \mu(\alpha^{-1})$, then 11_R reduces to a formula due to Gosper [25, Ex. 5.12], but 11_R is pretty easy to prove by using the transformation theory of basic hypergeometric series.

TABLE 4
Lattice $x(s) = q^{-s}$.

Polynomial type	Rational function type
$\sigma(s) = s_3 s_4 (1 - s_1 q^{-s})(1 - s_2 q^{-s})$, $\sigma(s) + \tau(s) \bigtriangledown x_1(s) = (q^{-s} - s_3)(q^{-s} - s_4)$.	$\sigma(s) = (1 - s_1 q^{-s})(1 - s_2 q^{-s})/s_1 s_2$, $\sigma(s) + \tau(s) \bigtriangledown x_1(s) = q^{-2s} \frac{(1-s_3 q^s)(1-s_4 q^s)(1-s_5 q^s)}{1-q^{s+1}/s_6}$, $\prod_1^6 s_k = q$.
(13_P) Askey–Roy's q-analogue of Barnes's first lemma (old),	(13_R) An extension of Askey–Roy formula,
(14_P) The q-Gauss summation formula (old),	(14_R) Nonterminating Saalschütz formula, (old)
(15_P) The bilateral q-Gauss formula (old?),	(15_R) Bilateral q-Saalschütz formula, (new)
(16_P) The q-version of Ramanujan's integral(?).	(16_R) ?

$$(13_P) \qquad \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{(\alpha s_3 e^{-i\theta}, \frac{q}{\alpha s_3} e^{i\theta}, \alpha s_4 e^{i\theta}, \frac{q e^{-i\theta}}{\alpha s_4}; q)_\infty}{(s_1 e^{i\theta}, s_2 e^{i\theta}, s_3 e^{-i\theta}, s_4 e^{-i\theta}; q)_\infty} d\theta$$

$$= \frac{(\alpha, q/\alpha, \alpha s_3/s_4, q s_4/\alpha s_3, s_1 s_2 s_3 s_4; q)_\infty}{(q, s_1 s_3, s_1 s_4, s_2 s_3, s_2 s_4; q)_\infty}, \quad |s_j s_k| < 1;$$

see [8].

$$(13_R) \quad \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{(\alpha e^{-i\theta}/s_1, q s_1 e^{i\theta}/\alpha, \alpha s_2 e^{i\theta}, q e^{-i\theta}/\alpha s_2, q e^{-i\theta}/s_6; q)_\infty}{(s_1 e^{i\theta}, s_2 e^{i\theta}, s_3 e^{-i\theta}, s_4 e^{-i\theta}, s_5 e^{-i\theta}; q)_\infty} d\theta$$

$$= \frac{(\alpha, q/\alpha, \alpha s_2/s_1, q s_1/\alpha s_2, q/s_3 s_6, q/s_4 s_6, q/s_5 s_6; q)_\infty}{(q, s_1 s_3, s_1 s_4, s_1 s_5, s_2 s_3, s_2 s_4, s_2 s_5; q)_\infty}, \qquad |s_j s_k| < 1;$$

see [24].

$$(14_P)\quad {}_2\phi_1\left[\begin{matrix} s_1s_3, s_2s_3 \\ qs_3/s_4 \end{matrix}; q, q/s_1s_2s_3s_4\right] = \frac{(q/s_1s_4, q/s_2s_4; q)_\infty}{(qs_3/s_4, q/s_1s_2s_3s_4; q)_\infty},$$

$$|q/s_1s_2s_3s_4| < 1;$$

$$(14_R)\quad {}_3\phi_2\left[\begin{matrix} s_1s_4, s_2s_4, s_3s_4 \\ qs_4/s_5, qs_4/s_6 \end{matrix}; q, q\right]$$

$$= \frac{(q/s_1s_5, q/s_2s_5, q/s_3s_5, s_6/s_4; q)_\infty}{(s_1s_6, s_2s_6, s_3s_6, qs_4/s_5; q)_\infty}$$

$$+ \frac{(s_1s_4, s_2s_4, s_3s_4; q)_\infty}{(q, qs_4/s_5, qs_4/s_6; q)_\infty} \frac{s_6}{s_4(1-s_2s_6)(1-s_3s_6)}$$

$$\cdot {}_3\phi_2\left[\begin{matrix} q/s_1s_4, q/s_1s_5, q \\ qs_2s_6, qs_3s_6 \end{matrix}; q, s_1s_6\right].$$

$$(15_P)$$

$${}_2\psi_2\left[\begin{matrix} \alpha s_1, \alpha s_2 \\ \alpha q/s_3, \alpha q/s_4 \end{matrix}; q, q/s_1s_2s_3s_4\right]$$

$$= \frac{(\alpha^2, q/\alpha^2, s_3s_4, q/s_3s_4, q/s_1s_3, q/s_1s_4, q/s_2s_3, q/s_2s_4, q; q)_\infty}{(\alpha s_3, q/\alpha s_3, \alpha s_4, q/\alpha s_4, q/\alpha s_1, q/\alpha s_2, \alpha q/s_3, \alpha q/s_4, q/s_1s_2s_3s_4; q)_\infty}$$

$$+ \alpha^2 \frac{(\alpha q/s_1, \alpha q/s_2, s_3/\alpha, s_4/\alpha; q)_\infty}{(q/\alpha s_1, q/\alpha s_2, \alpha s_3, \alpha s_4; q)_\infty}\ {}_2\psi_2\left[\begin{matrix} s_1/\alpha, s_2/\alpha \\ q/\alpha s_3, q/\alpha s_4 \end{matrix}; q, q/s_1s_2s_3s_4\right],$$

$$|q/s_1s_2s_3s_4| < 1,$$

$$(15_R)$$

$${}_3\psi_3\left[\begin{matrix} \alpha s_1, \alpha s_2, \alpha s_3 \\ \alpha q/s_4, \alpha q/s_5, \alpha q/s_6 \end{matrix}; q, q\right]$$

$$- \alpha^2 \frac{(\alpha q/s_1, \alpha q/s_2, \alpha q/s_3, s_4/\alpha, s_5/\alpha, s_6/\alpha; q)_\infty}{(q/\alpha s_1, q/\alpha s_2, q/\alpha s_3, \alpha s_4, \alpha s_5, \alpha s_6; q)_\infty}\ {}_3\psi_3\left[\begin{matrix} s_1/\alpha, s_2/\alpha, s_3/\alpha \\ q/\alpha s_4, q/\alpha s_5, q/\alpha s_6 \end{matrix}; q, q\right]$$

$$= \frac{(\alpha^2, q/\alpha^2, s_4s_5, q/s_4s_5, q/s_1s_4, q/s_1s_5; q)_\infty}{(\alpha s_4, q/\alpha s_4, \alpha s_5, q/\alpha s_5, q/\alpha s_1, q/\alpha s_2; q)_\infty}$$

$$\cdot \frac{(q/s_2s_4, q/s_2s_5, q/s_3s_4, q/s_3s_5, s_6/\alpha, q; q)_\infty}{(q/\alpha s_3, \alpha q/s_4, \alpha q/s_5, s_1s_6, s_2s_6, s_3s_6; q)_\infty}$$

$$+\alpha^{-1}\frac{(\alpha s_1,\alpha s_2,\alpha s_3;q)_\infty}{(\alpha q/s_4,\alpha q/s_5,\alpha q/s_6;q)_\infty}$$
$$\cdot\left[1-\alpha^4\frac{(s_1/\alpha,q\alpha/s_1,s_2/\alpha,\alpha q/s_2,s_3/\alpha,\alpha q/s_3;q)_\infty}{(\alpha s_1,q/\alpha s_1,\alpha s_2,q/\alpha s_2,\alpha s_3,q/\alpha s_3;q)_\infty}\right.$$
$$\left.\cdot\frac{(s_4/\alpha,\alpha q/s_4,s_5/\alpha,\alpha q/s_5,s_6/\alpha,\alpha q/s_6;q)_\infty}{(\alpha s_4,q/\alpha s_4,\alpha s_5,q/\alpha s_5,\alpha s_6,q/\alpha s_6;q)_\infty}\right]$$
$$\cdot\frac{s_6}{(1-s_2s_6)(1-s_3s_6)}\,{}_3\phi_2\left[\begin{matrix}q/s_1s_4,q/s_1s_5,q\\ qs_2s_6,qs_3s_6\end{matrix};q,s_1s_6\right].$$

As before, the ${}_3\phi_2$ series in (14_R) and (15_R) can be transformed to balanced ${}_3\phi_2$ series so that (14_R) is essentially the well-known q-Saalschütz formula [25, eq. (2.10.12)]. However, (15_R) appears to be new, although we believe it can be derived from the general basic bilateral transformation formula [25, eq. (5.4.3)], as can the simpler formula (15_P).

It would be natural to expect that the q-linear case is much easier to handle than the q-quadratic case. One might be tempted to argue that the formulas for the q-linear case should be derived from those for the q-quadratic case by a limiting process. A comparison between the formulas 11_P and 15_P or between 11_R and 15_R should convince the reader that this is not necessarily so. In fact, the formulas for the q-linear case are much harder to derive than the q-quadratic case, especially for the bilateral series. There are two main reasons for this unexpected difficulty. First, the asymptotics for the summands and integrands at $\pm\infty$ are much more delicate when $x(s)=q^{-s}$. Second, the task of taking the limit $n\to\infty$ after n iterations is no longer as straightforward as is the case for the other lattices—instead of one series one may end up with as many as four series! This is what we believe may happen in (16_P) and (16_R) that we have left as open problems. We have some ideas about (16_P) but almost none for (16_R). This is probably the time we could use a helping hand from Ramanujan. A little nudge from Richard Askey, with his near-prophetic feeling for formulas, would also be helpful.

Acknowledgment. We would like to thank the referee for pointing out the important reference [28]. Many helpful comments from Professors R. Askey and M.E.H. Ismail are also appreciated.

REFERENCES

[1] R. P. AGARWAL, *On integral analogues of certain transformations of well-poised basic hypergeometric series*, Quart. J. Math. Oxford, 4 (1953), pp. 161–167.

[2] G. E. ANDREWS AND R. ASKEY, *Another q-extension of the beta functions*, Proc. Amer. Math. Soc., 81 (1981), pp. 97–100.

[3] R. ASKEY, *A q-extension of Cauchy's form of the beta integral*, Quart. J. Math. Oxford Ser. 2, 32 (1981), pp. 255–266.

[4] ———, *Beta integrals and the associated orthogonal polynomials*, in Number Theory, K. Alladi, ed., Lecture Notes in Math. 1395, Springer-Verlag, New York, 1989, pp. 84-121.

[5] ———, *Beta integrals in Ramanujan's Papers, His Unpublished Work and Further Examples, Ramanujan Revisited*, G. E. Andrews et al., eds., Academic Press, New York, 1988, pp. 561–590.

[6] ———, *Beta integrals and q-extension*, Papers of the Ramanujan Centennial International Conference, Ramanujan Mathematical Society, 1987, pp. 85–102.

[7] ———, *Continuous Hahn polynomials*, J. Phys. A., 19 (1985), pp. L1017–L1019.

[8] R. ASKEY AND R. ROY, *More q-beta integrals*, Rocky Mountain J. Math., 16 (1986), pp. 365–372.

[9] R. ASKEY AND J. A. WILSON, *Some basic hypergeometric polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985).

[10] N. M. ATAKISHIYEV AND S. K. SUSLOV, *The Hahn and Meixner polynomials of an imaginary argument and some of their applications*, J. Phys. A, 18 (1985), pp. 1583–1596.

[11] ———, *Difference hypergeometic functions, Approximation Theory*: U.S.A.-U.S.S.R. A. A. Gonchar and E. B. Saff, eds., Springer-Verlag, New York, 1992, pp. 1–35.

[12] N. M. ATAKISHIYEV, M. RAHMAN, AND S. K. SUSLOV, *A definition and a classification of the classical orthogonal polynomials*, to appear.

[13] N. M. ATAKISHIYEV AND S. K. SUSLOV, *On the Askey–Wilson polynomials*, Constr. Approx., 8 (1992), pp. 363–369.

[14] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, UK, 1935; reprinted by Stechert–Hafner, New York, 1964.

[15] E. W. BARNES, *A new development of the theory of hypergeometric functions*, Proc. London Math. Soc., 6 (1908), pp. 141–177.

[16] ———, *A transformation of generalized hypergeometric series*, Quart. J. Math. Oxford Ser., 41 (1910), pp. 136–140.

[17] G. BIRKHOFF AND G. C. ROTA, *Ordinary Differential Equations*, fourth ed., John Wiley, New York, 1989.

[18] L. DE BRANGES, *Tensor product spaces*, J. Math. Anal. Appl., 38 (1972), pp. 109–148.

[19] A.-L. CAUCHY, *Sur les intégrales définèes prises entre des limites imaginaires*, Bulletin de Ferussoc, T. III (1825), pp. 214–221; Oeuvres de Cauchy, 2^e série, T. II, Gauthier-Villars, Paris, 1958, pp. 57–65.

[20] ———, *Mémoire sur les fonctions dont plusieurs valeurs sont lieés entre elles par une équation linéaire, et sur diverses transformations de produits composés d'un nombre indéfini de facteurs*, C.R. Acad. Sci. Paris, T. XVII (1845), p. 523; Oeuvres de Cauchy, 1^{re} série, T. VIII, Gauthier-Villars, Paris, 1893, pp. 42–50.

[21] J. DOUGALL, *On Vandermonde's theorem and some more general expansions*, Proc. Edinburgh Math. Soc., 25 (1907), pp. 114–132.

[22] A. ERDÉLYI, ED., *Tables of Integral Transforms*, Bateman Manuscript Project, Vol. 2, McGraw-Hill, New York, 1954.

[23] G. GASPER, *Solutions to problem #6497 (q-Analogues of a gamma function identity, by R. Askey)*, Amer. Math. Monthly, 94 (1987), pp. 199–201.

[24] G. GASPER, *q-Extensions of Barnes', Cauchy's, and Euler's beta integrals*, in Topics in Mathematical Analysis, T. M. Rassias, ed., World Scientific Publishing Co., London, Singapore, and Teaneck, NJ, 1989, pp. 294–314.

[25] G. GASPER AND M. RAHMAN, *Basic Hypergeometric Series*, in Encyclopedia of Mathematics and Its Applications, Vol. 35, Cambridge University Press, Cambridge, UK, 1990.

[26] R. WM. GOSPER, Letters to R. Askey, March 27 and July 21, 1988.

[27] R. A. GUSTAFSON, *A generalization of Selberg's beta integral*, Bull. Amer. Math. Soc., 22 (1990), pp. 97–106.

[28] ———, *Some q-beta and Mellin–Barnes integrals on compact Lie groups and Lie algebras*, Trans. Amer. Math. Soc., 1993, to appear.

[29] G. H. HARDY *Ramanujan*, Cambridge University Press, Cambridge, UK, 1940; reprinted by Chelsea, New York, 1978.

[30] E. HEINE, *Untersuchungen über die Reiche ...*, J. Reine Angew. Math., 34 (1847), pp. 285–328.

[31] F. H. JACKSON, *A generalization of the functions $\Gamma(n)$ and x^n*, Proc. Roy. Soc. London, 74 (1904), pp. 64–72.

[32] ———, *On q-definite integrals*, Quart. J. Pure Appl. Math., 41 (1910), pp. 193–203.

[33] E. G. KALNINS AND W. MILLER, *Symmetry techniques for q-series: Askey–Wilson polynomials*, Rocky Mountain J. Math., 19 (1989), pp. 1–8.

[34] ———, *q-series and orthogonal polynomials associated with Barnes' first lemma*, SIAM J. Math. Anal., 19 (1988), pp. 1216–1231.

[35] D. E. KNUTH, *The Art of Computer Programming*, Vol. I, Addison-Wesley, Reading, MA, 1973.

[36] T. H. KOORNWINDER, *Special functions: q-analogues and interpretations on groups and quantum groups*, Lecture Notes, CWI, Amsterdam, 1991.

[37] W. MILLER, *A note on Wilson polynomials*, SIAM J. Math. Anal., 18 (1987), pp. 1221–1226.

[38] B. NASSRALLAH AND M. RAHMAN, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, SIAM J. Math. Anal., 16 (1985), pp. 186–197.

[39] A. F. NIKIFOROV, S. K. SUSLOV, AND V. B. UVAROV, *Racah polynomials and dual Hahn polynomials as a generalization of classical orthogonal polynomials of a discrete variable*, preprint 165, Keldysh Institute Appl. Math., Moscow, 1982. (In Russian.)

[40] ———, *Classical orthogonal polynomials of a discrete variable*, Nauka, Moscow, 1985. (In Russian.) English translation, Springer-Verlag, New York, 1991.

[41] A. F. NIKIFOROV AND S. K. SUSLOV, *Systems of classical orthogonal polynomials of a discrete variable on nonuniform lattices*, preprint 8, Keldysh Institute Appl. Math., Moscow, 1985. (In Russian.)

[42] A. F. NIKIFOROV AND V. B. UVAROV, *Classical orthogonal polynomials of a discrete variable on nonuniform lattices*, preprint 17, Keldysh Institute Appl. Math., Moscow, 1983, (In Russian.)

[43] ———, *Fundamentals of the theory of classical orthogonal polynomials of a discrete variable*, preprint 56, Keldysh Institute Appl. Math., Moscow, 1987. (In Russian.)

[44] ———, *Construction of q-analogues of classical orthogonal polynomials of a discrete variable on nonuniform lattices*, preprint 179, Keldysh Institute Appl. Math., Moscow, 1987. (In Russian.)

[45] T. J. OSLER, *A further extension of the Leibnitz rule to fractional derivatives and its relation to Parseval's formula*, SIAM J. Math. Anal., 3 (1972), pp. 1–16.

[46] M. RAHMAN, *Families of biorthogonal rational functions in a discrete variable*, SIAM J. Math. Anal., 12 (1981), pp. 355–367.

[47] ———, *The linearization of the product of continuous q-Jacobi polynomials*, Canad. J. Math., 33 (1981), pp. 255–284.

[48] ———, *An integral representation of a ${}_{10}\Phi_9$ and continuous bi-orthogonal rational functions*, Canad. J. Math., 38 (1986), pp. 605–618.

[49] M. RAHMAN AND S. K. SUSLOV, *Classical bi-orthogonal rational functions*, in Methods in Approximation Theory in Complex Analysis and Mathematical Physics, A. A. Gonchar and E. B. Saff, eds., Lecture Notes in Mathematics, Vol. 1550, Springer-Verlag, Berlin, 1993, pp. 131–166.

[50] S. RAMANUJAN, *A class of definite integrals*, Quart. J. Math., 48 (1920), pp. 294–310; reprinted in Collected Papers of Srinivasa Ramanujan, G. H. Hardy, P. V. Seshu Aiyar, and B. M. Wilson, eds., Cambridge University Press, Cambridge, 1927; reprinted by Chelsea, New York, 1962.

[51] A. A. SAMARSKII, *Teoriya Raznostnykh Skhem* (*Theory of Difference Schemes*), Nauka, Moscow, 1977. (In Russian.)

[52] L. J. SLATER, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, UK, 1966.

[53] S. K. SUSLOV, *The theory of difference analogues of special functions of hypergeometric type*, Russian Math. Surveys, 44 (1989), pp. 227–278. (In English.)

[54] G. SZEGÖ, *Orthogonal Polynomials*, fourth ed., Amer. Math. Soc. Colloq. Publ. 23, Providence, RI, 1975.

[55] J. THOMAE, *Beiträge zur Theorie der durch die Heinesche Reiche ...*, J. Reine Angew. Math., 70 (1869), pp. 258–281.

[56] ———, *Les séries Heinéennes supérieres, ou les séries de la forme*, Ann. Mat. Pure Appl., 4 (1870), pp. 105–138.

[57] G. N. WATSON, *The continuation of functions defined by generalized hypergeometric series*, Trans. Cambridge Philos. Soc., 21 (1910), pp. 281–299.

[58] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, fourth ed., Cambridge University Press, Cambridge, UK, 1965.

[59] J. A. WILSON, *Some hypergeometric orthogonal polynomials*, SIAM J. Math. Anal., 11 (1980), pp. 690–701.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 694–711, March 1994

028

THE CLASSICAL UMBRAL CALCULUS*

G.-C. ROTA†‡ AND B. D. TAYLOR†§

Abstract. A rigorous presentation of the umbral calculus, as formerly applied heuristically by Blissard, Bell, Riordan, and others is given. As an application, the basic identities for Bernoulli numbers, as well as their generalizations first developed by Nörlund are derived.

Key words. umbral calculus, Bernoulli numbers, special functions

AMS subject classifications. 05, 33

1. Introduction. It is seldom recognized that the algebraic notation that we use today in mathematics is not the only possible one, nor perhaps even the best. The history of mathematics is littered with abandoned notation, some of which did not deserve such neglect. In this paper we consider one such notation, one that somehow has managed to survive in a no man's land and that has refused to die altogether because of its usefulness. In computing with sequences of numbers a_n, indexed by the nonnegative integers $n = 0, 1, 2, \ldots$, it is often convenient to treat the subscripts as if they were powers. This requires a_0 to equal 1 but is not otherwise a great strain. It was realized early in the development of combinatorics that this simple device would yield insight into quite a number of formulas and, what is more, that it would often suggest the right proofs.

To the best of our knowledge, the first mathematicians who extensively used such a device were Edouard Lucas, in the first volume of his *Théorie des nombres* [L] (the second volume never appeared), and the Rev. John Blissard, in a series of nine papers that appeared in the *Quarterly Journal of Mathematics* between 1862 and 1868 [Bl1]–[Bl9]. The editors of the *Quarterly Journal* must be given credit for publishing papers that thoroughly lacked in rigor but relied on a suggestive and often powerful notation.

Since that time, the umbral calculus, as it came to be called, has not been accepted in mathematical society. Sylvester held the umbral calculus in high esteem but made no attempt to present it; he simply used it. Only Eric Temple Bell [Be1] among mathematicians early in this century had the daring to show public appreciation for this notation. In a book and several papers he attempted to display the full power of the method and at the same time to give it a presentation that would meet the standards of algebraic rigor of the twentieth century. The last of his papers, written in 1940 [Be2], purports to give an explicit axiomatic basis of the umbral calculus. As happens with some writings of E. T. Bell, it is not quite clear whether these axioms make any sense, much less whether they are rigorous at all. Nonetheless, umbral calculus has managed to eke out a meager existence. John Riordan used it extensively in *An Introduction to Combinatorial Analysis* [Ri], the first modern textbook of combinatorics, and although he gives no justification of the method, every formula he displays is correct, though at times his proofs have an air of witchcraft about them. As a matter of fact, the feeling of witchcraft that has hovered over umbral calculus is probably what has kept it from dying out altogether.

* Received by the editors March 16, 1993; accepted for publication April 20, 1993.
† Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.
‡ This author was supported by National Science Foundation grant MCS-8104855.
§ This author was supported by an Office of Naval Research National Defense Science and Engineering Graduate Fellowship.

To the best of our knowledge, it was one of the present authors [Ro] who first pointed out explicitly what nowadays had become completely obvious, namely, that the shift from a^n to a_n can be "explained" as the application of a linear functional on the space of polynomials in the variable a. Unfortunately, it took the same author another twenty years to realize that this simple truth could not alone explain the stunts of calculations performed by the umbral mathematicians. At long last, it was realized that umbral calculus could be made entirely rigorous by using the language of Hopf algebras, and this was done in a lengthy treatment [RR]. However, although the notation of Hopf algebra satisfied the most ardent advocate of spic-and-span rigor, the translation of "classical" umbral calculus into the newly found rigorous language made the method altogether unwieldy and unmanageable. Not only was the eerie feeling of witchcraft lost in the translation, but, after such a translation, the use of calculus to simplify computation and sharpen our intuition was lost by the wayside. In retrospect, what the authors should have done is to translate the language of Hopf algebra into classical umbral language, a feat that was logically impossible at the time since it was not known whether classical umbral language made any sense.

In the present paper we have finally decided to bite the bullet and give a rigorous, simple presentation of umbral calculus as it was meant by the founders. We have kept new notation both minimal and indispensable to avoid the misunderstandings of the past. Our basic construct is a polynomial ring in several variables, on which no less than three equivalence relations are defined (the "classics" blithely denoted all three of these equivalence relations by the same symbol $=$, not realizing that this was the beginning of the end). The peculiarity of two of these three equivalence relations is that they are not invariant under substitution of variables, nor under multiplication. In fact, one of them is not even invariant under addition. More precisely, they are invariant under addition or multiplication provided that the variables involved are distinct, a fact that should make some algebraists slightly uncomfortable. Moreover, relative to one of these equivalence relations, the equivalence classes form something like an Abelian group, where, however, every element has an infinite set of inverses, all "equivalent" to each other, an unheard-of phenomenon in algebra. In fact, if one extracts a representative from each equivalence class and is willing to restrict oneself to sums of distinct representatives, then the resulting group is easily seen to be isomorphic to the group of formal power series in a single variable with constant coefficient 1, under multiplication. Weirdest of all, there are two ways of multiplying an element of the ring by a "scalar" integer, one of which is defined in a way that defies the current tenets of algebraic sobriety.

But whoever is willing to swallow this strange but (at last) correct definition, and to realize, after a little training, that the calculus is quite manageable, may find it rewarding to start performing combinatorial computations using umbral calculus without fear. Or else he or she may wish to reread the classics with the present key in hand and to realize that these writings make complete sense in retrospect.

We give only one example of computation with umbral calculus: by developing the theory of Bernoulli numbers in umbral notation. All classical identities relating to these numbers are found to have one-line proofs, which, we would like to believe, are the natural ones.

We hope to present other simplifications by the use of umbral calculus in forthcoming publications.

2. Foundations of umbral calculus. An umbral calculus consists of the following data:

(1) A set A, called the *alphabet*, whose elements are called letters or *umbrae*;

(2) A commutative integral domain D, whose quotient field is of characteristic zero;

(3) A linear functional **eval**, called the *evaluation*, defined on the polynomial ring $D[A]$ and taking values in D, such that

(a) $\mathbf{eval}(1) = 1$, where 1 is the identity of D;

(b) if $\alpha, \beta, \ldots, \gamma$ are *distinct* umbrae and if $i, j, \ldots, k$ are nonnegative integers, then

$$\mathbf{eval}(\alpha^i \beta^j \cdots \gamma^k) = \mathbf{eval}(\alpha^i)\mathbf{eval}(\beta^j) \cdots \mathbf{eval}(\gamma^k);$$

(4) A distinguished element ε of the alphabet A, called the *augmentation*, such that

$$\mathbf{eval}(\varepsilon^n) = \delta_{n,0},$$

where δ is the Kronecker delta.

Elements of the polynomial ring $D[A]$ are called *umbral polynomials*. When an umbral polynomial p is written as a linear combination of distinct monomials with nonzero coefficients, the *support* of p is defined to be the set of all umbrae that occur in some such monomial with a positive power. If $\{\alpha, \beta, \ldots, \gamma\}$ is the support of an umbral polynomial p, we may write p as $p(\alpha, \beta, \ldots, \gamma)$.

A set of umbral polynomials such that the supports of any two of them are disjoint is said to be *unrelated*.

A sequence $a_0, a_1, a_2, \ldots$, denoted (a_i), is said to be *umbrally represented* by an umbra α when

$$\mathbf{eval}(\alpha^n) = a_n$$

for $n = 0, 1, 2, \ldots$. Note that if the sequence (a_n) is umbrally represented by an umbra α, then we necessarily have $a_0 = 1$.

If p and q are umbral polynomials, we shall say that p and q are *umbrally equivalent*, in symbols $p \simeq q$, when $\mathbf{eval}(p) = \mathbf{eval}(q)$. For example, if the umbra α represents the sequence (a_n), then we have $\alpha^n \simeq a_n$ for all n, as well as

$$(\alpha + 1)^n \simeq \sum_{i \geq 0} \binom{n}{i} a_i.$$

Two umbral polynomials p and q are said to be *exchangeable*, in symbols $p \equiv q$, when $\mathbf{eval}(p^n) = \mathbf{eval}(q^n)$ for $n = 0, 1, 2, \ldots$. Two equal polynomials are exchangeable, and two exchangeable polynomials are umbrally equivalent, but neither converse holds.

Two umbrae α and β are said to be *inverse* to each other when $\alpha + \beta \equiv \varepsilon$; we say that α is an inverse of β. The inverse of an umbra is not unique, but any two inverse umbrae to the umbra α are exchangeable.

We note that ε behaves much like 0, namely, if p is any umbral polynomial, then $\varepsilon + p \equiv p$ and $\varepsilon \cdot p \equiv \varepsilon$.

By judicious use of the equivalence relations of umbral equivalence and exchangeability, one can dispense with explicit mention of the linear functional **eval**, as we shall do whenever possible.

Suppose the umbral polynomials p and q are exchangeable with the umbral polynomials f and g, respectively. It is true that $p + q$ is umbrally equivalent to $f + g$ but

not that pq is umbrally equivalent to fg. Neither such property holds for exchangeability. Let α and β be exchangeable umbrae representing the sequence a_n; then $(\alpha+\alpha)^3 \simeq 8a_3$ but $(\alpha+\beta)^3 \simeq 2a_3+6a_1a_2$. In other words, umbral polynomials with respect to the equivalence relations of umbral equivalence and exchangeability are not invariant under substitution of exchangeable variables. This failure of the substitution property, which is usually taken for granted in dealing with ordinary equality, has in the past been the stumbling block to a rigorous presentation of umbral calculus.

We have, however, the following proposition, which is easily verified.

PROPOSITION 2.1. *Let $p(x,y,\dots,z)$ be a polynomial in the variables $x,y,\dots,z$ with coefficients in D, and let $f,g,\dots,h,r$ be umbral polynomials, where f and r are exchangeable. If f is unrelated to each of $g,\dots,h$ and if r is also unrelated to each of $g,\dots,h$, then the polynomials $p(f,g,\dots,h)$ and $p(r,g,\dots,h)$ are exchangeable.*

In particular, we have the following corollary.

COROLLARY 2.2. *If the umbral polynomial f is exchangeable with an umbral polynomial g, if the umbral polynomial r is exchangeable with the umbral polynomial s, if f and r are unrelated, and if g and s are unrelated, then $f+r$ is exchangeable with $g+s$ and fr is exchangeable with gs.*

3. Saturated umbral calculi. We come now to one of the problems that vitiated the validity of umbral notation in the past. Briefly, one would like to write a sum of the form

$$\alpha' + \alpha'' + \cdots + \alpha''',$$

(where $\alpha', \alpha'', \dots, \alpha'''$ are any set of n distinct umbrae, each of them exchangeable with a given umbra α) in the form of the integer n multiplied by α, in some sense or other. However, the element $n\alpha \in D[A]$ is in no way exchangeable with $\alpha' + \alpha'' + \cdots + \alpha'''$. For example, if $\mathbf{eval}(\alpha^k) = a_k$, we have

$$\mathbf{eval}\Big((2\alpha)^k\Big) = 2^k\mathbf{eval}(\alpha^k) = 2^k a_k,$$

whereas

$$\mathbf{eval}\Big((\alpha'+\alpha'')^k\Big) = \sum_{i\geq 0}\binom{k}{i}(\alpha')^i(\alpha'')^{k-i} \simeq \sum_{i\geq 0}\binom{k}{i}a_i a_{k-i}.$$

Thus $\alpha'+\alpha''$, though umbrally equivalent to the expression 2α, is neither equal to nor exchangeable with 2α for general α. This puzzling phenomenon was, to the best of our knowledge, first explicitly pointed out by E. T. Bell, who was not able to provide a consistent notation.

What is needed, we believe, is a new notion of multiplication of an umbra by an integer, accompanied by the systematic use of the three kinds of equivalence relations introduced in the preceeding. We shall shortly give a rigorous definition of this concept of multiplication, but a few words of motivation may be in order.

We shall denote by the symbol $n\,.\,\alpha$ an umbra, often called an *auxiliary umbra*, that is exchangeable with the sum

$$\alpha' + \alpha'' + \cdots + \alpha''',$$

where $\alpha', \alpha'', \dots, \alpha'''$ are a set of n distinct umbrae, each of which is exchangeable with the given umbra α. Similarly, for every umbra α and for every positive integer n,

we introduce an umbra, written as $-n.\alpha$, that is exchangeable with $\beta'+\beta''+\cdots+\beta'''$, where $\beta', \beta'', \ldots, \beta'''$ is any set of n distinct umbrae exchangeable with β, and where β and α are inverse umbrae. Finally, we shall denote by $0 \,.\, \alpha$ an umbra exchangeable with the augmentation ε.

A similar notation $n \,.\, p$ is introduced for any umbral polynomial p whose support does not contain auxiliary umbrae. However, we have chosen not to allow such expressions as $m \,.\, (n \,.\, \alpha)$.

When all is said and done, it turns out that the "scalar multiplication," $n \,.\, \alpha$, satisfies the expected properties, stated in Proposition 3.2, which follows.

This motivation, together with the requirement that the alphabet shall be endowed with sufficiently many umbrae exchangeable with any expression whatsoever, leads us to the following definition.

DEFINITION 3.1. *A* saturated umbral calculus *with base alphabet A is an umbral calculus on an alphabet $A \cup B$, where the letters of the alphabet B (the auxiliary alphabet) are denoted by $n.p$ as p ranges over all the polynomials in $D[A]$ and n ranges over all integers. Further, a saturated umbral calculus shall satisfy the following:*

(1) *For every umbral polynomial q in $D[A \cup B]$, there exists an infinite set of umbrae α in A exchangeable with q.*

(2) *For every umbra α in A and for every positive integer n, the umbra $n.\alpha$ in B is exchangeable with $\alpha' + \alpha'' + \cdots + \alpha'''$, where $\alpha', \alpha'', \ldots, \alpha'''$ is any set of n distinct umbrae in A, each of them exchangeable with α.*

(3) *For every umbra α in A and for every positive integer n, the umbra $-n \,.\, \alpha$ in B is exchangeable with $\beta' + \beta'' + \cdots + \beta'''$, where $\beta', \beta'', \ldots, \beta'''$ is any set of n distinct umbrae in A, each of them inverse to α.*

(4) *For every umbra α in A, the umbra $0 \,.\, \alpha$ is exchangeable with the augmentation ε.*

(5) *For every umbral polynomial p in $D[A]$, for every integer n, and for every umbra α in A exchangeable with p, the umbra $n.p$ in B is exchangeable with the umbra $n \,.\, \alpha$.*

We shall hereafter assume that the umbral calculus we are dealing with is saturated. It can be shown that saturated umbral calculi exist and that every umbral calculus can be embedded in a saturated umbral calculus, but we shall spare the reader such proof.

In dealing with a saturated umbral calculus, the term "umbra" will denote an element of the alphabet A and not an element of the auxiliary alphabet B, unless otherwise specified.

The following statements are easily proved.

PROPOSITION 3.2.

(1) $(n+m).\alpha \equiv n.\alpha' + m.\alpha''$ *for any two integers n and m and any two distinct umbra α' and α'' exchangeable with α.*

(2) *If $n \,.\, \alpha \equiv n \,.\, \beta$ for some integer $n \neq 0$, then $\alpha \equiv \beta$.*

(3) *If c is any element of D, then $n \,.\, (c\alpha) \equiv c(n \,.\, \alpha)$.*

(4) *If c is any element of D, then $n \,.\, (c) \equiv nc$ for any integer n.*

The following proposition lists two other useful identities applicable in a saturated umbral calculus.

PROPOSITION 3.3.

(1) *If α and β are inverse umbrae, then $-\beta p(\alpha + \beta) \simeq \alpha p(\alpha + \beta)$ for any polynomial p in one variable with coefficients in D.*

(2) *If α and β are inverse umbrae and if p is a polynomial in one variable over*

D, then for all $n \in \mathbb{Z}$,

$$(n \,.\, \beta)p(n \,.\, \alpha + n \,.\, \beta) \simeq n\Big(\beta p(\alpha + \beta)\Big).$$

Proof (sketch). The left-hand side of the equation in the first part of the proposition may be subtracted from both sides of the identity, preserving umbral equivalence. Then it may be seen that the resulting right-hand side is actually *equal* to a polynomial in $\alpha + \beta$ with zero constant term. But this is equivalent to ε, thus proving the statement.

The key point to proving the second part involves observing that $n \,.\, \beta$ is by definition exchangeable with a sum of n distinct umbrae each exchangeable with β and that to effect a substitution *both* occurrences of $n \,.\, \beta$ in $(n \,.\, \beta)p(n \,.\, \alpha + n \,.\, \beta)$ must be replaced with this sum. Distributing yields a sum of n exchangeable terms, a representative one being $\beta' p\big((\beta' + \beta'' + \cdots + \beta''') + n \,.\, \alpha\big)$, the inner summation containing n umbrae exchangeable with β. But this is exchangeable with $\beta' p(\beta' + \alpha)$. The result follows for n greater than 0. For n negative, the result is an easy consequence of the positive version and the first part of the proposition. □

Umbral calculus leads naturally to certain sequences of polynomials. A sequence $p_k(x)$ of polynomials is termed an *Appell sequence* if for all k, $p_k(x) \simeq (x + \alpha)^k$ for some umbra α.

It can be shown that any sequence satisfying both the relation $p_k'(x) = kp_{k-1}(x)$ and $p_0(x) = 1$ may be defined as $(x + \alpha)^k$ for some umbra α. In fact, the following proposition characterizes Appell sequences.

PROPOSITION 3.4. *Suppose* $p_0(x), p_1(x), p_2(x), \ldots$ *is a sequence of polynomials with* $p_k(x)$ *being of degree* k *and* $p_0(x) = 1$. *Then* $p_k(x) \simeq (x + \alpha)^k$ *for some* α *if and only if* $p_k'(x) = kp_{k-1}(x)$.

Proof. The following calculation proves the proposition in the *only if* direction:

$$Dp_k(x) \simeq D(x + \alpha)^k = k(x + \alpha)^{k-1} \simeq kp_{k-1}(x).$$

The converse is omitted. □

We may define a linear operator T on polynomials in the variable x, over the integral domain of umbral polynomials, by setting $T(x^k) = (x + \alpha)^k$. One verifies that $T = e^{\alpha D}$. Thus $e^{\alpha D}(x^k) = (x + \alpha)^k$. Furthermore, one easily verifies that if $F(t)$ is the exponential generating function for the sequence (a_i), then $F(D)x^n = p_n(x)$.

For the remainder of the present exposition, we shall assume that we are dealing with a saturated umbral calculus. The quotient field of the integral domain D will always contain the complex numbers as a subfield.

4. The Bernoulli umbra. A useful heuristic principle, widely used since the eighteenth century, views Bernoulli numbers as a device for transforming differences into derivatives. We shall now give a precise meaning to such a principle. An umbra β will be said to be a *Bernoulli umbra* if for every polynomial $p(x)$ we have

$$\Delta p(\beta) = p(\beta + 1) - p(\beta) \simeq p'(\varepsilon). \tag{1}$$

Here the operator Δ is the forward difference operator, defined as

$$\Delta f(x) = f(x + 1) - f(x).$$

PROPOSITION 4.1. *A Bernoulli umbra exists and is unique up to exchangeability.*

Indeed, setting $p(x) = x^n$, we obtain the recurrence

$$(\beta + 1)^n \simeq \beta^n, \qquad n > 1,$$

and $\beta^0 \simeq 1$.

Thus the umbra β umbrally represents a unique sequence of complex numbers. Such a sequence (B_n) is the sequence of *Bernoulli numbers.* In other words, for $n > 1$ we have

$$\sum_{k\geq 0} \binom{n}{k} B_k = B_n \quad \text{or} \quad \sum_{k=0}^{n-1} \binom{n}{k} B_k = 0.$$

These results are trivially valid for $n = 0$.

Let $f(x, \beta)$ be a formal power series in the variable x, whose coefficients are umbral polynomials in the Bernoulli umbra β (and possibly other umbrae as well). By comparing coefficients we see that (1) generalizes to

$$f(x, \beta + 1) - f(x, \beta) \simeq f'(x, \varepsilon), \tag{2}$$

where the derivative is taken relative to the second variable. Here, the symbol $\simeq$ is to be understood as coefficientwise equivalence for the coefficients of two formal power series in x.

Setting $f(x, \beta) = e^{\beta x}$, the latter being, of course, interpreted as a formal power series, we obtain

$$(e^x - 1)e^{\beta x} \simeq xe^{\varepsilon x} \simeq x,$$

or, equivalently,

$$e^{\beta x} \simeq \frac{x}{e^x - 1},$$

which gives the exponential generating function of the Bernoulli numbers.

We come now to the main property of the Bernoulli umbra, which is expressed in the following theorem.

THEOREM 4.2. *The Bernoulli umbra satisfies the following umbral equation:*

$$\beta + 1 \equiv -\beta.$$

Proof. In (1) set

$$p(x) = \frac{-(-x + \beta' + 1)^{n+1}}{n+1},$$

where β' is an umbra exchangeable with β, thereby obtaining $p(\beta + 1) - p(\beta) \simeq (\beta' + 1)^n$. The left-hand side of the preceding may be rewritten as

$$\frac{-(-\beta + \beta')^{n+1} + (-\beta + \beta' + 1)^{n+1}}{n+1}. \tag{3}$$

Now let $q(x)$ be the polynomial

$$q(x) = \frac{(x - \beta)^{n+1}}{n+1}$$

and apply (1) again with $q(x)$ in place of $p(x)$ and β' in place of β, thereby obtaining $q(\beta'+1)-q(\beta') \simeq (-\beta')^n$. But observe that the left side of the previous equation is identical to (3). We infer that

$$(\beta'+1)^n \simeq (-\beta')^n$$

for all $n \in \mathbb{N}$; hence the conclusion follows after we replace β' by the exchangeable umbra β. □

It must be remarked that the Bernoulli umbra is not the only umbra α satisfying the umbral equation $\alpha+1 \equiv -\alpha$.

From Theorem 4.2 we infer that for $n>1$, $\beta^n \simeq (\beta+1)^n \simeq (-1)^n\beta^n$ and hence $B_{2n+1}=0$ for $n \geq 1$.

Recall that the operator J, sometimes called the *Bernoulli operator*, is defined as

$$Jp(x) = \int_x^{x+1} p(t)dt.$$

We have now the following proposition.

PROPOSITION 4.3. *Equation* (1) *can be equivalently rewritten in either of the forms*

$$Jp(\beta) \simeq p(\varepsilon) \quad \textit{or} \quad Jp(\varepsilon) \simeq p(\gamma),$$

where γ *is an inverse umbra of* β.

Proof. Indeed, suppose $p(x) = Dq(x)$ for some polynomial $q(x)$, where D is the ordinary derivative. Since $JD = \Delta$, we have $Jp(x) = \Delta q(x)$ and thus

$$Jp(\beta) = \Delta q(\beta) \simeq Dq(\varepsilon) = p(\varepsilon).$$

But every polynomial $p(x)$ can be written in the form $Dq(x)$ for some polynomial $q(x)$; hence the first conclusion. Applying D to both sides (by recalling that J and D commute) yields the converse. To show that the second equation is equivalent to the first, set $p(x) = q(\gamma+x)$, where $q(x)$ is any univariate polynomial. Applying the result just proved to this choice of $p(x)$, we obtain

$$Jq(\gamma+\beta) \simeq q(\gamma+\varepsilon) \simeq q(\gamma).$$

But $Jq(\gamma+\beta) \simeq Jq(\varepsilon)$, as desired. □

COROLLARY 4.4. *For any polynomial* $q(x)$ *we have*

(1) $\Delta q(x+\beta) \simeq Dq(x)$,

(2) $Jq(x+\beta) \simeq q(x)$,

(3) $Jq(x) \simeq q(x+\gamma)$.

Proof. Set $p(x) = q(x+c)$, where c is any constant. Now apply (1), thereby obtaining

$$\Delta q(c+\beta) \simeq Dq(c+\varepsilon) \simeq Dq(c).$$

This equality holds for all constants c; hence the first assertion. The remaining two assertions follow similarly. □

In other words, given the difference equation

$$\Delta f(x) = q'(x),$$

where $q(x)$ is an arbitrary polynomial and a polynomial solution $f(x)$ is desired, that solution is determined up to an additive constant and is given by the polynomial

$$q(x+\beta).$$

This explicit formula for the solution of a difference equation can be extended to a far more general class of functions $q(x)$.

For example, we obtain the classical formula

$$0^n + 1^n + 2^n + \cdots + (x-1)^n \simeq \frac{(x+\beta)^{n+1} - \beta^{n+1}}{n+1}.$$

The preceding formula is (up to a constant) an immediate consequence of the preceding remarks concerning difference equations. To determine the added constant $-\beta^{n+1}/(n+1)$, let the expression on the left-hand side be denoted by the function $s(x)$, taking its arguments from the positive integers. Note that in order for $\Delta s(x) = x^n$ to hold at $x = 0$ we have $s(0) = 0$. So it suffices to perform the obvious check that the constant term in the umbral polynomial on the right-hand side is 0.

As an application of Corollary 4.4, we obtain what is perhaps the simplest proof of the Euler–Maclaurin summation formula.

PROPOSITION 4.5 (Euler–Maclaurin). *For any polynomial $p(x)$ we have the identity*

$$p(x) \simeq Jp(x) + \beta\Delta p(x) + \frac{\beta^2}{2!}\Delta Dp(x) + \frac{\beta^3}{3!}\Delta D^2 p(x) + \cdots.$$

Proof. Again, let γ be an inverse to the Bernoulli umbra β. By Taylor's formula we have

$$p(x) \simeq p(\beta + x + \gamma) = \sum_{k\geq 0} \frac{\beta^k}{k!} D^k p(x+\gamma).$$

By Corollary 4.4 we have $D^k p(x+\gamma) \simeq D^k Jp(x) = D^{k-1}\Delta p(x)$ for $k > 0$. Expanding this relation yields the conclusion. □

5. The Nörlund umbra. A great many sequences occurring in combinatorics and number theory can be represented by sums or differences of umbrae exchangeable with the Bernoulli umbra. Such sequences are often obtained by specializing the following doubly indexed sequence.

DEFINITION 5.1. *Let n be any integer and let k be a nonnegative integer. The number $B_k^{(n)}$ defined as*

$$B_k^{(n)} \simeq (n\,.\,\beta)^k$$

will be called the kth Nörlund number of order n. *Correspondingly, any umbra exchangeable with $n\,.\,\beta$ for some integer n is called a* Nörlund umbra.

For $n = 1$ we may omit the superscript since $B_k^{(1)} \simeq \beta^k \simeq B_k$.

Recall that the operator Δ on polynomials of one variable may be written as $D + D^2/2! + D^3/3! + \cdots$. We denote the invertible operator $(\sum_{i\geq 1} D^{i-1}/i!)$ by Δ/D. It is easily verified that $J = \Delta/D$. The following is a generalization of (1).

THEOREM 5.2.

$$f\Big((n+j)\,.\,\alpha\Big) \simeq \left(\frac{\Delta}{D}\right)^{-n} f(j\,.\,\alpha) = J^{-n} f(j\,.\,\alpha)$$

holds for all polynomials f in one variable and all integers j, n, where the umbra α is a Bernoulli umbra. Conversely, if the equation holds for every polynomial f and some fixed integers $n \neq 0$ and j, then α is a Bernoulli umbra.

Proof. Assume $\alpha \equiv \beta$. Then Proposition 4.3 shows that $f(\beta) \simeq (\Delta/D)^{-1} f(\varepsilon)$. It suffices to assume $n > 0$ and $j = 0$. By induction on n we have

$$\begin{aligned} f(n\,.\,\beta) &\simeq \left(\frac{\Delta}{D}\right)^{-1} f\Big((n-1)\,.\,\beta\Big) \\ &\simeq \left(\frac{\Delta}{D}\right)^{-1} \left(\frac{\Delta}{D}\right)^{-(n-1)} f(\varepsilon). \end{aligned}$$

Replacing the polynomial $f(t)$ by $f(t + j\,.\,\beta)$ yields the result

$$f\Big((n+j)\,.\,\beta\Big) \simeq \left(\frac{\Delta}{D}\right)^{-n} f(j\,.\,\beta).$$

Conversely, the validity of the theorem both for the Bernoulli umbra β and some arbitrary umbra α for fixed $n \neq 0$ and j will enable us to prove that $n\,.\,\alpha \equiv n\,.\,\beta$. By Proposition 3.2 the conclusion follows.

Thus it suffices to show that if $p(t)$ is any polynomial in one variable then $p(n.\beta) \simeq p(n\,.\,\alpha)$. But

$$p(n\,.\,\beta) \simeq \left(\frac{\Delta}{D}\right)^{n} p(n\,.\,\alpha + n\,.\,\beta) \simeq p(n\,.\,\alpha). \qquad \square$$

It is worth stating several corollaries implicit in some of the preceding arguments.

COROLLARY 5.3.

(1) $\Delta^n f(n\,.\,\beta) \simeq D^n f(\varepsilon)$,
(2) $\Delta^n f(\varepsilon) \simeq D^n f(-n\,.\,\beta)$,
(3) $f(1) - f(0) \simeq f'(-1\,.\,\beta)$.

The third property allows the inverse umbra $-1\,.\,\beta$ to be immediately evaluated. Again, let γ be an umbra exchangeable with $-1\,.\,\beta$. Then

$$1^{n+1} - 0^{n+1} \simeq (n+1)(\gamma)^n,$$

so that

$$\gamma^n \simeq \frac{1}{n+1}.$$

6. Nörlund sequences. The Appell polynomial sequence associated with the Bernoulli numbers is the sequence of Bernoulli polynomials. We now concern ourselves with sequences of polynomials associated with the Nörlund umbrae. In particular, the Bernoulli polynomials form such a sequence.

DEFINITION 6.1. *Associated to the Nörlund umbra $n\,.\,\beta$ (for any integer n) is a sequence $(B_k^{(n)}(x))$ of* Nörlund polynomials of order n *and defined by*

$$B_k^{(n)}(x) \simeq (n\,.\,\beta + x)^k \simeq \sum_{i \geq 0} \binom{k}{i} B_i^{(n)} x^{k-i}.$$

The zeroth-order Nörlund polynomials are $B_k^{(0)}(x) = x^k$. The first-order sequence of Nörlund polynomials is the sequence of Bernoulli polynomials.

Applying Theorem 5.2 to $f(t) = (t+x)^k$, we obtain

$$\Delta^n B_k^{(i)}(x) = D^n B_k^{(i-n)}(x).$$

In particular,

$$\Delta^n B_k^{(n)}(x) = (k)_n x^{k-n}, \tag{4}$$

an identity that may be successfully used in the solution of difference equations.

COROLLARY 6.2. *Let $\phi^{(n)}(x)$ be the nth derivative of a polynomial $\phi(x)$ of degree d. Then the difference equation*

$$\Delta^n f(x) = \phi^{(n)}(x)$$

has $f(x) \simeq \phi(x + n \,.\, \beta)$ as a solution.

Observe that

$$\int_0^1 B_k^{(n)}(x)dx = B_k^{(n-1)}$$

since $JB_k^{(n)}(x)|_{x=0} \simeq (x + n \,.\, \beta + \gamma)^k|_{x=0} \simeq \big((n-1) \,.\, \beta\big)^k$.

Since the generating function for the Bernoulli numbers is $e^{\beta t} \simeq t/(e^t - 1)$ and since $e^{(n.\beta)t}e^{(m.\beta')t} \simeq e^{((n+m).\beta)t}$ for integers n and m, we have $e^{(n.\beta)t} \simeq (t/(e^t-1))^n$ and $e^{(x+n.\beta)t} \simeq e^{xt}\,(t/(e^t-1))^n$. These are the exponential generating functions for the Nörlund numbers and Nörlund polynomials each of order n, respectively.

7. Nörlund polynomials. We prove a number of facts concerning Nörlund polynomials that follow solely from the fact that the Nörlund polynomials of order n form an Appell sequence.

PROPOSITION 7.1. *If $n, n_1, n_2, \ldots, n_s$ are integers such that $\sum_{i=1}^s n_i = n$, then*

$$B_k^{(n)}(x_1 + \cdots + x_s) = \sum_{i_1,\ldots,i_s} \binom{k}{i_1,\ldots,i_s} B_{i_1}^{(n_1)}(x_1) \cdots B_{i_s}^{(n_s)}(x_s).$$

Proof. In umbral notation, the preceding identity is

$$(x_1 + \cdots + x_s + n \,.\, \beta)^k \simeq \Big((x_1 + n_1 \,.\, \beta') + \cdots + (x_s + n_s \,.\, \beta'')\Big)^k,$$

where $\beta', \ldots, \beta''$ are distinct umbrae, each exchangeable with β. □

Setting all $x_i = 0$ and all $n_i = 1$, we find

$$B_k^{(n)} = \sum_{i_1,\ldots,i_n} \binom{k}{i_1,\ldots,i_n} B_{i_1} \cdots B_{i_n}.$$

Nörlund polynomials of order n are an Appell sequence; hence the following proposition.

PROPOSITION 7.2. *For any integer n and for $k \geq 1$*

$$D_x\Big(B_k^{(n)}(x)\Big) \simeq D_x(x + n \,.\, \beta)^k = k(x + n \,.\, \beta)^{k-1} \simeq kB_{k-1}^{(n)}(x).$$

PROPOSITION 7.3 (inversion formula). *Suppose* $(r_i)_{i\in\mathbf{N}}$ *and* $(s_i)_{i\in\mathbf{N}}$ *are sequences of complex numbers such that* $r_0 = s_0 = 1$. *Then for* $n > 0$

$$r_k = \sum_{j=0}^{k} s_j \binom{k}{j} B_{k-j}^{(n)} \quad \text{if and only if} \quad s_k = \sum_{j=0}^{k} r_j \binom{k}{j} B_{k-j}^{(-n)}.$$

The coefficients $B_{k-j}^{(-n)}$ *of the sum on the right are all positive.*

Proof. Let ρ and σ be umbrae representing $(r_i)_{i\in\mathbf{N}}$ and $(s_i)_{i\in\mathbf{N}}$, respectively. Then $\rho \equiv \sigma + n \,.\, \beta$ if and only if $\rho + -n \,.\, \beta \equiv \sigma$, which is the desired conclusion. □

We conclude with an example of how the preceding techniques readily combine with facts particular to the Nörlund polynomials. Let k be greater than 0, and let n be any integer. Then

$$\Delta B_k^{(n)}(x) = D B_k^{(n-1)}(x) = k B_{k-1}^{(n-1)}(x).$$

8. Applications. We now prove some of the less trivial properties of Nörlund polynomials. Again, the proofs are, we believe, as short as they can be.

PROPOSITION 8.1. *For all integers* n *and for* $k \geq 0$

$$B_k^{(n)}(n - x) = (-1)^k B_k^{(n)}(x).$$

Proof.

$$\begin{aligned}(n - x + n \,.\, \beta)^k &\simeq \Big(- x + n \,.\, (\beta + 1)\Big)^k \\ &\simeq \Big(- x + n \,.\, (-\beta)\Big)^k \\ &\simeq (-1)^k (x + n \,.\, \beta)^k.\end{aligned}$$

The second line follows from Theorem 4.2, and the first and third results are from Proposition 3.2. □

As a further application of umbral techniques, apply Theorem 5.2 to $\sum_{j=0}^{m-1} B_k(x + j)$ to yield a telescoping sum.

LEMMA 8.2.

$$\begin{aligned}\sum_{j=0}^{m-1} (x + j + -1 \,.\, \beta)^k &\simeq \sum_{j=0}^{m-1} \frac{(x + j + 1)^{k+1} - (x + j)^{k+1}}{k + 1} \\ &= \frac{(x + m)^{k+1} - x^{k+1}}{k + 1} \simeq m^{k+1} (x/m + -1 \,.\, \beta)^k.\end{aligned}$$

This result forms the basis for another classical result.

PROPOSITION 8.3.

$$\sum_{j=0}^{m-1} B_k(x + (j/m)) = m^{-k+1} B_k(mx).$$

Proof. Let β, β' be exchangeable umbrae, and let $\gamma \equiv -1 \,.\, \beta$. Then

$$\sum_{j=0}^{m-1}\Big((x+\beta)+(j/m)\Big)^k \simeq m^{-k}\sum_{j=0}^{m-1}\Big(m(x+\beta)+\beta'+j+\gamma\Big)^k$$
$$\simeq m^{-k+1}\Big(mx+m\beta+\beta'+m\gamma\Big)^k$$
$$\simeq m^{-k+1}(mx+\beta')^k,$$

where the second umbral equality is by Lemma 8.2. □

Just as the Euler–Maclaurin summation formula was immediate from a computation with the Bernoulli umbra, we may easily derive a generalization of this formula by using the Nörlund umbrae.

We start with Taylor's formula,

$$f(x+y)=\sum_{k\geq 0}\frac{x^k}{k!}f^{(k)}(y), \tag{5}$$

and set $x = n\,.\,\beta$ and $y = x + n\,.\,\gamma$ to obtain

$$f(x)\simeq f(n\,.\,\beta+x+n\,.\,\gamma)\simeq\sum_{k\geq 0}\frac{(n\,.\,\beta)^k}{k!}f^{(k)}(x+n\,.\,\gamma).$$

Rewriting this by using Theorem 5.2 yields the following proposition.

PROPOSITION 8.4 (general Euler–Maclaurin summation formula). *Given a polynomial $f(x)\in D[x]$, the quotient field of D, as usual, containing $\mathbb{C}$, is*

$$f(x)\simeq\sum_{k=0}^{n-1}\frac{(n\,.\,\beta)^k}{k!}f^{(k)}(x+n\,.\,\gamma)+\sum_{k\geq n}\frac{(n\,.\,\beta)^k}{k!}\Delta^n f^{(k-n)}(x)$$

for $n\geq 0$.

Similarly, for any integer n we make use of Proposition 3.2 and expand

$$f(x+n)\simeq f\Big(n\,.\,(\beta+1)+x+n\,.\,\gamma\Big)\simeq f\Big(-1(n\,.\,\beta)+x+n\,.\,\gamma\Big)$$

to derive Proposition 8.5.

PROPOSITION 8.5 (Euler–Maclaurin second form). *For any integer $n\geq 0$*

$$f(x+n)\simeq\sum_{k=0}^{n-1}(-1)^k\frac{(n\,.\,\beta)^k}{k!}f^{(k)}(x+n\,.\,\gamma)+\sum_{k\geq n}(-1)^k\frac{(n\,.\,\beta)^k}{k!}\Delta^n f^{(k-n)}(x).$$

Setting $n = 1$ and summing the two preceding formulas as x ranges through $x, x+1,\dots,x+m$ and $x, x+1,\dots,x+m-1$, respectively, yields classical versions of the Euler–Maclaurin summation formula, to wit:

$$\sum_{i=0}^{m}f(x+i)=\int_x^{x+m+1}f(t)dt+\sum_{k\geq 1}\frac{B_k}{k!}\Big[f^{(k-1)}(x+m+1)-f^{(k-1)}(x)\Big]$$

and

$$\sum_{i=0}^{m}f(x+i)=\int_x^{x+m}f(t)dt+\frac{1}{2}\Big[f(x+m)+f(x)\Big]+\sum_{k\geq 2}(-1)^k\frac{B_k}{k!}\Big[f^{(k-1)}(x+m)-f^{(k-1)}(x)\Big].$$

The term $f(x)$ has been added to both sides in the second formula.

We next prove a theorem designed to simplify expressions of the form $(n.\beta)p(n.\beta)$ and examine some consequences. We begin with a basic result for the inverse Bernoulli umbra γ.

PROPOSITION 8.6. *If γ is exchangeable with $-1.\beta$, where β is a Bernoulli umbra, then for any polynomial p*

$$\gamma p'(\gamma) \simeq p(1) - p(\gamma) \simeq p(0) + p'(\gamma) - p(\gamma).$$

Proof. For $n \geq 0$ we have

$$n\gamma^n \simeq \frac{n}{n+1} \simeq 1 - \frac{1}{n+1} \simeq 1^n - \gamma^n.$$

The first identity follows by linearity, and the second is an application of the defining relation (1) of Bernoulli umbra. □

Let γ' be exchangeable with γ. Recall that $q'(\gamma) \simeq q(1) - q(0)$. Letting $q(t) = p(1 - t\gamma')$, we obtain $-\gamma p'(1 - \gamma\gamma') \simeq p(1 - \gamma) - p(1) \simeq p(\gamma) - p(1)$, the last equivalence coming from Theorem 4.2. By Proposition 8.6 this is equivalent to $-\gamma p'(\gamma)$. Evaluating at $p(t) = t^n$, we obtain the elegant identity

$$\sum_{k\geq 0} \binom{n-1}{k} (-1)^k \frac{1}{(k+1)(k+2)} = \frac{1}{n+1}.$$

By recalling Proposition 3.3, Proposition 8.6 may be rewritten as

$$\beta p'(\gamma + \beta) \simeq -\gamma p'(\gamma + \beta) \simeq -\Big[p(\beta) + p'(\varepsilon) - p(\varepsilon)\Big]. \tag{6}$$

We thereby obtain the following theorem.

THEOREM 8.7.

$$(n\,.\,\beta)p'(n\,.\,\beta + n\,.\,\gamma) \simeq -n\Big[p(\beta) + p'(\varepsilon) - p(\varepsilon)\Big].$$

Proof. By Proposition 3.3 the left-hand side in Theorem 8.7 is equivalent to $n[\beta p'(\beta + \gamma)]$. By (6) this is equivalent to the right-hand side. □

Now we use Theorem 8.7 to find an alternative formula for $(x + n\,.\,\beta)^k$, with $n \neq 0$.

$$\begin{aligned}
(x + n\,.\,\beta)^k &= (x + n\,.\,\beta)(x + n\,.\,\beta)^{k-1} \\
&= n\,.\,\beta(x + n\,.\,\beta)^{k-1} + x(x + n\,.\,\beta)^{k-1} \\
&\simeq -\frac{n}{k}\left[\Big(x + (n+1)\,.\,\beta\Big)^k + k(x + n\,.\,\beta)^{k-1} - (x + n\,.\,\beta)^k\right] \\
&\quad + x(x + n\,.\,\beta)^{k-1} \\
&\simeq -\frac{n}{k}\left[\Big(x + (n+1)\,.\,\beta\Big)^k + \frac{k}{n}(n - x)(x + n\,.\,\beta)^{k-1} - (x + n\,.\,\beta)^k\right].
\end{aligned}$$

Rewriting this as a recursion in the Nörlund polynomials, we obtain Corollary 8.8.

COROLLARY 8.8. *For $n \neq 0$ and $k \geq 1$*

$$B_k^{(n+1)}(x) = \frac{k}{n}(x - n)B_{k-1}^{(n)} + \left(1 - \frac{k}{n}\right) B_k^{(n)}(x).$$

Making the appropriate substitutions yields Corollary 8.9.

COROLLARY 8.9. *For $n \neq 0$ and $k \geq 0$*

$$B_k^{(n+1)}(n) = (-1)^k \left(1 - \frac{k}{n}\right) B_k^{(n)}.$$

Proof. By Corollary 8.8 the left-hand side in Corollary 8.9 is umbrally equivalent to $(1-k/n)(n+n\,.\,\beta)^k$. But since $(n+n\,.\,\beta) \equiv n\,.\,(1+\beta) \equiv n\,.\,(-\beta)$, this is umbrally equivalent to the right-hand side. □

9. Stirling numbers. Recall that the Stirling numbers of the second kind $S(n,k)$ are defined as the number of partitions of a set of n elements into k parts. An easy computation shows that

$$k!S(n,k) = [\Delta^k x^n]_{x=0}.$$

Remarkably, the Stirling numbers are very simply related to the Nörlund sequence.

PROPOSITION 9.1. *If $n \geq k \geq 0$, then*

$$\binom{n}{k} B_{n-k}^{(-k)} = S(n,k).$$

Proof. Suppose $n \geq k \geq 0$, and let p be the polynomial $p(t) = t^n$. Then by Theorem 5.2

$$(n)_k (x + -k\,.\,\beta)^{n-k} \simeq D^k p(x + -k\,.\,\beta) \simeq \Delta^k p(x).$$

Setting $x = 0$ and rewriting yields

$$B_{n-k}^{(-k)} \simeq \frac{\Delta^k \varepsilon^n}{(n)_k} \simeq \frac{k!S(n,k)}{(n)_k},$$

where $(n)_k = (n)(n-1)\cdots(n-k+1)$ denotes n falling factorial k. □

PROPOSITION 9.2. *For $n > 0$*

$$B_n^{(n+1)}(x) = (x-1)_n \quad \text{and} \quad B_n^{(n+1)} = (-1)^n n!. \tag{7}$$

Proof. The second half of the proposition is immediate from setting $x = 0$ in the first half.

To prove the first half, it suffices to show that for $n > 1$, $B_n^{(n+1)}(x) = (x - n)B_{n-1}^{(n)}(x)$. But this is precisely Corollary 8.8 for $k = n$. □

Now from Proposition 9.2 we obtain

$$(x-1)\cdots(x-n) \simeq (x + (n+1)\,.\,\beta)^n \simeq \sum_{k=0}^{n} \binom{n}{k} x^k \Big((n+1)\,.\,\beta\Big)^{n-k}. \tag{8}$$

By replacing x by $-x$ and n by $n-1$, this may be rewritten as

$$(-1)^n x(x+1)\cdots(x+n-1) = \sum_{k=1}^{n} \binom{n-1}{k-1} (-1)^k x^k B_{n-k}^{(n)}.$$

We have $x(x+1)\cdots(x+n-1) = \sum_{k=0}^{n} s(n,k)x^k$. The coefficients $s(n,k)$ are called the Stirling numbers of the first kind, which have numerous combinatorial interpretations. We have thus expressed the Stirling numbers of the first kind in terms of the Nörlund sequence.

PROPOSITION 9.3. *For $n \geq k \geq 1$*

$$(-1)^{n-k}s(n,k) = \binom{n-1}{k-1} B_{n-k}^{(n)}.$$

This relationship between Stirling numbers and higher-order Bernoulli numbers allows easy results on the one to be transformed into results on the other.

PROPOSITION 9.4. *If $n \geq k \geq 0$, then*

$$\sum_{i=0}^{k} \binom{k}{i} (-1)^i (k-i)^n = (n)_k B_{n-k}^{(-k)}.$$

Proof. The left-hand side in Proposition 9.4 is equivalent to $\Delta^k f(k \,.\, \beta')$, where $f(t) = (-k \,.\, \beta + t)^n$. Since the right-hand side is equivalent to $f^{(k)}(t)$, the result follows by Theorem 5.2. □

In terms of Stirling numbers, if $n \geq k \geq 0$, we have proved the classical identity

$$\sum_{i=0}^{k} \binom{k}{i} (-1)^i (k-i)^n = k!S(n,k).$$

Recall Newton's formula, namely, for any polynomial f

$$f(x) = \sum_{k \geq 0} \binom{x}{k} \Delta^k f(0). \tag{9}$$

Applying Newton's formula to the polynomial $f(t) = (t + n \,.\, \beta)^d$, for any integer n, yields

$$\begin{aligned}(x + n \,.\, \beta)^d &= \sum_{k \geq 0} \binom{x}{k} \Delta^k (n \,.\, \beta)^d \\ &\simeq \sum_{k \geq 0} \binom{x}{k} D^k \Big((n-k) \,.\, \beta\Big)^d \\ &\simeq \sum_{k \geq 0} \binom{x}{k} (d)_k \Big((n-k) \,.\, \beta\Big)^{d-k}.\end{aligned}$$

From the preceding computation we obtain the expansion of the Nörlund polynomials as linear combinations of the lower factorials $(x)_k$.

PROPOSITION 9.5. *For any integer n*

$$B_d^{(n)}(x) = \sum_{k=0}^{d} \binom{d}{k} (x)_k B_{d-k}^{(n-k)}.$$

Recalling Proposition 9.1, for $n = 0$, we obtain a classical identity for the Stirling numbers of the second kind.

Since $D^{n-k}\big(x+(n+1).\beta\big)^n \simeq (n)_{n-k}\big(x+(n+1).\beta\big)^k$, differentiating the left-hand side of equation (7) $n-k$ times yields the following corollary.

COROLLARY 9.6. *For* $n \geq k \geq 0$

$$B_k^{(n+1)}(x) = \frac{k!}{n!} D^{n-k}(x-1)(x-2)\cdots(x-n).$$

Corollary 9.6 has the following elegant consequence.

PROPOSITION 9.7.

$$B_{n-1}^{(n+1)}(x) = \frac{1}{n}(x-1)(x-2)\cdots(x-n)\left(\frac{1}{x-1}+\frac{1}{x-2}+\cdots+\frac{1}{x-n}\right),$$

in which setting n *to* $n+1$ *and* x *to* 0 *yields*

$$B_n^{(n+2)} = (-1)^n n!\left(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n+1}\right).$$

Finally, Proposition 9.2 yields the following proposition.

PROPOSITION 9.8.

$$B_n^{(n)}(x) = \int_x^{x+1}(t-1)(t-2)\cdots(t-n)dt$$

and

$$B_n^{(n)} = \int_0^1 (t-1)(t-2)\cdots(t-n)dt.$$

Proof. The first equation in Proposition 9.8 clearly implies the second. To prove the first, rewrite the equation in umbral form by using Corollary 9.6 to produce

$$\begin{aligned}\int_x^{x+1}(t-1)(t-2)\cdots(t-n)dt &= \int_x^{x+1} B_n^{(n+1)}(t)dt \\ &\simeq \Big(x+(n+1)\,.\,\beta+\gamma\Big)^n \\ &\simeq B_n^{(n)}(x),\end{aligned}$$

where the second line is from Corollary 4.4. □

REFERENCES

[Be1] E. T. BELL, *Algebraic Arithmetic*, American Mathematical Society, New York, 1927.

[Be2] ———, *Postulational bases for the umbral calculus*, Amer. J. Math., 62 (1940), pp. 717–724.

[Bl1] J. BLISSARD, *Theory of generic equations*, Quart. J. Pure Appl. Math., 4 (1861), pp. 279–305; 5 (1862), pp. 58–75, 184–208.

[Bl2] ———, *Note on certain remarkable properties of numbers*, Quart. J. Pure Appl. Math., 5 (1862), p. 184.

[Bl3] ———, *On the discovery and properties of a peculiar class of algebraic formulae*, Quart. J. Pure Appl. Math., 5 (1862), pp. 325–335.

[Bl4] ———, *Examples of the use and application of representative notation*, Quart. J. Pure Appl. Math., 6 (1864), pp. 49–64.

[Bl5] ———, *On the generalization of certain formulae investigated by Mr. Walton*, Quart. J. Pure Appl. Math., 6 (1864), pp. 167–179.

[Bl6] J. BLISSARD, *Researches in analysis*, Quart. J. Pure Appl. Math., 6 (1864), pp. 142–257.

[Bl7] ———, *On the properties of the $\Delta^n 0^n$ class of numbers and of others analogous to them, as investigated by means of representative notation*, Quart. J. Pure Appl. Math., 8 (1867), pp. 85–110; 9 (1868), pp. 82–94, 154–171.

[Bl8] ———, *Note on a certain formula*, Quart. J. Pure Appl. Math., 9 (1868), pp. 71–76.

[Bl9] ———, *On certain properties of the gamma function*, Quart. J. Pure Appl. Math., 9 (1868), pp. 280–296.

[J] C. JORDAN, *Calculus of Finite Differences*, Chelsea, New York, 1960.

[L] E. LUCAS, *Théorie des nombres* I, Gauthier–Villars, Paris, 1876.

[M-T] L. M. MILNE-THOMSON, *The Calculus of Finite Differences*, Macmillan, London, 1951.

[MR] R. MULLIN AND G.-C. ROTA, *Theory of binomial enumeration*, in Graph Theory and Its Applications, Academic Press, New York, 1970.

[Ni] N. NIELSEN, *Traité élémentaire des nombres de Bernoulli*, Gauthier–Villars, Paris, 1923.

[Nö] N. E. NÖRLUND, *Vorlesungen Über Deifferenzenrechnung*, Chelsea, New York, 1954.

[Ri] J. RIORDAN, *An Introduction to Combinatorial Analysis*, John Wiley, New York, 1958.

[R] S. ROMAN, *The Umbral Calculus*, Academic Press, Orlando, FL, 1984.

[RR] S. ROMAN AND G.-C. ROTA, *The umbral calculus*, Adv. Math., 27 (1978), pp. 95–188.

[Ro] G.-C. ROTA, *The number of partitions of a set*, Amer. Math. Monthly, 71 (1964), pp. 498–504.

[RKO] G.-C. ROTA, D. KAHANER, AND A. ODLYZKO, *Finite operator calculus*, J. Math. Anal. Appl., 42 (1973), pp. 685–760.

[S] R. STANLEY, *Enumerative Combinatorics*, vol. I. Wadsworth and Brooks/Cole, Monterey, CA, 1986.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 712–719, March 1994

029

SPECIALIZATIONS OF GENERALIZED LAGUERRE POLYNOMIALS*

R. SIMION† AND D. STANTON‡

Abstract. Three specializations of a set of orthogonal polynomials with "8 different q's" are given. The polynomials are identified as q-analogues of Laguerre polynomials, and the combinatorial interpretation of the moments gives infinitely many new Mahonian statistics on permutations.

Key words. Laguerre polynomials, permutation statistics

AMS subject classifications. 05E35, 33C45

1. Introduction.

The Laguerre polynomials $L_n^{\alpha}(x)$ have been extensively studied, analytically [3] and combinatorially [5], [9]. There is also a classical set of q-Laguerre polynomials [7]. Recently, a set of orthogonal polynomials generalizing the Laguerre polynomials has been studied [8]. These polynomials in some sense have "8 different q's." Various specializations of them give orthogonal polynomials associated with many types of combinatorial objects. The purpose of this paper is to present the specializations that are true q-analogues of $L_n^0(x)$. By this we mean that the nth moments, instead of being $n!$, are basically $n!_q$.

Here we present three specializations whose moments lead to new Mahonian statistics on permutations (Theorems 2, 3, and 4). In fact, infinitely many Mahonian statistics can be derived from those presented here. Moreover, the theorems obtained from our specializations follow easily from classical analytic facts, but are combinatorially nontrivial.

We shall use the terminology and notation found in [6], and let

$$[n]_q = \frac{1-q^n}{1-q}, \qquad [n]_{r,s} = \frac{r^n - s^n}{r-s}.$$

2. The polynomials and their moments.

Any set of monic orthogonal polynomials satisfies the three-term recurrence relation

$$p_{n+1}(x) = (x - b_n)p_n(x) - \lambda_n p_{n-1}(x). \tag{2.1a}$$

For the set of orthogonal polynomials with 8 different "q's" considered in [8], the coefficients are

$$b_n = a[n+1]_{r,s} + b[n]_{t,u}, \qquad \lambda_n = ab[n]_{p,q}[n]_{v,w}. \tag{2.1b}$$

*Received by the editors March 23, 1993; accepted for publication (in revised form) April 13, 1993.

†Department of Mathematics, George Washington University, Washington, DC 20052. This work was carried out in part during this author's visits at the Mittag–Leffler Institute and the University of Québec at Montréal, Montréal, Canada, and with partial support through National Science Foundation grant DMS-9108749.

‡School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported by the Mittag–Leffler Institute and by National Science Foundation grant DMS-9001195.

We refer to the polynomials defined by (2.1) as the *octabasic Laguerre polynomials.* They generalize the Laguerre polynomials and are the polynomials whose specializations we consider in this paper.

The fundamental combinatorial fact (Theorem 1) that we need here concerns the moments for these polynomials. They are generating functions for permutations according to certain statistics.

For the definition of these statistics, it is convenient to represent a permutation σ as a word $\sigma(1)\sigma(2)\cdots\sigma(n)$ consisting of increasing runs, separated by the descents of the permutation. For example, the permutation $\sigma = 26|357|4|189$ has 4 runs separated by 3 descents, and we write $\mathrm{run}(\sigma) = 4$. The runs of length 2 or more will be called *proper runs* and those of length 1 will be called *singleton runs.*

The elements $\sigma(i)$ of σ fall into four classes: the elements that begin proper runs (openers), the elements that close proper runs (closers), the elements that form singleton runs (singletons), and the elements that continue runs (continuators). We shall abbreviate these classes of elements "op," "clos," "sing," and "cont," respectively. In the example, $\mathrm{op}(\sigma) = \{3,2,1\}$, $\mathrm{clos}(\sigma) = \{6,7,9\}$, $\mathrm{sing}(\sigma) = \{4\}$, and $\mathrm{cont}(\sigma) = \{5,8\}$.

DEFINITION 1. For $\sigma \in S_n$, the statistics $\mathrm{lsg}(\sigma)$ and $\mathrm{rsg}(\sigma)$ are defined by

$$\mathrm{lsg}(\sigma) = \sum_{i=1}^{n} \mathrm{lsg}(i), \qquad \mathrm{rsg}(\sigma) = \sum_{i=1}^{n} \mathrm{rsg}(i),$$

where $\mathrm{lsg}(i)$ = the number of runs of σ strictly to the left of i which contain elements smaller and greater than i, and $\mathrm{rsg}(i)$ = the number of runs of σ strictly to the right of i which contain elements smaller and greater than i.

We also define the lsg and rsg on the openers of σ

$$\mathrm{lsg(op)}(\sigma) = \sum_{i\in\mathrm{op}(\sigma)} \mathrm{lsg}(i), \qquad \mathrm{rsg(op)}(\sigma) = \sum_{i\in\mathrm{op}(\sigma)} \mathrm{lsg}(i).$$

The statistics lsg and rsg have analogous definitions on each of the remaining three classes of elements.

For example, if $\sigma = 26|357|4|189$, then $\mathrm{lsg}(7) = 0$, $\mathrm{rsg}(7) = 1$, $\mathrm{lsg(op)}(\sigma) = 0+1+0 = 1$, $\mathrm{rsg(op)}(\sigma) = 1+1+0 = 2$, $\mathrm{lsg(clos)}(\sigma) = 0$, $\mathrm{rsg(clos)}(\sigma) = 2+1+0 = 3$, etc.

THEOREM 1. *The* n*th moment* μ_n *for the octabasic Laguerre polynomials is*

$$\mu_n = \sum_{\sigma\in S_n} r^{\mathrm{lsg(sing)}(\sigma)} s^{\mathrm{rsg(sing)}(\sigma)} t^{\mathrm{lsg(cont)}(\sigma)} u^{\mathrm{rsg(cont)}(\sigma)} p^{\mathrm{lsg(op)}(\sigma)} q^{\mathrm{rsg(op)}(\sigma)} v^{\mathrm{lsg(clos)}(\sigma)} w^{\mathrm{rsg(clos)}(\sigma)} a^{\mathrm{run}(\sigma)} b^{n-\mathrm{run}(\sigma)}.$$

Proof. The Viennot theory [9],[10] gives μ_n as a generating function for Motzkin paths from $(0,0)$ to $(n,0)$. From (2.1b) the paths have four types of edges (or steps):

(1) northeast (NE) edges starting at level k with weight in $a[k+1]_{p,q}$,
(2) southeast (SE) edges starting at level k with weight in $b[k]_{v,w}$,
(3) east (E solid) edges starting at level k with weight in $a[k+1]_{r,s}$,
(4) east (E dotted) edges starting at level k with weight in $b[k]_{t,u}$,

where "weight in $\alpha[m]_{c,d}$" means that the weight of the edge is one of the monomials $\alpha c^{m-1}, \alpha c^{m-2}d, \ldots, \alpha d^{m-1}$, which appear in $\alpha[m]_{c,d}$.

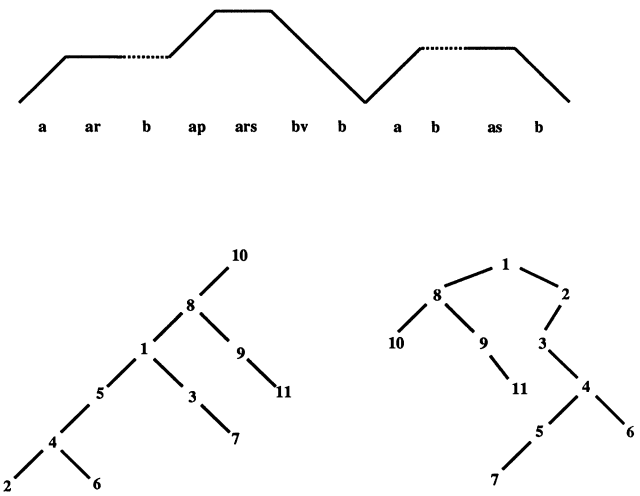


FIG. 1. *The permutation* 10 8 9 11 1 3 7 5 4 6 2 *as a weighted Motzkin path, runs, and a binary tree.*

The weight of a Motzkin path is then the product of the weights of its edges. An example of such a weighted path is given in Fig. 1.

In the statement of Theorem 1 we claim that μ_n is the sum over all permutations in S_n of monomial weights defined in terms of our permutation statistics. To prove the theorem we construct a bijection between the weighted Motzkin paths of length n and permutations in S_n, so that the weight of a path is equal to the weight of its corresponding permutation.

Given a weighted Motzkin path of length n, its corresponding permutation will be constructed in n stages. We begin at the origin and with the empty permutation. We traverse the path from left to right, and the ith step will determine where to insert i in the current (partial) permutation $\sigma_{i-1} \in S_{i-1}$. The end result will be a permutation $\sigma_n = \sigma \in S_n$. Depending on the type of the ith step of the path, i will belong to one or another of the four classes of elements of σ:

(1) NE: $i \in \text{op}(\sigma)$,
(2) SE: $i \in \text{clos}(\sigma)$,
(3) E solid: $i \in \text{sing}(\sigma)$,
(4) E dotted: $i \in \text{cont}(\sigma)$.

The weight $\alpha c^j d^k$ of the ith step of the path determines, as described below, the exact position in σ_{i-1} where we insert i as a point in the appropriate class of elements.

A run in σ_{i-1} will be called an *active run* if its maximum is a (future) opener or continuator in σ. Notice that the first step of the path starts at level 0 and that σ_0, being the empty permutation, has no active runs. Inductively, assume that the level h at which the ith step starts is equal to the number of active runs in σ_{i-1}, and recall the relation between the weight $\alpha c^j d^k$ of a step and its starting level h: for NE and E solid steps we have $j + k = h$, while for other steps we have $j + k = h - 1$.

If the ith step is E dotted or SE, then the element i is adjoined to the $(j+1)$st active run of σ_{i-1}, as a continuator or closer, respectively. This is well defined since, as discussed above, $j + k$ is one unit less than the number of active runs in σ_{i-1}. It follows as well that the new partial permutation, σ_i, will have as many active runs as the starting level of the $(i+1)$st step of the path.

If the ith step is NE or E solid, then $j+k$ is equal to the number of active runs in σ_{i-1}, and we insert i in the leftmost position possible such that it will have j of the *active* runs of σ_{i-1} strictly to its left and k of the *active* runs of σ_{i-1} strictly to its right. The position where i is inserted in this case ensures that i will be the initial element of a run, and it is again true that σ_i will have as many active runs as the starting level of the $(i+1)$st step of the path.

Note that in the final permutation σ the values of $\mathrm{lsg}(i)$ and $\mathrm{rsg}(i)$ are completely determined by the runs that were active in σ_{i-1} and the position, relative to these runs, where i was inserted. So, if the ith step has weight $\alpha c^j d^k$, then $\mathrm{lsg}(i) = j$ because each of the j open runs of σ_{i-1} that remain to the left of i upon its insertion will eventually be extended by at least one element greater than i. Similarly, $\mathrm{rsg}(i) = k$. Finally, $\alpha = a$ in the weight of the ith step corresponds precisely to i being the initial (possibly the only) element of a run. Consequently, the weight of a Motzkin path is equal to our intended weight for the corresponding permutation σ.

It remains to verify that this correspondence is bijective. We claim that, given $\sigma \in S_n$, we can reconstruct its associated Motzkin path, step by step, beginning at the right end of the path, $(n,0)$, since each stage of our construction is reversible.

Using our rules (1)–(4), each permutation in S_n will produce a Motzkin path of length n (not yet weighted), since $|\mathrm{op}(\sigma)| = |\mathrm{clos}(\sigma)|$ and each closer is larger than its corresponding opener. We must check that the weight $\alpha c^{\mathrm{lsg(i)}} d^{\mathrm{rsg(i)}}$ (where $\alpha = a$ if i is the initial element of a run, and $\alpha = b$ otherwise) is a valid weight for the ith step of the path. This follows immediately from the equality $\mathrm{lsg}(i) + \mathrm{rsg}(i) =$ the level of the left endpoint e_{i+1} of the partial path reconstructed from the values $n, n-1, \ldots, i+1$. The equality holds for $i = n$ since $\mathrm{lsg}(n) = \mathrm{rsg}(n) = 0$ in all permutations of S_n, and the left endpoint of the one-point path consisting just of $(n,0) = e_{n+1}$ is at level 0. Suppose the equality holds for $i+1$ and we will prove it for i. Observe that the level of e_{i+1} is equal to the number of SE steps minus the number of NE steps in the partial path from e_{i+1} to $(n,0)$. That is, the level of e_{i+1} is equal to the number of proper runs in σ whose maximum is larger than i, minus the number of proper runs in σ whose minimum is larger than i. But this is equal in turn with the number of proper runs in σ with maximum larger than i and minimum smaller than i, i.e., it is equal to $\mathrm{lsg}(i) + \mathrm{rsg}(i)$. It now becomes clear that our map from weighted Motzkin paths to permutations is indeed invertible. □

Figure 1 shows the weighted Motzkin path that corresponds with the permutation $\sigma = 10|89\,11|137|5|46|2$. We have also included a binary tree representation of the permutation, deeming it of possible interest to the readers familiar with [9]. The bijection constructed in the proof above is related to Viennot's correspondence between Motzkin paths and permutations [9]. As an intermediate step in Viennot's correspondence, Motzkin paths and permutations are encoded by increasingly labeled binary trees.

We will be concerned with specializations of the 8 q's under which the moments in Theorem 1 become multiples of $n!_q = [n]_q[n-1]_q \cdots [1]_q$. It is clear that the parameter b can be rescaled to 1. Also, in considering specializations, we can take advantage of the property — obvious from (2.1) — that the moments are fixed under the interchange of $\{r,s\}$, $\{t,u\}$, $\{p,q\}$, and $\{v,w\}$, and also fixed if p and q are interchanged with v and w.

3. The specializations. In this section we state three different specializations of the polynomials defined by (2.1). The polynomials that arise are monic little q-Jacobi, sums of two little q-Jacobi, and classical q-Laguerre. Each of these three cases

will have moments that are basically $n!_q$.

First we choose the parameters so that the polynomials coincide with the monic form of the little q-Jacobi polynomials [6], $p_n(xq(1-q);q^\alpha,0;q)$, which have

$$b_n = q^{n-1}[n+1+\alpha]_q + q^{n+\alpha-1}[n]_q, \qquad \lambda_n = q^{2n-3+\alpha}[n]_q[n+\alpha]_q. \tag{3.1}$$

The appropriate specialization occurs only for $\alpha = 0$, $p_n(xq(1-q);1,0;q)$, and is $r = t = p = v = q^2$, $s = u = q = w$, $a = 1/q$, $b = 1$. The explicit formula for the polynomials, which is just the definition of the little q-Jacobi polynomials, is

$$p_n(x) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix}_q [n]_q \cdots [n-k+1]_q (-1)^k x^{n-k} q^{\binom{k-1}{2}-1}. \tag{3.2}$$

The measure for $p_n(x;q^\alpha,q^\beta;q)$ is purely discrete, with masses of

$$\frac{q^{(\alpha+1)i}(q^{\beta+1};q)_i(q^{\alpha+1};q)_\infty}{(q;q)_i(q^{\alpha+\beta+2};q)_\infty}$$

at $x = q^i$. An easy calculation shows that the moments are given by $\mu_n = (q^{\alpha+1};q)_n/(q^{\alpha+\beta+2};q)_n$. Thus for $p_n(xq(1-q);1,0;q)$, we have $\mu_n = q^{-n}n!_q$. Based upon these remarks, Theorem 1 becomes the following theorem. An equivalent theorem has been given in [2, Prop. 5.2].

THEOREM 2. *For $\sigma \in S_n$, let*

$$s(\sigma) := n - \mathrm{run}(\sigma) + 2\mathrm{lsg}(\sigma) + \mathrm{rsg}(\sigma).$$

Then

$$\sum_{\sigma \in S_n} q^{s(\sigma)} = n!_q.$$

Moreover, we see from the symmetry of (2.1) with respect to the 4 pairs of "q's" that Theorem 2 holds for 16 statistics related to $s(\sigma)$. These 16 are obtained by choosing the coefficients 1 and 2 for lsg and rsg independently for the four types of elements of σ. This means, for example, that

$$\begin{aligned} s'(\sigma) = n - \mathrm{run}(\sigma) &+ \mathrm{lsg(sing)} + 2\mathrm{rsg(sing)} + \mathrm{lsg(op)} + 2\mathrm{rsg(op)} \\ &+ 2\mathrm{lsg(cont)} + \mathrm{rsg(cont)} + 2\mathrm{lsg(clos)} + \mathrm{rsg(clos)} \end{aligned}$$

also satisfies Theorem 2. We will see later that in fact there are infinitely many equidistributed statistics related to $s(\sigma)$.

For our second specialization we consider

$$b_n = q^{n+1}[n+1]_q + q^{n-1}[n]_q, \qquad \lambda_n = q^{2n-1}[n]_q[n]_q. \tag{3.3}$$

The appropriate values are $r = t = p = v = q^2$, $s = u = q = w$, $a = q$, $b = 1$. The polynomials turn out to be a sum of two little q-Jacobi polynomials,

$$\begin{aligned} p_n(x) = n!_q q^{\binom{n}{2}}(-1)^n \Bigg[{}_2\phi_1 \left(\begin{matrix} q^{-n} & 0; \\ & q \end{matrix} \; q, \; xq(1-q) \right) \\ - (1-q^n)\,{}_2\phi_1 \left(\begin{matrix} q^{1-n} & 0; \\ & q^2 \end{matrix} \; q, \; xq(1-q) \right) \Bigg], \end{aligned}$$

or equivalently,

(3.4)
$$p_n(x) = x^n + \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix}_q [n]_q \cdots [n-k+2]_q ([n-k]_q + q^n)(-1)^k x^{n-k} q^{\binom{k}{2}} + (-1)^n q^{\binom{n+1}{2}} n!_q.$$

We omit the proof of these formulas. It is a verification of the recurrence relation (2.1) from the recurrence relation for the little q-Jacobi polynomials.

Since these polynomials do not explicitly appear in the literature, we cannot compute the moments by quoting the relevant facts about their measure. Nonetheless, the moments and measure are easily determined.

PROPOSITION 1. *The moments for the polynomials in* (3.3) *are* $\mu_0 = 1$, *and* $\mu_n = q\, n!_q, n > 0$. *The measure is purely discrete, with masses of* $q^i(q;q)_\infty/(q;q)_{i-1}$ *at* $q^{i-1}/(1-q)$, $i \geq 1$, *and a mass of* $1-q$ *at* 0.

Proof. The q-binomial theorem clearly shows that the total mass $\mu_0 = (1-q)+q = 1$. It also implies

$$\begin{aligned} \mu_n &= \sum_{i=1}^{\infty} \frac{q^{(i-1)n} q^i (q;q)_\infty}{(1-q)^n (q;q)_{i-1}} + (1-q)\delta_{n,0} \\ &= q n!_q + (1-q)\delta_{n,0}. \end{aligned}$$

Thus the stated measure has the right moments. To show that the polynomials are orthogonal with respect to this measure, note that it is easy to check, from the explicit formula (3.4), that the moments annihilate $p_1, p_2, \ldots$. Hence, the linear functional defined by the measure annihilates $p_1, p_2, \ldots$. Finally, the three-term recurrence now shows that the polynomials are indeed orthogonal. □

We then get a companion theorem to Theorem 2. As in the case of Theorem 2, we have 16 equivalent versions of the statistic $s(\sigma)$, by assigning coefficients 1 and 2 to lsg and rsg independently on openers, continuators, closers, and singletons.

THEOREM 3. *For* $\sigma \in S_n$, *let*

$$s(\sigma) := \mathrm{run}(\sigma) - 1 + 2\mathrm{lsg}(\sigma) + \mathrm{rsg}(\sigma).$$

Then

$$\sum_{\sigma \in S_n} q^{s(\sigma)} = n!_q.$$

We remark that Theorems 2 and 3 are valid for an infinite number of variations of the statistic $s(\sigma)$. It is easy to verify that for each $\sigma \in S_n$,

(3.5) $$\mathrm{lsg(op)}(\sigma) + \mathrm{rsg(op)}(\sigma) = \mathrm{lsg(clos)}(\sigma) + \mathrm{rsg(clos)}(\sigma).$$

(In fact there is a specialization of $\{r, s, t, u, p, q, v, w\}$ with one free parameter giving (3.5).) Therefore the value of $s(\sigma)$ remains the same if the coefficients $\{1, 2\}$ are replaced on the openers with $\{1+c, 2+c\}$, and on the closers with $\{1-c, 2-c\}$. This provides a variation of Theorems 2 and 3 for each choice of the real parameter c. For example, $c = 1$ gives the unusual choice of coefficients $\{2, 3\}$ and $\{0, 1\}$.

Our third choice for specialization is the set of the classical q-Laguerre polynomials $L_n^\alpha(x(1-q);q)$ [7], [6, p. 194], whose monic form has

(3.6) $$b_n = q^{-2n-\alpha}[n]_q + q^{-2n-1-\alpha}[n+1+\alpha]_q, \qquad \lambda_n = q^{1-4n-2\alpha}[n]_q[n+\alpha]_q.$$

The appropriate values are $r = t = p = v = q^{-2} = b$, $s = u = q = w := q^{-1} = a$ for $L_n^0(x(1-q), q)$. The explicit formula for the monic form of $L_n^\alpha(x(1-q); q)$ is

$$p_n(x) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix}_q [n+\alpha]_q \cdots [n+\alpha-k+1]_q (-1)^k x^{n-k} q^{k(k-\alpha-2n)}. \tag{3.7}$$

Again a measure of these polynomials is explicitly known [7, Thm. 1], and the moments for $L_n^\alpha(x(1-q); q)$ can be found as

$$\mu_n = (q^{\alpha+1}; q)_n q^{-n\alpha-\binom{n+1}{2}}/(1-q)^n.$$

For $\alpha = 0$ this is $q^{-\binom{n+1}{2}} n!_q$. However, the combinatorial version of this theorem is equivalent to Theorem 2, if q is replaced with $1/q$. Thus no new combinatorial theorem results.

4. The "odd" polynomials. If $r = p$, $s = q$, $t = v$, and $u = w$, then the polynomials defined by (2.1) are the "even" polynomials for the polynomials defined by (see [1, p. 41])

$$b_n = 0, \quad \lambda_{2n} = b[n]_{t,u}, \quad \lambda_{2n+1} = a[n+1]_{r,s}.$$

The "odd" polynomials have the coefficients

$$b_n = a[n+1]_{r,s} + b[n+1]_{r,s}, \qquad \lambda_n = ab[n+1]_{r,s}[n]_{t,u}. \tag{4.1}$$

The moments for these "odd" polynomials satisfy $\mu_n(\text{odd}) = \mu_{n+1}(\text{even})/\mu_1(\text{even})$. Since all of our specializations in §3 satisfied $r = p$, $s = q$, $t = v$, and $u = w$, these "odd" polynomials also have moments that are multiples of $(n+1)!_q$. There is a version of Theorem 1 for the "odd" polynomials that yields more statistics related to $s(\sigma)$. We do not state this combinatorial theorem here; rather, in this section we state what these "odd" polynomials are, give their moments, and state in Theorem 4 what the statistics related to $s(\sigma)$ are. Clearly the "odd" polynomials are analogues of the Laguerre polynomials $L_n^1(x)$.

We keep the parameters r, s, t, and u. This specialization gives the "even" and "odd" polynomials a combinatorial interpretation as weighted versions of injective maps (see [5], $L_n^0(x)$, and $L_n^1(x)$). This family with "4 q's" also contains other families of orthogonal polynomials of combinatorial interest that are discussed in [8].

Here we list the "odd" polynomials for the three cases in §3, and the respective moments. In each case the polynomials are monic forms of the given polynomials.

(1) little q-Jacobi $p_n(xq(1-q); q, 0; q)$, $\mu_n = q^{-n}(n+1)!_q$;
(2) little q-Jacobi $p_n(x(1-q); q, 0; q)$, $\mu_n = (n+1)!_q$;
(3) q-Laguerre $L_n^1(x(1-q); q)$, $\mu_n = q^{-(n^2+3n)/2}(n+1)!_q$.

The combinatorial theorem that results is Theorem 4 below. To describe the suitable statistic $s(\sigma)$ we shall need an auxillary statistic, $n(\sigma)$, defined as follows. For a given permutation $\sigma \in S_n$, let d be the largest element in the same run as 1, $d \geq 1$. Now partition the elements of σ into three classes: elements to the left of 1, elements in the same run as 1, and those to the right of d. Suppose that over the portion of σ to the right of d no left-to-right minimum constitutes a singleton run. Put $n(\sigma) = 0$. Otherwise, let c be the last (rightmost, smallest) singleton, which is a left-to-right minimum on the portion of σ to the right of d. Let $\text{nleft}(\sigma) := \#\{i : i <$

$\sigma^{-1}(1),\ c < \sigma(i) < d\}$. In this case put $n(\sigma) = 2(d-c) - \text{nleft}(\sigma)$. For example, $n(9 \mid 1\ 5\ 7 \mid 2\ 6 \mid 4 \mid 3\ 8) = 0$, while for $\sigma = 7\ 12 \mid 1\ 6\ 9 \mid 3 \mid 2\ 10\ 11 \mid 5 \mid 4\ 8$ we have $d = 9$, $c = 3$, $\text{nleft}(\sigma) = 1$ (from the element 7), and we get $n(\sigma) = 2(9-3) - 1 = 11$.

We also need variations lsg^* and rsg^* on the statistics lsg and rsg. These differ from the original statistics in two respects. For closers and singletons (the maxima of the runs), the run containing the element 1 is ignored in the calculation of lsg^* and rsg^*. For openers and continuators, the run containing 1 is always counted in lsg^* (if it is to the left) or in rsg^* (if it is to the right).

For example, if $\sigma = 7\ 12 \mid 1\ 6\ 9 \mid 3 \mid 2\ 10\ 11 \mid 5 \mid 4\ 8$, then $\text{lsg}^*(5) = 1$, $\text{lsg}^*(10) = 2$, $\text{lsg}^*(\sigma) = 10$, and $\text{rsg}^*(\sigma) = 8$.

THEOREM 4. *For $\sigma \in S_n$, let*

$$s(\sigma) := \text{run}(\sigma) - 1 + 2\text{lsg}^*(\sigma) + \text{rsg}^*(\sigma) + n(\sigma).$$

Then

$$\sum_{\sigma \in S_n} q^{s(\sigma)} = n!_q.$$

Theorem 4 also holds if $\text{run}(\sigma) - 1$ is replaced by $n - \text{run}(\sigma)$. Moreover, the role of closers and openers can be interchanged in Theorem 4, and there are also 16 variations, although complicated ones.

A version of Theorem 4 holds for permutations in S_{n+1} that satisfy the following condition: 1 and $n+1$ belong to the same run, and no left-to-right minimum to the right of $n+1$ constitutes a singleton run. There are $n!$ such permutations in S_{n+1}.

Finally, we remark that these specializations are the only ones we have found for which the moments factor into an analogue of $n!$. They are also the only specializations that give the three sets of polynomials that were considered. A more extensive study of the specializations of (2.1) appears in [8].

REFERENCES

[1] T. CHIHARA, *An Introduction to Orthogonal Polynomials*, Mathematics and Its Applications, Vol. 13, Gordon and Breach, New York, 1978.

[2] A. DE MEDICIS AND X. VIENNOT, *Moments des q-polynômes de Laguerre et la bijection de Foata–Zeilberger*, Stud. Appl. Math., to appear.

[3] A. ERDÉLYI, *Higher Transcendental Functions*, McGraw-Hill, New York, 1953.

[4] P. FLAJOLET, *Combinatorial aspects of continued fractions*, Discrete Math., 32 (1980), pp. 126–161.

[5] D. FOATA AND V. STREHL, *Combinatorics of Laguerre polynomials*, in Enumeration and Design, Waterloo Jubilee Conference, Academic Press, New York, 1984, pp. 123–140.

[6] G. GASPER AND M. RAHMAN, *Basic Hypergeometric Series*, in Encyclopedia of Mathematics and Its Applications, Vol. 35, Cambridge University Press, New York, 1990.

[7] D. MOAK, *The q-analogue of the Laguerre polynomials*, J. Math. Anal. Appl., 81 (1981), pp. 20–47.

[8] R. SIMION AND D. STANTON, *Octabasic Laguerre polynomials and permutation statistics*, manuscript.

[9] G. VIENNOT, *Une théorie combinatoire des polynômes orthogonaux généraux*, Lecture Notes, Université du Québec au Montréal, Canada, 1983.

[10] ———, *A combinatorial theory for general orthogonal polynomials with extensions and applications*, in Polynômes Orthogonaux et Applications, Bar-le-Duc, Springer Lecture Notes in Mathematics, Vol. 1171, Springer-Verlag, New York, 1984, pp. 139–157.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 720–732, March 1994

030

DISCRETE AND CONTINUOUS LIOUVILLE–GREEN–OLVER APPROXIMATIONS: A UNIFIED TREATMENT VIA VOLTERRA–STIELTJES INTEGRAL EQUATIONS*

RENATO SPIGLER† AND MARCO VIANELLO‡

Abstract. A unified treatment of the Liouville–Green–Olver approximation theory for linear second-order differential *and* difference equations is presented. This is based on reduction to Volterra–Stieltjes integral equations with respect to *complex* measures. The present approach embodies and improves several previous results. Moreover, error bounds are obtained for *recessive* solutions of certain difference equations, for which only qualitative results were known. The theory can be applied, for instance, to the asymptotics of certain families of orthogonal polynomials.

Key words. Volterra–Stieltjes integral equations, complex-valued measures, Liouville–Green (WKBJ) approximation, linear difference equations, linear differential equations

AMS subject classifications. 39A12, 34E20, 45D05, 28A10

1. Introduction.

The well-known and widely used Liouville–Green (or WKBJ) approximations for solutions to linear second-order *differential* equations, to which F. W. J. Olver has been able to associate precise estimates for the error terms since 1961 [10], have been recently extended to the case of *difference* equations [4], [12], [13]. Olver's analysis was based on integral equations of the Volterra type satisfied by the relevant error terms. In this paper we proceed similarly, using Volterra integral equations with respect to *complex* Lebesgue–Stieltjes measures to treat both differential equations (absolutely continuous measures) *and* difference equations (discrete measures) at the same time.

As an application, *error bounds* are obtained for *recessive* solutions to a class of linear difference equations of the second order. Only qualitative results were known for this class. Our results can be applied, for instance, to the asymptotics of certain orthogonal polynomials off their essential spectrum.

The paper is organized as follows. In §2 we prove some theorems on existence, uniqueness, and estimates for the solutions of certain types of Volterra–Stieltjes integral equations with respect to complex-valued measures. In §3 these results are related to the Liouville–Green approximation for the solutions of both differential *and* difference equations and several examples and applications are presented.

2. The main theorems.

In this section we prove some theorems yielding existence, uniqueness, and estimates for the solutions to certain linear Volterra integral equations with respect to Lebesgue–Stieltjes complex-valued measures (for generalities on complex-valued measures, we refer the reader to [16, Chap. 11]). These results will be related to the asymptotic theory of *differential* as well as *difference* equations. Throughout the paper we stipulate, for convenience, that

$$\int_x^{+\infty} f\,d\mu \equiv \int_{[x,+\infty)} f\,d\mu$$

* Received by the editors May 18, 1992; accepted for publication (in revised form) April 13, 1993. This work was supported by MURST Mathematical Analysis funds, MURST Numerical Analysis funds, the GNFM-CNR, and the Istituto Nazionale di Alta Matematica F. Severi.

† Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università di Padova, via Belzoni 7, 35131, Padova, Italy.

‡ Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, 35131, Padova, Italy.

for the general (complex) measure μ.

THEOREM 2.1. *Consider the linear integral equation*

$$\varepsilon(x) = \int_x^{+\infty} K(x,t)[\phi(x,t) + \varepsilon(t)]d\mu, \qquad x \in [a, +\infty), \tag{1}$$

of the Volterra type, where $a \in (-\infty, +\infty)$ *and* μ *is a complex (finite) measure. Suppose that for each fixed* $x \in [a, +\infty)$

(i) $K(x,\cdot)$, $\phi(x,\cdot)$ *are* μ*-measurable complex-valued functions*;

(ii) $|K(x,t)|$, $|K(x,t)\phi(x,t)| \le h(x,t)$ $|\mu|$*—almost everywhere for* $t \ge x$, *where* $h(x,\cdot) \in L^1([x,+\infty);\mu)$, *and, moreover,*

$$V(x) := \int_x^{+\infty} h(x,t)d|\mu| \tag{2}$$

is nonincreasing and $\lim_{x\to+\infty} V(x) < 1$.

Then there exists a unique solution $\varepsilon(x)$ *of* (1) *for* $x > x_1$, *where*

$$x_1 := \inf\{x : x \ge a,\ V(x) < 1\}, \tag{3}$$

and the estimate

$$|\varepsilon(x)| \le \frac{V(x)}{1 - V(x)}, \qquad x > x_1 \tag{4}$$

holds.

Proof. Consider the sequence

$$\begin{aligned} \varepsilon_0(x) &\equiv 0, \\ \varepsilon_{s+1}(x) &= \int_x^{+\infty} K(x,t)[\phi(x,t) + \varepsilon_s(t)]d\mu, \qquad s = 0,1,2,\ldots, \end{aligned} \tag{5}$$

that is well defined since it is easily proved, by induction on s, that

$$|\varepsilon_s(x)| \le C_s, \quad C_0 = 0, \quad C_s = (C_{s-1} + 1)V(a), \quad s = 1,2,3,\ldots. \tag{6}$$

Define, formally,

$$\varepsilon(x) := \sum_{s=0}^{\infty} [\varepsilon_{s+1}(x) - \varepsilon_s(x)]. \tag{7}$$

Now,

$$|\varepsilon_1(x)| \le V(x) \tag{8}$$

holds, and, assuming as an inductive hypothesis that

$$|\varepsilon_s(x) - \varepsilon_{s-1}(x)| \le [V(x)]^s, \tag{9}$$

we obtain

$$|\varepsilon_{s+1}(x) - \varepsilon_s(x)| \le \int_x^{+\infty} |K(x,t)|[V(t)]^s d|\mu| \le [V(x)]^s \int_x^{+\infty} |K(x,t)|d|\mu|$$

$$\leq [V(x)]^s \int_x^{+\infty} h(x,t)d|\mu| = [V(x)]^{s+1}. \tag{10}$$

From (8), (9) then (4) follows, where x_1 is given in (3). The estimates (8) and (9) also show that the series in (7) converges uniformly with respect to x for $x \geq \xi$ and for every fixed $\xi > x_1$. Therefore,

$$\begin{aligned} \varepsilon(x) &= \varepsilon_1(x) + \sum_{s=1}^{\infty} \int_x^{+\infty} K(x,t)[\varepsilon_s(t) - \varepsilon_{s-1}(t)]d\mu \\ &= \varepsilon_1(x) + \int_x^{+\infty} K(x,t) \sum_{s=1}^{\infty} [\varepsilon_s(t) - \varepsilon_{s-1}(t)]d\mu, \end{aligned} \tag{11}$$

and hence $\varepsilon(x)$ given by (7) solves (1) for $x > x_1$. It is finally immediately proved that uniqueness holds for $x \geq \xi$, ξ being any fixed number with $\xi > x_1$. □

Remark 2.2. When $V(x_1) < 1$, i.e., "inf" can be replaced by "min" in (3), all results hold up to and include x_1. When h in (2) is *continuous* as a function of x *and* μ is *absolutely continuous*, then if $x_1 > a$, it follows that $V(x_1) = 1$ (and thus, if $V(x_1) < 1$, it is necessary that $x_1 = a$), $V(x)$ being a *continuous* function in this case.

When μ is absolutely continuous, a variant of Theorem 2.1 that has some interest for differential equations can be proved.

THEOREM 2.3. *Suppose that in Theorem* 2.1 μ *is absolutely continuous with density* $p(x)$ *and condition* (ii) *is replaced by*

(ii′) $|K(x,t)|$, $|K(x,t)\phi(x,t)| \leq M_0(t)N_0(x)$ *almost everywhere for* $t \geq x$, *where* $M_0 \in L^1((a,+\infty);\mu)$, *and* N_0 *is* nondecreasing *for* $x \geq a$.

Then, if we set

$$U_0(x) := \int_x^{+\infty} M_0(t)|p(t)|dt, \tag{12}$$

there exists a unique bounded solution $\varepsilon(x)$ *of* (1) *for* $x \geq a$, *with*

$$|\varepsilon(x)| \leq \exp\{N_0(x)U_0(x)\} - 1. \tag{13}$$

Proof. The proof follows the lines of the previous one, the only difference being that (9) and (10) are replaced with

$$|\varepsilon_s(x) - \varepsilon_{s-1}(x)| \leq \frac{[N_0(x)U_0(x)]^s}{s!}, \qquad s = 1,2,\ldots, \tag{14}$$

and

$$\begin{aligned} |\varepsilon_{s+1}(x) - \varepsilon_s(x)| &\leq N_0(x) \int_x^{+\infty} M_0(t) \frac{[N_0(t)U_0(t)]^s}{s!} |p(t)|dt \\ &\leq -[N_0(x)]^{s+1} \int_x^{+\infty} \frac{[U_0(t)]^s}{s!} U_0'(t)dt = \frac{[N_0(x)U_0(x)]^{s+1}}{(s+1)!}, \end{aligned} \tag{15}$$

where (12) has been used. Therefore, (7) leads to the exponential estimate in (13). Moreover, there is a *unique* solution to (1). In fact, suppose that $\varepsilon(x)$ and $\eta(x)$ are

two such solutions. Then, if we define $\delta(x) := \varepsilon(x) - \eta(x)$, there is a constant C such that $|\delta(x)| \le C$ for $x \ge a$, and $\delta(x)$ can be estimated as

$$
\begin{aligned}
|\delta(x)| &\le \int_x^{+\infty} |K(x,t)|\,|\delta(t)|\,|p(t)|dt \\
&\le CN_0(x)\int_x^{+\infty} M_0(t)|p(t)|dt = CN_0(x)U_0(x),
\end{aligned} \tag{16}
$$

and again, by successive resubstitutions (see [11, p. 141]),

$$
|\delta(x)| \le C\frac{[N_0(x)U_0(x)]^s}{s!} \quad \text{for } s \ge 1 \text{ and } x \ge a, \tag{17}
$$

and hence $\delta(x) \equiv 0$ for $x \ge a$. □

3. Applications to differential and difference equations. As is well known, the measure μ appearing in Theorem 2.1, being finite and complex-valued on $\mathbf{R}$ (and thus Lebesgue–Stieltjes), can be represented as the sum of *three* (complex) measures, the first being *absolutely continuous*, the second being *discrete*, and the third being *singular*. When μ reduces merely to the first one, one is led to integral equations that can be related to linear *differential equations*; when μ reduces to the second one, one is led to integral equations that can be related to linear *difference equations*. In this section we analyze these two special cases in detail.

3.1. Differential equations. In this case some regularity results are needed for the solution $\varepsilon(x)$. These can be estabilished by requiring some additional properties on $K(x,t)$, $\phi(x,t)$ and on the density of μ.

THEOREM 3.1. *Suppose that* (1) *is given, the measure* μ *being absolutely continuous with density* $p(x)$. *Moreover, assume that for* $r \in \mathbf{N}$

(i) K, $\phi \in C^r(T)$, T *being the sector* $x \in [a,+\infty)$, $t \ge x$;

(ii) $p \in C^{r-1}([a,+\infty))$ *when* $r > 0$;

(iii) *there exist* $2r+2$ *nonnegative functions* $M_j(t)$, $N_j(x)$, $j = 0,1,\dots,r$, *with* $M_j \in L^1((a,+\infty);\mu)$, $N_j \in C^0([a,+\infty))$, *and*

$$
\left|\frac{\partial^j K}{\partial x^j}\right|, \ \left|\frac{\partial^j (K\phi)}{\partial x^j}\right| \le M_j(t)N_j(x), \quad j = 0,1,\dots,r, \quad \text{a.e. for } t \ge x. \tag{18}
$$

Then the solution $\varepsilon(x)$ *to equation* (1) *is of class* $C^r([a,+\infty))$.

Proof. We notice first that (18) with $j = 0$ ensures that there is *a unique bounded* solution to (1) by Theorem 2.3.

It is then easy to derive from (1) the representation

$$
\begin{aligned}
\varepsilon^{(r)}(x) = &\int_x^{+\infty} \left[\frac{\partial^r (K\phi)}{\partial x^r}(x,t) + \frac{\partial^r K}{\partial x^r}(x,t)\varepsilon(t)\right] p(t)dt \\
&- \sum_{j=0}^{r-1} \frac{d^j}{dx^j}\left\{\left[\frac{\partial^{r-1-j}(K\phi)}{\partial x^{r-1-j}}(x,x) + \varepsilon(x)\frac{\partial^{r-1-j}K}{\partial x^{r-1-j}}(x,x)\right] p(x)\right\}
\end{aligned} \tag{19}
$$

for the rth derivative of $\varepsilon(x)$. The proof of this can be based essentially on legitimate differentiations under the sign of integral, in view of the dominated convergence theorem, which also shows that $\varepsilon \in C^r$. Details are left to the reader. □

Remark 3.2. Observe that Theorem 3.1 for $r = 0$ ensures uniqueness of solutions to (1) *without* requiring their boundedness a priori. In fact, all solutions would be continuous and thus locally bounded while uniqueness holds in a neighborhood of $+\infty$ since by (12) the integral operator in (1) is a contraction for x sufficiently large.

The following variant of Theorem 3.1 may be useful.

THEOREM 3.3. *Theorem* 3.1 *holds if assumption* (ii) *is replaced by*

$$\text{(ii')} \qquad \frac{\partial^j K}{\partial x^j}(x,x) \equiv 0 \quad \textit{for } j = 1,2,\dots,r,$$

when $r > 0$. *Moreover, if* $N_0(x)$ *is also nondecreasing for* $x \geq a$, *the following estimates hold for the derivatives of* $\varepsilon(x)$:

$$|\varepsilon^{(j)}(x)| \leq N_j(x)U_j(x)\exp\{N_0(x)U_0(x)\}, \qquad x \geq a, \tag{20}$$

$$U_j(x) := \int_x^{+\infty} M_j(t)|p(t)|dt, \qquad j = 1,2,\dots,r. \tag{21}$$

Proof. The first part of the theorem can be proved immediately by observing that (19) still holds true with all terms in the sum equal to zero. As for the second part, similarly to (19), from (5) the following is obtained:

$$\begin{aligned} &\varepsilon_0(x) \equiv 0, \\ &\varepsilon_{s+1}^{(j)}(x) = \int_x^{+\infty} \left[\frac{\partial^j (K\phi)}{\partial x^j}(x,t) + \frac{\partial^j K}{\partial x^j}(x,t)\varepsilon_s(t)\right] p(t)dt, \qquad s = 0,1,2,\dots. \end{aligned} \tag{22}$$

Therefore, by using the monotonicity of $U_0(x)$ (and of $N_0(x)$) it can be proved by induction on s and (14) that

$$|\varepsilon_{s+1}^{(j)}(x) - \varepsilon_s^{(j)}(x)| \leq N_j(x)U_j(x)\frac{[N_0(x)U_0(x)]^s}{s!} \quad \text{for } s = 0,1,2,\dots, \tag{23}$$

and thus (20) follows. In fact, exchanging series and derivatives in (7) is permissible because (23) shows the *uniform* convergence of $\sum_{s=0}^{\infty}[\varepsilon_{s+1}^{(j)}(x) - \varepsilon_s^{(j)}(x)]$ for each j since $N_0(x)U_0(x) \leq N_0(a)U_0(a)$ for $x \geq a$. □

Moreover, we have the following theorem.

THEOREM 3.4. *Under all hypotheses of Theorem* 3.1, *with condition* (ii') *replacing* (ii), *and the additional estimates*

$$\left|\frac{\partial^j K}{\partial x^j}\right|, \left|\frac{\partial^j (K\phi)}{\partial x^j}\right| \leq P_j(x,t), \qquad j = 1,2,\dots,r, \quad \textit{a.e. for } t \geq x, \tag{24}$$

where $P_j(x,\cdot) \in L^1((x,+\infty);\mu)$ *and*

$$W_j(x) := \int_x^{+\infty} P_j(x,t)|p(t)|dt \tag{25}$$

is nonincreasing, then

(I) *if the assumptions of Theorem* 2.1 *are also satisfied, we obtain*

$$|\varepsilon^{(j)}(x)| \leq \frac{W_j(x)}{1 - V(x)} \quad \textit{for } x > x_1, \tag{26}$$

$V(x)$ *being defined in* (2) *and* x_1 *in* (3), *whereas,*

(II) *if* $N_0(x)$ (*see Theorem* 3.1) *is nonincreasing, we obtain*

$$|\varepsilon^{(j)}(x)| \le W_j(x)\exp\{V(x)\} \quad \textit{for } x \ge a. \tag{27}$$

Proof. The proof is similar to that of Theorem 3.3, with inequality (23) being replaced by

$$|\varepsilon_{s+1}^{(j)}(x) - \varepsilon_s^{(j)}(x)| \le W_j(x)[V(x)]^s, \qquad s = 0,1,2,\dots \tag{28}$$

in case (I) and by

$$|\varepsilon_{s+1}^{(j)}(x) - \varepsilon_s^{(j)}(x)| \le W_j(x)\frac{[V(x)]^s}{s!}, \qquad s = 0,1,2,\dots \tag{29}$$

in case (II). □

Theorems 3.1, 3.3, and 3.4 turn out to be useful in the framework of linear differential equations. Here we consider linear second-order differential equations like

$$y'' + [f(x) + g(x)]y = 0, \qquad x \ge 1, \quad f, g \in C^0([1,+\infty)), \tag{30}$$

where f is real valued, g is complex valued, and

(a) $f(x) \ne 0$ in $[1,+\infty)$, $f \in C^2$, and

$$\mathcal{V}(x) = \int_x^{+\infty} |f^{-1/4}(f^{-1/4})'' - gf^{-1/2}|dt < \infty, \tag{31}$$

or

(b) $f(x) \equiv 0$ in $[1,+\infty)$ and

$$\mathcal{M}_k(x) = \int_x^{+\infty} t^k|g(t)|dt < \infty \quad \text{for } k = 1 \text{ or } 2. \tag{32}$$

Case (a) is the typical case considered by F. W. J. Olver in connection with the Liouville–Green approximation theory (see [11, Chap. 6]). If $f(x) > 0$ (oscillatory case), when we look for solutions to (30) of the form

$$y_j(x) = f^{-1/4}(x)\exp\left\{(-1)^j i\int^x f^{1/2}(t)dt\right\}[1 + \varepsilon_j(x)], \qquad j = 1,2, \tag{33}$$

it turns out that the error term, $\varepsilon_j(x)$, must be a C^2-solution to an integral equation like (1), with $K(x,t) = (1/2i)[1 - \exp\{(-1)^j 2i\int_x^t f^{1/2}(s)ds\}]$, $\phi(x,t) \equiv 1$, $d\mu = [f^{-1/4}(f^{-1/4})'' - gf^{-1/2}]dt$. Choosing in (ii′) of Theorem 2.3, $M_0(t) \equiv 1$, $N_0(x) \equiv 1$, one gets $U_0(x) = \mathcal{V}(x)$ and hence Olver's result [11, Thm. 2.1, Chap. 6],

$$|\varepsilon_j(x)| \le \exp\{\mathcal{V}(x)\} - 1, \tag{34}$$

holds. As for the derivatives, one can choose, in Theorem 3.3, $M_1(t) \equiv 1$, $N_1(x) = f^{1/2}(x)$, and thus $U_1(x) = U_0(x)$ and $|\varepsilon_j'(x)| \le f^{1/2}(x)U_0(x)\exp\{U_0(x)\}$. The latter implies that $f^{-1/2}(x)\varepsilon_j'(x) = O(U_0(x))$, as in Olver's theorem [11, Thm. 2.1, Chap. 6].

Alternatively, one could apply Theorem 2.1 (and Theorem 3.4). In this case the estimates hold, in general, only for $x > x_1$ (x_1 being defined in (3)). However, $\varepsilon_j(x) = O(V(x))$ and $V(x) \le U_0(x)$ when $h = |K| = |K\phi| = |\sin(\int_x^t f^{1/2}(s)ds)|$ is

chosen. Moreover, the geometric estimate (4) *may* provide better estimates for fixed x, with respect to certain parameters. This fact can be related to the *double asymptotic* nature of the Liouville–Green approximations with respect to both the independent variable and the parameters entering $f+g$. Here is a simple example, for the purpose of illustration, in which all calculations can be carried out explicitly. Let $f(x) \equiv 1$, $g(x) = x^{-\gamma}$, $\gamma > 2$ on $x \geq 1$. Then $|K| = |K\phi| = |\sin(t-x)|$ and $|p| = x^{-\gamma}$. Choosing $h(x,t) = t-x$ and $M_0(t) \equiv 1$, $N_0(x) \equiv 1$, we get $V(x) = x^{2-\gamma}/(\gamma-1)(\gamma-2)$ and $U_0(x) = x^{1-\gamma}/(\gamma-1)$. Now, the fact that the geometric estimate (4) yields an error of order of $O(\gamma^{-2})$ for fixed x while the exponential estimate (13) gives an order of $O(\gamma^{-1})$ suggests that the former may be better. Indeed, if we compare the two estimates for x such that, e.g., $V(x) \leq \frac{1}{2}$, clearly $V/(1-V) \leq 2V < U_0 < \exp\{U_0\} - 1$ as long as $[2/(\gamma-1)(\gamma-2)]^{1/(\gamma-2)} \leq x < \gamma/2 - 1$, and thus the geometric estimate performs better. Case (a) with $f(x) < 0$ (nonoscillatory case) can be handled in a similar way, and Olver's results are recovered again.

Case (b) has been studied, for instance, in [6], [13]. If (32) holds for $k = 1$, when one looks for a (recessive) solution of the form

$$y_1(x) = 1 + \varepsilon_1(x), \tag{35}$$

$\varepsilon_1(x)$ turns out to be a C^2-solution of (1) with $K(x,t) = t-x$, $\phi(x,t) \equiv 1$, $d\mu = g\,dt$. In this case it is known that a second (dominant) solution $y_2(x) \sim x$ as $x \to +\infty$ exists. If (18) holds for $k = 2$, in addition, when one looks for a (dominant) solution like

$$y_2(x) = x + \varepsilon_2(x), \tag{36}$$

one finds that $\varepsilon_2(x)$ must be a C^2-solution of (1) with $K(x,t) = t-x$, $\phi(x,t) = t$, and $d\mu$ as before. As in case (a), choosing in (ii′) of Theorem 2.3 $M_0(t) = t^k$, $N_0(x) \equiv 1$ (see [6]), we obtain $U_0(x) = \mathcal{M}_k(x)$. Concerning the derivatives, Theorem 3.3 can be applied with $M_1(t) = t^{k-1}$, $N_1(x) \equiv 1$.

It is also possible to apply Theorems 2.1 and 3.4 with $h(x,t) = t^{k-1}(t-x)$, $P_1(x,t) = t^{k-1}$ (see [13]). The advantage obtainable by using the geometric versus the exponential estimates is more pronounced here. In fact, for $g(x) = x^{-\gamma}$, $\gamma > 2$, we get (for $k = 1$) $V(x) = U_0(x)/(\gamma-1)$ for all x. Therefore, comparing the two estimates, (4) and (13), for x such that, e.g., $V(x) \leq \frac{1}{2}$, we find that the geometric estimate is certainly sharper for $\gamma > 3$. Such an estimate is of the order of $O(\gamma^{-2})$, whereas the exponential one is of the order of $O(\gamma^{-1})$, as in case (a), but this holds now without any further limitation on x.

3.2. Difference equations. When the measure μ in Theorem 2.1 is merely discrete, with discrete support $\mathbf{Z}_\nu := \{n \in \mathbf{Z} : n \geq \nu\}$, ν being a given integer, the integral equation (1) plays a role in studying linear difference equations. Hereafter we shall be concerned with the asymptotic solution of second-order difference equations like

$$\Delta^2 y_n + (\alpha + g_n) y_n = 0, \qquad n \in \mathbf{Z}_\nu, \tag{37}$$

$\Delta y_n := y_{n+1} - y_n$, where we consider the following.

(A) $\alpha > 0$ and $\sum_{n=\nu}^{\infty} |g_n| < \infty$;

(B) $\alpha = 0$ and $\sum_{n=\nu}^{\infty} n^k |g_n| < \infty$, $k = 1$ or 2;

(C) $\alpha < 0$, $\alpha \neq -1$, and $\sum_{n=\nu}^{\infty} |g_n| < \infty$.

It is worth noting that the importance of equations of the type

$$\Delta^2 y_n + q_n y_n = 0 \tag{38}$$

stems from the fact that they represent a kind of canonical form for all difference equations like

$$Y_{n+2} + A_n Y_{n+1} + B_n Y_n = 0. \tag{39}$$

Indeed, the transformation

$$Y_n = \alpha_n y_n, \quad \alpha_n := \alpha_{\nu+1} \prod_{k=\nu}^{n-2} \left(-\frac{A_k}{2}\right), \quad n \geq \nu + 2, \tag{40}$$

α_ν and $\alpha_{\nu+1} \neq 0$ being arbitrary constants, takes (39) into (38) with a suitable q_n, provided that $A_k \neq 0$ (at least for k sufficiently large); see [12], [13].

Case (A), in which all real solutions turn out to be *oscillatory*, was studied in some detail in [12]. If one looks for a solution of the form $y_n = \lambda^n(1 + \varepsilon_n)$, λ being one of the roots of the characteristic equation of (37) with $g_n \equiv 0$, $\lambda = 1 \pm i\sqrt{\alpha}$, the "integral equation"

$$\varepsilon_n = \frac{1}{2\lambda(\lambda - 1)} \sum_{k=n}^{\infty} \left(1 - (\lambda/\overline{\lambda})^{k-n+1}\right) g_k(1 + \varepsilon_k) \tag{41}$$

for the error term ε_n is obtained. In [12] it was proved *directly* that such an equation has a solution estimated as

$$|\varepsilon_n| \leq \frac{V_n}{1 - V_n}, \tag{42}$$

$$V_n := \frac{1}{[\alpha(\alpha + 1)]^{1/2}} \sum_{k=n}^{\infty} |g_k|. \tag{43}$$

Both the existence of the solution and the estimate (42), (43) hold for $n \geq n_1 := \min\{n \in \mathbf{Z}_\nu : V_n < 1\}$. This result represents an extension to the discrete domain of the Liouville–Green–Olver approximation theorem for oscillatory-type differential equations (see [11, Thm. 2.2, p. 196]).

Note the formal analogy of (37) with (30) when we take $f(x) \equiv \text{const.} > 0$ (case (a), §3.1). In fact, the unified treatment presented in this paper leads to (42), (43) when we choose in Theorem 2.1 $\phi(x,t) \equiv 1$, $K(x,t) = (1/2\lambda(\lambda - 1))(1 - (\lambda/\overline{\lambda})^{t-x+1})$, and $\mu = \sum_{k=\nu}^{\infty} g_k \delta_k$, where δ_k is the Dirac measure centered in $t = k$. The function $h(x,t)$ estimating K, $K\phi$ in (ii) of Theorem 2.1 can be chosen equal to the constant $1/|\lambda(\lambda - 1)| = [\alpha(\alpha + 1)]^{-1/2}$. It is easily seen that $V(x)$ is left-continuous and piecewise constant and that $x_1 = n_1$; also, $a = \nu$, $V_n = V(n)$. Finally, observe that the solution to (1) restricted to the integers $n \geq n_1$ also solves (41) and that the estimate (42), (43) holds.

Case (B) represents the discrete analogue of case (b) of §3.1 and was studied in [13]. In [13] a unified approach was followed; it was based, however, on a special *subclass* of integral equations like those in (1). Indeed, when one looks for solutions to (37) with $\alpha = 0$ of the form $y_n = n^{k-1} + \varepsilon_n$, with $k = 1$ or 2, an "integral equation"

of type (1) is obtained, with $K(x,t) = x - t + 1$, $\phi(x,t) = t^{k-1}$, and $\mu = \sum_{k=\nu}^{\infty} g_k \delta_k$ ($\nu = 1$). If one takes $h(x,t) = t^{k-1}(t - x + 1)$, estimates are obtained for the error terms corresponding to the *recessive* and the *dominant* solution ($k = 1$ *or* 2 also refers to the finiteness of the first *or* the second moment of $|g_n|$: When $\sum^{\infty} n^2|g_n| = +\infty$ but $\sum^{\infty} n|g_n| < \infty$ the estimate for the recessive solution still holds true).

A Liouville–Green–Olver approximation result for case (C) seems to be missing so far. Note that, unlike in the previous cases, *qualitative* asymptotics for the solutions to (37) with $\alpha < 0$ could be obtained here by Poincaré's or Perron's theorems [9]. Deriving *precise error bounds*, however, is our goal, in the spirit of Olver's approach. This problem represents a discrete analogue of the differential case with solutions of the exponential type in [11, Thm. 2.1, p. 193]. We state our result as a theorem.

THEOREM 3.5. *Suppose that* (37) *is given with* $\alpha = -\beta < 0$, $\beta \neq 1$, *and* $\sum^{\infty} |g_n| < \infty$ (*see case* (*C*)). *Then there exist* $n_1 \in \mathbf{Z}_\nu$ *and two linearly independent solutions to* (37), $y_n^{\pm}$, *such that*

$$y_n^- = (\lambda_-)^n[1 + \varepsilon_n], \quad n \geq n_1; \qquad y_n^+ \sim (\lambda_+)^n, \quad n \to \infty, \tag{44}$$

where $\lambda_\pm = 1 \pm \sqrt{\beta}$ *are the roots of the characteristic equation associated to* (37) *with* $g_n \equiv 0$. *For the error term* ε_n *the estimate*

$$|\varepsilon_n| \leq \frac{V_n}{1 - V_n}, \quad V_n := \frac{1}{2\sqrt{\beta}(\sqrt{\beta} - 1)} \sum_{k=n}^{\infty} |g_k|, \quad n \geq n_1 \tag{45}$$

holds and

$$n_1 = \min\{n \in \mathbf{Z}_\nu : V_n < 1\}. \tag{46}$$

Moreover, when g_n *is real,* $y_n^{\pm}$ *are real.*

Remark 3.6. In view of (45), y_n^- and y_n^+ are *recessive* and *dominant* solutions, respectively. Note that, unlike the corresponding case for differential equations, while $y_n^+ \sim (1 + \sqrt{\beta})^n$ grows exponentially, $y_n^- \sim (1 - \sqrt{\beta})^n$ and thus decays exponentially when $0 < \beta < 1$, but it exhibits *oscillations* exponentially growing (when $\beta > 4$) *or* exponentially decreasing (when $1 < \beta < 4$); when $\beta = 4$, $y_n^- \sim (-1)^n$. The case $\beta = 1$ is pathological in that the unperturbed equation (see (37) with $g_n \equiv 0$) degenerates, having the lowest-order coefficient vanishing (see [8] and [9]), and *cannot* be treated by the present approach.

Remark 3.7. The qualitative behavior $y_n^{\pm} \sim (\lambda_\pm)^n$, $n \to \infty$, *cannot* be obtained directly from Poincaré's or Perron's theorems, despite the fact that such theorems can be applied, since $\lambda_+ \neq \lambda_-$ [8, §5.3, p. 221].

Remark 3.8. Observe even here the *double asymptotic nature* of the Liouville–Green–Olver approximations with respect to both n and β (as $n \to \infty$, and as $\beta \to +\infty$); see (45). In particular, $V_n = O(\beta^{-1})$, whereas the corresponding quantity for the analogous differential equation, $y'' + (-\beta + g(x))y = 0$, $g \in L^1(1, +\infty)$, is $O(\beta^{-1/2})$; see [11] and [12].

Proof of Theorem 3.5. Looking for a solution of the form $y_n^- = (\lambda_-)^n[1 + \varepsilon_n]$, one is led to the difference equation

$$(\lambda_-)^2 \Delta^2 \varepsilon_n + 2\lambda_-(\lambda_- - 1)\Delta\varepsilon_n + g_n(1 + \varepsilon_n) = 0 \tag{47}$$

for the error term. It is then easily proved that any solution of the linear discrete Volterra-type "integral" equation

$$\varepsilon_n = \sum_{k=n}^{\infty} (1 - \rho^{k-n+1}) g_k \sigma(1 + \varepsilon_k), \tag{48}$$

$$\rho = \frac{\lambda_-}{\lambda_+} = \frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}, \qquad \sigma = \frac{1}{2\lambda_-(\lambda_- - 1)} = \frac{1}{2\sqrt{\beta}(\sqrt{\beta}-1)}, \tag{49}$$

also solves (47). The easy but lengthy verification is left to the reader; see [12]. At this point, consider equation (1), with $\phi(x,t) \equiv 1$, $K(x,t) = \sigma(1-\rho^{t-x+1})$, $x \geq a = \nu$, $\mu = \sum_{k=\nu}^{\infty} g_k \delta_k$, and choose $h(x,t) \equiv |\sigma|$, since $|\rho| < 1$. Then we use Theorem 2.1 to obtain (45) since it is clear that the solution to (1), restricted to the integers $\geq n_1$, also solves (48) and $V_n = V(n)$. Note that uniqueness of solutions to (48) does *not* follow immediately from Theorem 2.1 but can be proved directly for $n \geq n_1$.

As for the second (dominant) solution, we consider a solution of the form

$$z_n := y_n^- \sum_{k=n^*}^{n-1} \frac{C_k}{y_k^- y_{k+1}^-} \tag{50}$$

(see [8, §3.5, Thm. 3.9, p. 94]), n^* being the smallest integer $\geq n_1$ such that $y_n^- \neq 0$ for $n \geq n^*$, and C_k denoting the *Casoratian* of y_k^- and z_k. The latter is [8, §3.5]

$$C_k = C_{n^*} \prod_{j=n^*}^{k-1} (1 - \beta + g_j), \tag{51}$$

where C_{n^*} is a nonzero constant that we shall choose. Now,

$$\frac{z_n}{(\lambda_+)^n} \sim \left(\frac{\lambda_+}{\lambda_-}\right)^{-n} \sum_{k=n^*}^{n-1} \frac{C_k}{(\lambda_-)^{2k+1}[1+o(1)]}, \qquad n \to \infty. \tag{52}$$

When $0 < \beta < 1$, using *Cesaro's* theorem, we get

$$\frac{z_n}{(\lambda_+)^n} \sim \frac{C_n/(\lambda_-)^{2n+1}}{(\lambda_+/\lambda_-)^{n+1} - (\lambda_+/\lambda_-)^n} = \frac{(1-\beta)^{-n^*}}{2\sqrt{\beta}} C_{n^*} \prod_{j=n^*}^{n-1} \left(1 + \frac{g_j}{1-\beta}\right), \quad n \to \infty. \tag{53}$$

In fact, $(\lambda_+/\lambda_-)^n \uparrow +\infty$ as $n \to \infty$, and the product in (53) converges in view of the fact that $\sum^{\infty} |g_j| < \infty$. When $\beta > 1$, the same condition holds true, although the classical Cesaro's theorem cannot be invoked. Such a generalized version of Cesaro's theorem is contained in [14]. Finally, choosing the constant C_{n^*} in such a way that the limit of the right-hand side of (53) is equal to 1, one obtains $z_n = y_n^+$. □

Remark 3.9. When one looks for a dominant solution of the form $y_n^+ = (\lambda_+)^n(1+\eta_n)$, an error equation for η_n similar to (48) is obtained, with $\rho = \lambda_+/\lambda_-$, $\sigma = 1/2\lambda_+(\lambda_+ - 1)$. However, $|\rho| > 1$ prevents proving, in general, the existence itself of such a solution, as was done for ε_n. A noteworthy case, however, occurs when $g_n = \rho^{-n}u_n$, with $\sum^{\infty} |u_n| < \infty$, e.g., $g_n = c^{-n}$, with $|c| > |\rho|$. Here Theorem 2.1 can be applied with $\phi(x,t) \equiv 1$, $K(x,t) = \rho^{-t}(1-\rho^{t-x+1})$, $\mu = \sum_{k=\nu}^{\infty} u_k \delta_k$, and $h(x,t) \equiv 1$ for $x \geq 1$. The estimate

$$|\eta_n| \leq \frac{W_n}{1-W_n}, \qquad n \geq n_2 = \min\{n \in \mathbf{Z}_\nu : W_n < 1\},$$
$$W_n := \frac{1}{2\sqrt{\beta}|\sqrt{\beta}-1|} \sum_{k=n}^{\infty} |u_k| \tag{54}$$

is then obtained.

It may be useful, in closing, to reformulate the hypotheses of Theorem 3.5 in terms of the coefficients of (39). Since in (38) one obtains

$$q_n = -1 + \frac{4B_n}{A_n A_{n-1}} \tag{55}$$

for $n \geq \nu + 1$ (see [12] and [13]), they become

$$\lim_{n\to\infty} \frac{B_n}{A_n A_{n-1}} =: L, \qquad \text{with } L < \frac{1}{4}, \quad L \neq 0, \tag{56}$$

$$\sum_{k=\nu+1}^{\infty} \left| \frac{B_n}{A_n A_{n-1}} - L \right| < \infty. \tag{57}$$

We conclude with two simple applications.

Example 3.10 (perturbed Fibonacci equation). Consider

$$f_{n+2} - (1+\sigma_n) f_{n+1} - (1+\tau_n) f_n = 0, \qquad n \geq 0, \tag{58}$$

where

$$\sum_{n=0}^{\infty} (|\sigma_n| + |\tau_n|) < \infty, \tag{59}$$

and $\sigma_n \neq -1$ for $n \geq 0$. The latter restriction can be removed by confining ourselves to $n \geq \nu$ with ν sufficiently large. Then, setting $f_n = \alpha_n y_n$, with

$$\alpha_n = \prod_{k=0}^{n-2} \left(\frac{1+\sigma_k}{2} \right) \tag{60}$$

(cf. (40)) leads to (38) with

$$q_n = -1 - 4 \frac{1+\tau_n}{(1+\sigma_n)(1+\sigma_{n-1})}. \tag{61}$$

Now, $q_n \to -5$ as $n \to \infty$, and then (56) and (57) are satisfied with $L = -1$ and owing to (59). Since $\beta = 5$,

$$g_n = 4 \frac{(\sigma_n + \sigma_{n-1} + \sigma_n \sigma_{n-1} + \tau_n)}{(1+\sigma_n)(1+\sigma_{n-1})} \tag{62}$$

and $\lambda_\pm = 1 \pm \sqrt{5}$, we obtain a *recessive* solution to (58) like

$$f_n^- = \left(\frac{1-\sqrt{5}}{2} \right)^n \left(\prod_{k=0}^{n-2} (1+\sigma_k) \right) (1+\varepsilon_n) \quad \text{for } n \geq n_1, \tag{63}$$

with

$$|\varepsilon_n| \leq \frac{V_n}{1-V_n}, \qquad V_n := \frac{2}{\sqrt{5}(\sqrt{5}-1)} \sum_{k=n}^{\infty} \left| \frac{\sigma_n + \sigma_{n-1} + \sigma_n \sigma_{n-1} + \tau_n}{(1+\sigma_n)(1+\sigma_{n-1})} \right| \quad \text{for } n \geq n_1, \tag{64}$$

with n_1 defined in (46) (cf. Theorem 3.5). Moreover, a *dominant* solution f_n^+ exists, with

$$f_n^+ \sim \left(\frac{1+\sqrt{5}}{2}\right)^n \prod_{k=0}^{\infty}(1+\sigma_k), \qquad n \to \infty. \tag{65}$$

Note that the convergence of the infinite product in (65) is guaranteed by (59).

Example 3.11 (orthogonal polynomials). A field in which asymptotic representations of solutions to three-term recurrent equations is important is that of orthogonal polynomials. In [5] an application is made of the discrete WKB theory developed in [4] to a class of orthogonal polynomials obeying a recurrence with regularly and slowly varying coefficients. Here we show that Theorem 3.5 can be applied to obtain qualitative asymptotics for a well-known class of orthogonal polynomials, on the real line off their essential spectrum. Consider, in fact, the linear recurrence

$$P_{n+2}(x) - (x-\gamma_n)P_{n+1}(x) + \delta_n P_n(x) = 0, \tag{66}$$

with γ_n real, $\delta_n > 0$, $\gamma_n \to \gamma$ and $\delta_n \to \delta$, γ and δ both finite, which defines along with the initial conditions $P_{-1}(x) \equiv 0$, $P_0(x) \equiv 1$, a family of orthogonal polynomials having as essential spectrum the interval $[\gamma - 2\sqrt{\delta}, \gamma + 2\sqrt{\delta}]$; see, e.g., [2]. In this case we obtain from (56)

$$L = \frac{\delta}{(x-\gamma)^2}, \tag{67}$$

and thus $L < 1/4$ for x *off* the essential spectrum. Note then that $x \neq \gamma$. In view of the fact that $\gamma_n \to \gamma$ we see that also $x - \gamma_n \neq 0$ for all n sufficiently large ($n \geq \nu(x)$). This ensures that the transformation in (40),

$$\alpha_n(x) = \prod_{k=\nu(x)}^{n-2}\left(\frac{x-\gamma_k}{2}\right), \tag{68}$$

is applicable and the previous theory can be used. Notice also that a *uniform* lower bound for $\nu(x)$ can be found for x off the essential spectrum. Condition (47) becomes

$$\sum_{n=\nu(x)+1}^{\infty}\left|\frac{\delta_n}{(x-\gamma_n)(x-\gamma_{n-1})} - \frac{\delta}{(x-\gamma)^2}\right| < \infty. \tag{69}$$

Observe that the assumption $\sum^{\infty}(|\gamma_n - \gamma| + |\delta_n - \delta|) < \infty$, which, incidentally, ensures the orthogonality measure to be absolutely continuous in the essential spectrum, *implies* (69) and appears, e.g., in [7] as a key condition for the asymptotic analysis of the linear recurrence (66). Such a condition also appears, for instance, in [15]. Apart from having a weaker condition in (69), asymptotic results obtainable through Theorem 3.5 off the essential spectrum *coincide* with those reported in [7], where completely different techniques were adopted. Our theory, however, yields a precise *error estimate* for a *recessive* solution. This occurrence may be useful in connection to certain well-known numerical algorithms (see [3]).

REFERENCES

[1] F. V. ATKINSON, *Discrete and Continuous Boundary Value Problems*, Academic Press, New York, 1964.

[2] T. S. CHIHARA, *The three term recurrence relation and spectral properties of orthogonal polynomials*, in Orthogonal Polynomials: Theory and Practice, P. G. Nevai, ed., NATO ASI Series, Kluwer, Dordrecht, The Netherlands, 1990, pp. 99–114.

[3] W. GAUTSCHI, *Minimal solutions of three-term recurrence relations and orthogonal polynomials*, Math. Comp., 36 (1981), pp. 547–554.

[4] J. S. GERONIMO AND D. T. SMITH, WKB (*Liouville–Green*) *analysis of second order difference equations and applications*, J. Approx. Theory, 69 (1992), pp. 269–301.

[5] J. S. GERONIMO, D. SMITH, AND W. VAN ASSCHE, *Strong asymptotics for orthogonal polynomials with regularly and slowly varying recurrence coefficients*, J. Approx. Theory, 72 (1993), pp. 141–158.

[6] E. HILLE, *Non-oscillation theorems*, Trans. Amer. Math. Soc., 64 (1948), pp. 234–252.

[7] M. E. H. ISMAIL, D. R. MASSON, AND E. B. SAFF, *A minimal solution approach to polynomial asymptotics*, in Orthogonal Polynomials and Their Applications, C. Brezinski, L. Gori, and A. Ronveaux, eds., IMACS, Baltzer, Basel, Switzerland, 1991, pp. 299–303.

[8] W. G. KELLEY AND A. C. PETERSON, *Difference Equations. An Introduction with Applications*, Academic Press, San Diego, CA, 1991.

[9] V. LAKSHMIKANTHAM AND D. TRIGIANTE, *Theory of Difference Equations: Numerical Methods and Applications*, Academic Press, San Diego, CA, 1988.

[10] F. W. J. OLVER, *Error bounds for the Liouville–Green (or* WKB*) approximation*, Proc. Cambridge Philos. Soc., 57 (1961), pp. 790–810.

[11] ———, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[12] R. SPIGLER AND M. VIANELLO, *Liouville–Green approximations for a class of linear oscillatory difference equations of the second order*, J. Comput. Appl. Math., 41 (1992), pp. 105–116.

[13] ———, WKBJ*-type approximation for finite moments perturbations of the differential equation* $y'' = 0$ *and the analogous difference equation*, J. Math. Anal. Appl., 169 (1992), pp. 437–452.

[14] ———, *Cesaro's theorems for complex sequences*, J. Math. Anal. Appl., 179 (1993), to appear.

[15] W. VAN ASSCHE, *Asymptotics for orthogonal polynomials and three-term recurrences*, in Orthogonal Polynomials: Theory and Practice, P. G. Nevai, ed., NATO ASI Series, Kluwer, Dordrecht, The Netherlands, 1990, pp. 435–462.

[16] A. C. ZAANEN, *Integration*, North-Holland, Amsterdam, 1967.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 733–748, March 1994

031

TRUNCATION ERROR FOR LIMIT PERIODIC SCHUR ALGORITHMS*

W. J. THRON†

Abstract. The Schur algorithm arises in connection with the approximation of functions bounded in the unit disk. For a limit periodic Schur algorithm $\{T_n(z,w)\}$ it is shown that the truncation error is

$$|T_n(z,0) - T(z)| < c(R')^n,$$

where c depends on R' but is independent of n, and

$$1 > R' > \left|\frac{1+\bar{\gamma}zw_1}{1+\bar{\gamma}zw_2}\right|.$$

Here w_1 and w_2 are the fixed points of the limit transformation $t(z,w) = (\gamma + zw)/(1+\bar{\gamma}zw)$, $0 < |\gamma| < 1$. In the proof use is made of the facts

$$\lim_{n\to\infty} T_n^{-1}(\infty) = w_1, \qquad \lim_{n\to\infty} T^{(n)}(w) = w_2,$$

which are proved here.

Key words. Schur algorithm, limit periodic, truncation error

AMS subject classifications. 30B70, 40A15, 40A25, 65G05

1. Introduction.

The algorithm to be studied here was introduced by Schur [8] in 1917 to investigate functions holomorphic and bounded in the unit disk. For convenience the bound is taken to be 1. We are thus led to consider the family

$$U := [f : f(z) \text{ is holomorphic and } |f(z)| \le 1 \text{ for } |z| < 1].$$

Define

$$t_n(z,w) := \frac{\gamma_n + zw}{1+\bar{\gamma}_n zw}, \qquad n \ge 0, \tag{1,1}$$

and

$$t_n^{-1}(z,u) := \frac{1}{z}\,\frac{\gamma_n - u}{\bar{\gamma}_n u - 1}. \tag{1.2}$$

In terms of the sequence $\{t_n\}$ we define the sequence $\{T_n\}$ inductively as follows:

$$T_0(z,w) := t_0(z,w), \qquad T_n(z,w) := T_{n-1}(z, t_n(z,w)), \quad n \ge 1. \tag{1.3}$$

This can be expressed informally as

$$T_n(z,w) = t_0 \circ \cdots \circ t_n(z,w).$$

In the following we will frequently not indicate the dependence of the t_k and the T_n on z. Since the t_n are linear fractional transformations in w, so is T_n. Hence it can be written as

$$T_n(z,w) := \frac{C_n(z)zw + D_n(z)}{E_n(z)zw + F_n(z)}, \tag{1.4}$$

* Received by the editors November 18, 1991; accepted for publication (in revised form) July 14, 1992. This research was supported in part by the National Science Foundation under grant DMS-9103141.

† Department of Mathematics, University of Colorado, Campus Box 395, Boulder, Colorado 80309-0395.

where the C_n, D_n, E_n, F_n are polynomials in z satisfying the recursion relationships

$$C_0 = 1, \quad D_0 = \gamma_0, \quad E_0 = \bar{\gamma}_0, \quad F_0 = 1,$$

$$\begin{pmatrix} C_n \\ E_n \end{pmatrix} = z \begin{pmatrix} C_{n-1} \\ E_{n-1} \end{pmatrix} + \bar{\gamma}_n \begin{pmatrix} D_{n-1} \\ F_{n-1} \end{pmatrix}, \qquad n \geq 1, \tag{1.5}$$

$$\begin{pmatrix} D_n \\ F_n \end{pmatrix} = \begin{pmatrix} D_{n-1} \\ F_{n-1} \end{pmatrix} + \gamma_n z \begin{pmatrix} C_{n-1} \\ E_{n-1} \end{pmatrix}, \qquad n \geq 1.$$

Schur showed that for every $f \in U$ one can determine a sequence of functions $\{f_n\}$, $f_n \in U$, and a sequence of complex numbers $\{\gamma_n(f)\}$, $|\gamma_n(f)| \leq 1$ inductively, by the rule

$$f_0 =: f, \qquad f_n(z) := t_{n-1}^{-1}(z, f_{n-1}(z)), \quad n \geq 1, \qquad \gamma_n(f) := f_n(0), \quad n \geq 0.$$

The process terminates if a $\gamma_n(f)$ satisfies $|\gamma_n(f)| = 1$. Schur then showed that for the sequence $\{T_n\}$, where $\gamma_n = \gamma_n(f)$, $n \geq 0$,

$$\lim_{n\to\infty} T_n(z, 0) = f(z) \quad \text{for } |z| < 1.$$

It is also true that for any sequence $\{\gamma_n\}$, with $|\gamma_n| < 1$, $n \geq 0$, as the only restriction, $\{T_n(z, 0)\}$ converges uniformly on compact subsets of $|z| < 1$ to a function in U; see [3, Thm. 4].

In what follows we shall assume that

$$|\gamma_n| < 1, \qquad n \geq 0.$$

For any such sequence $\{\gamma_n\}$ we call the mapping $\{\gamma_n\} \to \{T_n\}$ a *Schur algorithm*. If $\gamma_n \to \gamma$, then we speak of a *limit periodic Schur algorithm* (lpS). Frequently we use this expression also for $\{T_n\}$. Schur referred to his algorithm as "kettenbruchartig," that is, as being closely related to the continued-fraction algorithm. Thus it is not unreasonable to look for results for lpS analogous to those found for limit periodic continued fractions. Since all Schur algorithms and thus, in particular, all lpS converge, we will here be concerned with the truncation error $T_n(z, 0) - T$ for lpS.

In §§2 and 3 we derive some preliminary results. In §4 we use the concept "invariance of the cross ratio under linear fractional transformations" to obtain a formula for the truncation error $T_n(0) - T$ as well as an expression for the ratio $(T - T_n(w_2))/(T - T_n(0))$, from which it can be deduced that $\{T_n(w_2)\}$ converges faster to T than does $\{T_n(0)\}$. In these formulas the sequences $\{T_n^{-1}(\infty)\}$ and the "tail" sequence $\{T^{(n)}\}$ play important roles. The convergence of $\{T_n^{-1}(\infty)\}$ to w_1 is proved in §5. In §6 we show that the tail sequence converges to w_2. The truncation error takes the form $K(R')(R')^n$, where

$$\left| \frac{1 + \bar{\gamma} z w_1}{1 + \bar{z} w_2} \right| < R' < 1.$$

Let

$$\delta_n = \gamma_n - \gamma, \qquad d_n = \max_{k \geq n} |\delta_k|. \tag{1.6}$$

Our results here are based on the assumption $d_n \to 0$. In another paper we shall study lpS under the stronger assumption $\sum d_n < \infty$ and obtain results on "separate convergence."

Some aspects of the Schur algorithm have previously been investigated [1], [5], [6], [9]. The algorithm can also be approached by way of *Schur fractions* (see [2] and the references therein). Truncation errors have been obtained for Schur fractions and the closely related positive PC fractions [1], [6].

2. The transformation $T(z,w)$. In our study the "limit" transformation

$$t(z,w) = \frac{\gamma + zw}{1 + \bar{\gamma} zw} \tag{2.1}$$

plays an important role. Frequently, the three cases

$$\gamma = 0, \quad 0 < |\gamma| < 1, \quad |\gamma| = 1$$

have to be treated separately. For the limit transformation we have

$$t(z,w) = zw \quad \text{if } \gamma = 0, \qquad t(z,w) \equiv e^{i\theta}, \quad w \neq -e^{i\theta} \quad \text{if } \gamma = e^{i\theta}. \tag{2.2}$$

If $0 < |\gamma| < 1$, the fixed points w_i of $t(z,w)$ are given by

$$w_i = \frac{\gamma + zw_i}{1 + \bar{\gamma} zw_i}, \tag{2.3}$$

or

$$\bar{\gamma} z w_i^2 + (1 - z) w_i - \gamma = 0. \tag{2.4}$$

Hence

$$w_i(z) := \frac{z - 1 + (-1)^i \sqrt{(z-1)^2 + 4|\gamma|^2 z}}{2\bar{\gamma} z}, \qquad \operatorname{Re} \sqrt{\ } > 0. \tag{2.5}$$

The case $|\gamma| = 1$ can be subsumed under the case $0 < |\gamma| < 1$ if we understand by $w_i(z)$ the quantities given by (2.5). From (2.2) we see that $w_2 = e^{i\theta}$ is indeed the attractive fixed point of the singular transformation $t(z,w) = e^{i\theta}$. The other solution of (2.4), that is, $-e^{i\theta}/z$, is not a fixed point but the exceptional point for which $t(z,w)$ is undefined. The fixed points of $t(z,w) = zw$ are 0 and ∞, and thus that case needs to be considered separately and will be excluded in the rest of this paper.

For the two cases $0 < |\gamma| < 1$ and $|\gamma| = 1$ the identities

$$w_1 w_2 = -\frac{\gamma}{\bar{\gamma} z}, \tag{2.6a}$$

$$w_i(1 + \bar{\gamma} z w_i) = \gamma + z w_i, \tag{2.6b}$$

$$z\bar{\gamma}(w_1 + w_2) = z - 1 \tag{2.6c}$$

are valid and will be useful in what follows.

Since for $0 < |\gamma| < 1$ and $|z| < 1$ the mapping $v = t(z,w)$ maps $|w| \leq 1$ into $|v| < 1$, it follows from the fixed-point theorem that one fixed point, which we shall see is w_2, satisfies $|w_2| < 1$. From the first identity in (2.6) it follows that

$$|w_1| = \frac{1}{|w_2 z|} > 1.$$

Thus for $0 < |\gamma| < 1$, $|z| < 1$, the transformation (2.1) cannot be parabolic.

Next, we explore when the transformation is elliptic. Since for $|\gamma| = 1$, $t(z, w)$ is singular, we consider only $0 < |\gamma| < 1$. It is known (see, for example, [4, p. 52]) that in this case the transformation is elliptic if and only if

$$\left|\frac{1+\bar{\gamma}zw_1}{1+\bar{\gamma}zw_2}\right| = 1, \qquad w_1 \neq w_2. \tag{2.7}$$

We set

$$Q(\gamma, z) := \left|\frac{1+\bar{\gamma}zw_1}{1+\bar{\gamma}zw_2}\right| \quad \text{if } z \neq 0, \quad Q(\gamma, 0) := 0.$$

One easily derives

$$Q(\gamma, z) = \left|\frac{1+z-\sqrt{(1+z)^2-(1-|\gamma|^2)4z}}{1+z+\sqrt{(1+z)^2-(1-|\gamma|^2)4z}}\right|.$$

Now either $z = -1$, in which case $Q(\gamma, -1) = 1$ and hence $t(-1, w)$ is elliptic, or $z \neq -1$ and one can write

$$Q(\gamma, z) = \left|\frac{1-\sqrt{1-(1-|\gamma|^2)\frac{4z}{(1+z)^2}}}{1+\sqrt{1-(1-|\gamma|^2)\frac{4z}{(1+z)^2}}}\right|.$$

The condition for $Q(\gamma, z)$ to equal 1 is then easily seen to be

$$\operatorname{Im} \frac{4z}{(1+z)^2} = 0 \qquad \operatorname{Re} \frac{4z}{(1+z)^2} > \frac{1}{1-|\gamma|^2}. \tag{2.8}$$

Now

$$\operatorname{Im} \frac{z}{(1+z)^2} = \frac{\operatorname{Im} z(1+\bar{z})^2}{|1+z|^4} = \frac{\operatorname{Im}(z+2|z|^2+\bar{z}|z|^2)}{|1+z|^4} = \frac{1-|z|^2}{|1+z|^4} \operatorname{Im} z.$$

Hence $\operatorname{Im}(z/(1+z)^2) = 0$ if and only if either $|z| = 1$ or $z \in \mathbf{R}$. If $z \in \mathbf{R}$, the function $4z/(1+z)^2$ assumes its maximum value at $z = 1$ and the maximum is 1, which is less than $1/(1-|\gamma|^2)$. Thus there is no real value of z for which (2.8) is satisfied. If $|z| = 1$, set $z = e^{i\varphi}$. Then

$$\begin{aligned}\frac{4z}{(1+z)^2} &= \frac{4e^{i\varphi}(1+e^{-i\varphi})^2}{|1+e^{i\varphi}|^4} \\ &= \frac{4e^{i\varphi}+8+4e^{-i\varphi}}{((1+\cos\varphi)^2+\sin^2\varphi)^2} \\ &= \frac{8(1+\cos\varphi)}{4(1+\cos\varphi)^2} = \frac{2}{1+\cos\varphi}.\end{aligned}$$

Hence (2.8) is satisfied if and only if

$$|z| = 1 \quad \text{and} \quad \cos(\arg z) < 1 - 2|\gamma|^2. \tag{2.9}$$

One reaches the conclusion that for $0 < |\gamma| < 1$ the transformation $t(z, w)$ is elliptic for all z on the interior of the arc A of the unit circle for which $\cos(\arg z) \leq 1 - 2|\gamma|^2$.

At the endpoints z_1, z_2 of the arc A the transformation is parabolic. The two points z_1, z_2 are also the branch points of the functions $w_1(z)$, $w_2(z)$. Both functions are holomorphic for

$$z \in D := \mathbf{C} \sim A, \qquad z \neq 0. \tag{2.10}$$

Moreover, $Q(\gamma, z) < 1$ for $z \in D$, because we have $z = 0 \in D$ and $Q(\gamma, 0) = 0$. Since $Q(\gamma, z) = 1$ only on A, our assertion follows.

It will now be useful to introduce

$$u_i(z) := (1 + \bar{\gamma} z w_i), \qquad z \neq 0, \quad u_1(1) := 0, \quad u_2(0) := 1. \tag{2.11}$$

Then it follows from (2.7) that

$$Q(\gamma, z) = \left| \frac{u_1(z)}{u_2(z)} \right|.$$

Set

$$r := r(z) = \frac{u_1(z)}{u_2(z)}.$$

Then for $z \in D$

$$\frac{u_1}{u_2} = r, \qquad |r| =: R < 1. \tag{2.12}$$

3. Two lemmas. From (1.5), (1.6), (2.10), and (2.6b) we obtain

$$\begin{aligned} C_n z w_i + D_n &= z C_{n-1} z w_i + (\bar{\gamma} + \bar{\delta}_n) D_{n-1} z w_i + D_{n-1} + (\gamma + \delta_n) z C_{n-1} \\ &= u_i (C_{n-1} z w_i + D_{n-1}) + \bar{\delta}_n D_{n-1} z w_i + \delta_n z C_{n-1} \\ &= u_i^n (C_0 z w_i + D_0) + \sum_{\nu=0}^{n-1} u_i^\nu (\bar{\delta}_{n-\nu} z w_i D_{n-\nu-1} + \delta_{n-\nu} z C_{n-\nu-1}) \\ &= u_i^n \left(z w_i + \gamma_0 + \sum_{\mu=0}^{n-1} \left(\bar{\delta}_{\mu+1} z w_i \frac{D_\mu}{u_i^{\mu+1}} + \delta_{\mu+1} \frac{z C_\mu}{u_i^{\mu+1}} \right) \right). \end{aligned}$$

Analogous formulas are known from continued-fraction theory (see, for example, [11], [13]). Now set

$$\frac{C_m}{u_2^{m+1}} =: \Gamma_m, \quad \frac{D_m}{u_2^{m+1}} =: \Delta_m, \quad \frac{E_m}{u_2^{m+1}} =: H_m, \quad \frac{F_m}{u_2^{m+1}} =: \Phi_m. \tag{3.1}$$

Then, if $u_1 \neq 0$,

$$u_2(\Gamma_n z w_1 + \Delta_n) = r^n \left(z w_i + \gamma_0 + \sum_{\mu=0}^{n-1} \left(\bar{\delta}_{\mu+1} z w_1 r^{-(\mu+1)} \Delta_\mu + \delta_{\mu+1} z r^{-(\mu+1)} \Gamma_\mu \right) \right),$$

$$u_2(\Gamma_n z w_2 + \Delta_n) = z w_2 + \gamma_0 + \sum_{\mu=0}^{n-1} (\bar{\delta}_{\mu+1} z w_2 \Delta_\mu + \delta_{\mu+1} z \Gamma_\mu).$$

Hence

$$u_2(w_2-w_1)z\Gamma_n = zw_2 - zr^n w_1 + \gamma_0(1-r^n) + \sum_{\mu=0}^{n-1}(\bar{\delta}_{\mu+1}z(w_2 - r^{n-\mu-1}w_1)\Delta_\mu + \delta_{\mu+1}z(1-r^{n-\mu-1})\Gamma_\mu) \tag{3.2}$$

and

(3.3)

$$u_2(w_2-w_1)\Delta_n = zw_1w_2(r^n-1) + \gamma_0(w_2r^n - w_1) + \sum_{\mu=0}^{n-1}(\bar{\delta}_{\mu+1}z(w_1w_2(r^{n-\mu-1}-1))\Delta_\mu + \delta_{\mu+1}z(w_2r^{n-\mu-1} - w_1)\Gamma_\mu).$$

Introduce

$$s_m := \sum_{k=0}^{m} r^k, \qquad m \geq 0, \quad s_{-1} := 0, \quad s_{-2} := -\frac{1}{r}. \tag{3.4}$$

Then for $m \geq 0$

$$\begin{aligned} 1 - r^m &= \frac{\bar{\gamma}z(w_2-w_1)}{u_2}s_{m-1}, \\ w_1w_2(r^m-1) &= \gamma\frac{w_2-w_1}{u_2}s_{m-1}, \\ w_2 - r^m w_1 &= \frac{w_2-w_1}{u_2}(u_2s_m - s_{m-1}), \\ w_2r^m - w_1 &= \frac{w_2-w_1}{u_2}(s_{m-1} - u_1s_{m-2}). \end{aligned} \tag{3.5}$$

Substituting the results of (3.5) into (3.2) and (3.3), one arrives at

$$u_2^2\Gamma_n = u_2s_n - s_{n-1} + \gamma_0\bar{\gamma}s_{n-1} + \sum_{\mu=0}^{n-1}(\bar{\delta}_{\mu+1}(u_2s_{n-\mu-1} - s_{n-\mu-2})\Delta_\mu + \delta_{\mu+1}\bar{\gamma}zs_{n-\mu-2}\Gamma_\mu) \tag{3.6}$$

and

$$u_2^2\Delta_n = \gamma zs_{n-1} + \gamma_0(s_{n-1} - u_1s_{n-2}) + \sum_{\mu=0}^{n-1}(\bar{\delta}_{\mu+1}\gamma zs_{n-\mu-2}\Delta_\mu + \delta_{\mu+1}z(s_{n-\mu-2} - u_1s_{n-\mu-3})\Gamma_\mu). \tag{3.7}$$

From (1.5) it also follows that for H_n and Φ_n one has

$$u_2^2H_n = \bar{\gamma}_0(u_2s_n - s_{n-1}) + \gamma s_{n-1} + \sum_{\mu=0}^{n-1}(\bar{\delta}_{\mu+1}(u_2s_{n-\mu-1} - s_{n-\mu-2})\Phi_\mu + \delta_{\mu+1}\bar{\gamma}zs_{n-\mu-2}H_\mu) \tag{3.8}$$

and

$$u_2^2\Phi_n = \gamma_0\gamma z s_{n-1} + s_{n-1} - u_1 s_{n-2}$$
$$+ \sum_{\mu=0}^{n-1}(\bar{\delta}_{\mu+1}\gamma z s_{n-\mu-2}\Phi_\mu + \delta_{\mu+1} z(s_{n-\mu-2} - u_1 s_{n-\mu-3})H_\mu). \tag{3.9}$$

We have proved the following lemma.

LEMMA 3.1. *With notation as defined in* (3.1), (1.5), (1.6), (2.10), (2.11), *and* (3.4) *one has, for* $z \in D$, $u_1 \neq 0$, *and* $0 < |\gamma| < 1$, *the sum formulas* (3.6), (3.7), (3.8), *and* (3.9).

Another useful lemma is the following.

LEMMA 3.2. *Let*

$$A_n \leq \alpha_0 + \sum_{k=0}^{n-1} a_{k+1}A_k, \qquad n \geq 1, \quad A_0 \leq \alpha_0,$$

where, for $n \geq 0$, A_n *and* α_n *all are nonnegative real numbers. Then*

$$A_n \leq \alpha_0 \prod_{k=1}^{n}(1 + \alpha_k), \qquad n \geq 1. \tag{3.10}$$

Proof. One has

$$A_1 \leq \alpha_0 + \alpha_1 A_0 \leq \alpha_0(1 + \alpha_1).$$

Set $P_0 := \alpha_0$ and

$$P_n := \alpha_0 \prod_{k=1}^{n}(1 + \alpha_k), \qquad n \geq 1,$$

and assume that $A_k \leq P_k$ for $0 \leq k \leq n-1$. Then

$$A_n \leq \alpha_0 + \sum_{k=0}^{n-1}((1 + \alpha_k) - 1)P_k = \alpha_0 + \sum_{k=0}^{n-1}(P_{k+1} - P_k) = \alpha_0 - P_0 + P_n = P_n.$$

The lemma is thus proved by induction. □

Lemma 3.2 is the discrete version of Gronwall's inequality, which can also be found in [11] and in slightly different forms in other places.

Combining the two lemmas, we can get bounds for $|\Gamma_n| + |\Delta_n|$ and $|H_n| + |\Phi_n|$.

From (3.6) and (3.7), using the fact that $|\gamma_0|$, $|\gamma|$ are less than 1 and $|s_m| \leq 1/R(1-R)$ for $m \geq -2$, one obtains

$$R(1-R)|u_2^2||\Gamma_n| \leq 2 + |u_2| + \sum_{\mu=0}^{n-1} d_{\mu+1}((1 + |u_2|)|\Delta_\mu| + |z||\Gamma_\mu|),$$

$$R(1-R)|u_2^2||\Delta_n| \leq |z| + 1 + |u_1| + \sum_{u=0}^{n-1} d_{\mu+1}(|z||\Delta_\mu| + (|z| + |zu_1|)|\Gamma_\mu|).$$

Combining the two inequalities, one arrives at

$$R(1-R)|u_2^2|(|\Gamma_n|+|\Delta_n|) \le |z|+3+|u_1|+|u_2| + \sum_{\mu=0}^{n-1} d_{\mu+1}(1+2|z|+|zu_1|+|u_2|)(|\Gamma_\mu|+|\Delta_\mu|).$$

The conditions of Lemma 3.2 then are satisfied with

$$\alpha_0 = \frac{|z|+3+|u_1|+|u_2|}{|u_2^2|R(1-R)}$$

and

$$\alpha_m = \frac{d_m(1+2|z|+|zu_1|+|u_2|)}{|u_2^2|R(1-R)}.$$

Thus (an analogous argument is valid for $|H_n|+|\Phi_n|$)

$$\left.\begin{matrix}|\Gamma_n|+|\Delta_n| \\ |H_n|+|\Phi_n|\end{matrix}\right\} \le c_0 \prod_{\mu=1}^{n}(1+c_1 d_\mu), \qquad n \ge 1. \tag{3.11}$$

4. A formula for the truncation error $T_n(0) - T$. In [10] we showed that many formulas useful in the theory of continued fractions can be derived from the invariance of the cross ratio of four distinct complex numbers u, v, w, z under a linear fractional transformation. Let F be such a mapping. Then the invariance can be expressed by the formula

$$\left(\frac{u-v}{u-w}\right)\left(\frac{w-z}{v-z}\right) = \left(\frac{F(u)-F(v)}{F(u)-F(w)}\right)\left(\frac{F(w)-F(z)}{F(v)-F(z)}\right). \tag{4.1}$$

We are interested in the case for which $F = T_n$, as defined in (1.3). It will be clear that the same approach will also work for more general linear fractional transformations t_n than those given by (1.1). In those cases one would, of course, have different values for $t_n^{-1}(0)$ and $t_{n+1}(0)$. A formula analogous to the one to be obtained here was found by Waadeland and the author [14] in 1983 for continued fractions $K(a_n/1)$.

In addition to setting $F = T_n$, let $z = g_n$, where

$$g_n := T_n^{-1}(\infty). \tag{4.2}$$

Then

$$\left(\frac{u-v}{u-w}\right)\left(\frac{w-g_n}{v-g_n}\right) = \frac{T_n(u)-T_n(v)}{T_n(u)-T_n(w)}. \tag{4.3}$$

Assume $\{T_n(0)\}$ converges to a limit T. This will certainly be the case if $|z| < 1$. Furthermore, let

$$T_{n+m}^{(n)}(w) = t_n \circ \cdots \circ t_{n+m}(w). \tag{4.4}$$

Then $|T_{n+m}^{(n)}(0)\}_{m=0}^{\infty}$ also converges. We denote the limit by $T^{(n)}$.

$$T^{(n)} = \lim_{m\to\infty} T_{n+m}^{(n)}(0). \tag{4.5}$$

Since $T_{n+m}(w) = T_n \circ T_{n+m}^{(n+1)}(w)$, we have

$$T = T_n(T^{(n+1)}). \tag{4.6}$$

In (4.3) st $u = 0$, $v = T^{(n+1)}$, $w = t_n^{-1}(0)$.

Then for $\gamma_n \neq 0$

$$\frac{T_n(0) - T}{T_n(0) - T_{n-1}(0)} = \frac{T_n(0) - T_n(T^{(n+1)})}{T_n(0) - T_n(t_n^{-1}(0))}$$
$$= \left(\frac{0 - T^{(n+1)}}{0 - t_n^{-1}(0)}\right)\left(\frac{t_n^{-1}(0) - g_n}{T^{(n+1)} - g_n}\right) = \frac{T^{(n+1)}(\gamma_n + zg_n)}{\gamma_n(g_n - T^{(n+1)})}.$$

Hence for $\gamma_n \neq 0$

$$T - T_n(0) = \frac{T^{(n+1)}(\gamma_n + zg_n)}{\gamma_n(g_n - T^{(n+1)})}(T_n(0) - T_{n-1}(0)). \tag{4.7}$$

Note that for $\gamma_n = 0$ one has $T_n(0) = T_{n-1}(0)$.

Next, let $u = 0$, $v = t_{n+1}(0)$, $w = t_n^{-1}(0)$ in (4.3). Then for $\gamma_n \neq 0$

$$\frac{T_n(0) - T_{n+1}(0)}{T_n(0) - T_{n-1}(0)} = \frac{T_n(0) - T_n(t_{n+1}(0))}{T_n(0) - T_n(t_n^{-1}(0))}$$
$$= \left(\frac{0 - \gamma_{n+1}}{0 - (-\gamma_n/z)}\right)\left(\frac{-\gamma_n/z - g_n}{\gamma_{n+1} - g_n}\right)$$
$$= \frac{\gamma_{n+1}(\gamma_n + zg_n)}{\gamma_n(\gamma_{n+1} - g_n)}.$$

Hence

$$T_{n+1}(0) - T_n(0) = \frac{\gamma_{n+1}(\gamma_n + zg_n)}{\gamma_n(g_n - \gamma_n)}(T_n(0) - T_{n-1}(0)). \tag{4.8}$$

It follows that

$$T - T_n(0) = \frac{T^{(n+1)}(\gamma_n + zg_n)}{\gamma_n(g_n - T^{(n+1)})} \prod_{m=1}^{n-1} \left(\frac{T_{m+1}(0) - T_m(0)}{T_m(0) - T_{m-1}(0)}\right)(T_1(0) - T_0(0))$$

$$= \frac{T^{(n+1)}(\gamma_n + zg_n)}{\gamma_n(g_n - T^{(n+1)})} \prod_{m=1}^{n-1} \frac{\gamma_{m+1}(\gamma_m + zg_m)}{\gamma_m(g_m - \gamma_{m+1})}\left(\frac{\gamma_0 + z\gamma_1}{1 + \bar{\gamma}_0 z\gamma_1} - \gamma_0\right) \tag{4.9}$$

$$= \frac{T^{(n+1)}(\gamma_n + zg_n)}{g_n - T^{(n+1)}} \prod_{m=1}^{n-1} \left(\frac{\gamma_m + zg_m}{g_m - \gamma_{m+1}}\right) \cdot \frac{z(1 - |\gamma_0|^2)}{1 + z\bar{\gamma}_0\gamma_1}.$$

This formula is valid provided that $T_n(0) \to T$ and $\gamma_n \neq 0$, $m \geq 0$.

If we are dealing with a lpS, that is, if $d_n \to 0$, then $g_n \to w_1$, as we shall show in §5. In this case the factors in the product in (4.9) approach

$$\frac{\gamma + zw_1}{w_1 - \gamma}.$$

For this expression we have

$$\frac{\gamma + zw_1}{w_1 - \gamma} = \frac{w_1(1+\bar{\gamma}zw_1)}{-\frac{\gamma}{\bar{\gamma}zw_2} - \gamma} = \frac{w_1 w_2 \bar{\gamma} z(1+\bar{\gamma}zw_1)}{-\gamma(1+\bar{\gamma}zw_2)}$$

$$= \frac{1+\bar{\gamma}zw_1}{1+\bar{\gamma}zw_2} = \frac{u_1}{u_2} = r.$$

Introduce R' so that $0 < R < R' < 1$, and set

$$\zeta_n = w_1 - g_n \to 0.$$

Then

$$\left|\frac{\gamma_n + zg_n}{g_n - \gamma_{n+1}}\right| = \left|\frac{\gamma + zw_1 + \delta_n - \zeta_n z}{w_1 - \gamma - \delta_{n+1} - \zeta_n}\right|$$

$$= R\left|\frac{1 + \frac{\delta_n - \zeta_n z}{\gamma + zw_1}}{1 + \frac{\delta_{n+1} + \zeta_n}{w_1 - \gamma}}\right|$$

$$= R'\frac{R}{R'}\left|\frac{1 + \frac{\delta_n - \zeta_n z}{\gamma + zw_1}}{1 - \frac{\delta_{n+1} + \zeta_n}{w_1 - \gamma}}\right|$$

$$< R' \quad \text{for } n > N(R').$$

In §6 we shall show that $T^{(n)} \to w_2$. Thus, finally, for lpS with $\gamma_m \neq 0$, $m \geq 0$, $|z| < 1$ the following truncation error estimate is valid:

$$|T - T_n(0)| < K(R')|w_2|\left|\frac{\gamma + zw_1}{w_1 - w_2}\right|(R')^{n-1}, \qquad R < R' < 1. \tag{4.10}$$

Note that if $\sum d_n < \infty$ and $\sum |\zeta_n| < \infty$, then $|T - T_n(0)| < KR^n$.

Returning once more to the formula (4.3) and setting $u = T^{(n+1)}$, $v = w_2$, $w = 0$, we arrive at

$$\frac{T - T_n(w_2)}{T - T_n(0)} = \frac{g_n(T^{(n+1)} - w_2)}{(g_n - w_2)T^{(n+1)}} \sim \frac{w_1(T^{(n+1)} - w_2)}{(w_1 - w_2)w_2}, \tag{4.11}$$

which is valid for lpS with $|z| < 1$. From (4.11) it follows that $\{T_n(w_2)\}$ converges faster to T than does $\{T_n(0)\}$.

5. Convergence of $\{g_n\}$. We show here that $g_n \to w_1$ provided that $d_n \to 0$ and $g_n - w_2 \neq 0$ for $n > N$. This last condition will be satisfied at least for $|z| < 1$. We recall that g_n was defined in (4.2) and is

$$g_n = T_n^{-1}(\infty).$$

Its introduction was motivated by the substantial simplification that can be achieved in the formula (4.1) if one sets $z = g_n$.

In §4 we defined

$$T_{n+m}^{(n)}(w) = t_n \circ \cdots \circ t_{n+m}(w).$$

Continuing in this vein, we now write

$$T_{n+m}^{(n)} =: \frac{C_{n+m}^{(n)} zw + D_{n+m}^{(n)}}{E_{n+m}^{(n)} zw + F_{n+m}^{(n)}} \tag{5.1}$$

and

$$\Gamma_{n+m}^{(n)} := \frac{C_{n_m}^{(n)}}{u_2^{m+1}}, \quad \Delta_{n+m}^{(n)} := \frac{D^{(n)}}{u_2^{m+1}}, \quad H_{n+m}^{(n)} := \frac{E_{n+m}^{(n)}}{u_2^{m+1}}, \quad \Phi_{n+m}^{(n)} := \frac{F_{n+m}^{(n)}}{u_2^{m+1}}. \tag{5.2}$$

We obtain the expressions (5.3)–(5.6) by rewriting (3.6)–(3.9). We replace γ_0 by γ_n and rearrange the "constant terms" as

$$\begin{aligned} u_2 s_m - s_{m-1} + \gamma_n \gamma s_{m-1} &= (u_2 - 1) s_{m-1} + \gamma_n \gamma s_{m-1} + u_2 r^m, \\ &= (\bar{\gamma} z w_2 + \gamma_n \gamma) s_{m-1} + u_2 r^m \end{aligned}$$

and similarly for the other "constant terms." This leads to

$$u_2^2 \Gamma_{n+m}^{(n)} =: (\bar{\gamma} z w_2 + \gamma_n \gamma) s_{m-1} + \sigma_{n,m}^{(1)}, \tag{5.3}$$

$$u_2^2 \Delta_{n+m}^{(n)} =: (\gamma z - \gamma_n \gamma z w_1) s_{m-1} + \sigma_{n,m}^{(2)}, \tag{5.4}$$

$$u_2^2 H_{n+m}^{(n)} =: (\bar{\gamma}_n \bar{\gamma} z w_2 + \bar{\gamma}) s_{m-1} + \sigma_{n,m}^{(3)}, \tag{5.5}$$

$$u_2^2 \Phi_{n+m}^{(n)} =: (\bar{\gamma}_n \gamma z - \bar{\gamma} z w_1) s_{m-1} + \sigma_{n,m}^{(4)}. \tag{5.6}$$

Here

$$\sigma_{n,m}^{(1)} = u_2 r^m + \sum_{\mu=0}^{m-1} (\bar{\delta}_{n+m+1} (u_2 s_{m-\mu-1} - s_{m-\mu-2}) \Delta_{n+m}^{(n)} + \delta_{n+\mu+1} \bar{\gamma} z s_{m-\mu-2} \Gamma_{n+\mu}^{(n)}),$$

$$\sigma_{n,m}^{(2)} = \gamma_n u_1 r^{m-1} + \sum_{\mu=0}^{m-1} (\bar{\delta}_{n+\mu+1} \gamma z s_{m-\mu-2} \Delta_{n_\mu}^{(n)} + \delta_{n+\mu+1} (s_{m-\mu-2} - u_1 s_{m-\mu-3}) \Gamma_{n+\mu}^{(n)}),$$

$$\sigma_{n,m}^{(3)} = \bar{\gamma}_n u_2 r^m + \sum_{\mu=0}^{m-1} (\bar{\delta}_{n+\mu+1} (u_2 s_{m-\mu-1} - s_{m-\mu-2}) \Phi_{n+\mu}^{(n)} + \delta_{n+\mu+1} \gamma z s_{m-\mu-2} H_{n+\mu}^{(n)}),$$

$$\sigma_{n,m}^{(4)} = u_1 r^{m-1} + \sum_{\mu=0}^{m-1} (\bar{\delta}_{n+\mu+1} \gamma z s_{m-\mu-2} \Phi_{n+\mu}^{(n)} + \delta_{n+\mu+1} (s_{m-\mu-2} - u_1 s_{m-\mu-3}) H_{n+\mu}^{(n)}).$$

One can write

$$\begin{aligned} g_N = g_{n+m} &= T_{n+m}^{-1}(\infty) = t_{n+m}^{-1} \circ \cdots \circ t_n^{-1}(g_{n-1}) \\ &= \left(\frac{C_{n+m}^{(n)} z g_{n-1} + D_{n+m}^{(n)}}{E_{n+m}^{(n)} z g_{n-1} + F_{n+m}^{(n)}} \right)^{-1} = \frac{1}{z} \frac{F_{n+m}^{(n)} g_{n-1} - D_{n+m}^{(n)}}{E_{n+m}^{(n)} g_{n-1} - C_{n+m}^{(n)}} \\ &= -\frac{1}{z} \frac{u_2^2 \Phi_{n+m}^{(n)} g_{n-1} - u_2^2 \Delta_{n+m}^{(n)}}{u_2^2 H_{n+m}^{(n)} g_{n-1} - u_2^2 \Gamma_{n+m}^{(n)}}. \end{aligned} \tag{5.7}$$

Then

$$g_{n+m} = -\frac{1}{z}\frac{(\bar\gamma_n\gamma z-\bar\gamma z w_1)s_{m-1}g_{n-1}+\sigma^{(4)}_{m,n}g_{n-1}-(\gamma z-\gamma_n\bar\gamma z w_1)s_{m-1}-\sigma^{(2)}_{m,n}}{(\bar\gamma_n\bar\gamma z w_2-\bar\gamma)s_{m-1}g_{n-1}+\sigma^{(3)}_{m,n}g_{n-1}-(\bar\gamma z w_2+\gamma_n\bar\gamma)s_{m-1}-\sigma^{(1)}_{m,n}}$$

$$= w_1\frac{(\bar\gamma+\bar\gamma_n\bar\gamma w_2 z)g_{n-1}-(\gamma_n\bar\gamma+\bar\gamma z w_2)-\sigma^{(4)}_{n,m}g_{n-1}-\sigma^{(2)}_{n,m})/w_1 s_{m-1}z}{(\bar\gamma+\bar\gamma_n\bar\gamma w_2 z)g_{n-1}-(\gamma_n\bar\gamma+\bar\gamma z w_2)-(\sigma^{(1)}_{n,m}-\sigma^{(3)}_{n,m}g_{n-1})/s_{m-1}}$$

$$= w_1\frac{(1+\bar\gamma z w_2)g_{n-1}-(\gamma+z w_2)+\bar\delta_n z w_2 g_{n-1}-\delta_n-(\sigma^{(4)}_{n,m}g_{n-1}-\sigma^{(2)}_{n,m})/w_1 s_{m-1}\bar\gamma z}{(1+\bar\gamma z w_2)g_{n-1}-(\gamma+z w_2)+\bar\delta_n z w_2 g_{n-1}-\delta_n-(\sigma^{(1)}_{n,m}-\sigma^{(3)}_{n,m}g_{n-1})/s_{m-1}\bar\gamma}$$

$$= w_1\frac{g_{n-1}-w_2+(\bar\delta_n z w_2 g_{n-1}-\delta_n)/u_2-(\sigma^{(4)}_{n,m}g_{n-1}-\sigma^{(2)}_{n,m})/w_1 s_{m-1}\bar\gamma z u_2}{g_{n-1}-w_2+(\bar\delta_n z w_2 g_{n-1}-\delta_n)/u_2-(\sigma^{(1)}_{n,m}-\sigma^{(3)}_{n,m}g_{n-1})/s_{m-1}\bar\gamma u_2}.$$

Hence

$$\frac{g_{n+m}}{w_1}-1 = \frac{-(\sigma^{(4)}_{n,m}g_{n-1}-\sigma^{(2)}_{n,m})/w_1 z+(\sigma^{(1)}_{n,m}-\sigma^{(3)}_{n,m}g_{n-1})}{(g_{n-1}-w_2)s_{m-1}\bar\gamma u_2+(\bar\delta_n z w_2 g_{n-1}-\delta_n)-(\sigma^{(1)}_{n,m}-\sigma^{(3)}_{n,m}g_{n-1})}$$

$$= \frac{w_2\bar\gamma(\sigma^{(4)}_{n,m}-\sigma^{(2)}_{n,m}/g_{n-1})/\gamma+(\sigma^{(1)}_{n,m}/g_{n-1}-\sigma^{(3)}_{n,m})}{(1-w_2/g_{n-1})s_{m-1}\bar\gamma u_2+(\bar\delta_n z w_2-\delta_n/g_{n-1})s_{m-1}\gamma-(\sigma^{(1)}_{n,m}/g_{n-1}-\sigma^{(3)}_{n,m})}$$

provided that $|1/g_{n-1}|<1$, $w_1\neq 0$, $|w_2/g_{n-1}|<1$. We note that $w_1=0$ if and only if $\gamma=0$, which we are excluding. Furthermore, for $|z|<1$ the function $t_n^{-1}(u)$ maps $|u|>1$ into $|w|>1$. Hence $g_n=T_n^{-1}(\infty)$ must satisfy $|g_n|>1$. Finally, for $|z|<1$, $|w_2|\leq 1$.

It follows that, at least for $|z|<1$,

$$|g_{n+m}-w_1| \leq \frac{|w_1|\max(1,|w_2|)\sum_{j=1}^4|\sigma^{(j)}_{n,m}|}{|1-w_2/g_{n-1}||\gamma u_2|-d_n(|z w_2|+1)/(1-R)-|\sigma^{(1)}_{n,m}|-\sigma^{(3)}_{n,m}|}. \tag{5.8}$$

From Lemma 3.2 and the inequality (3.11) one deduces that $|\Gamma^{(n)}_{n+\mu}|$, $|\Delta^{(n)}_{n+\mu}|$, $|H^{(n)}_{n+\mu}|$, and $|\Phi^{(n)}_{n+\mu}|$ are, for $0<\mu<m$, all bounded by

$$c_0\prod_{\nu=0}^{\mu-1}(1+c_1 d_{n+\nu})\leq c_0(1+c_1 d_n)^m.$$

Thus

$$\sum_{j=1}^4|\sigma^{(j)}_{n,m}|<KR^m+\Theta m d_n(1+c_1 d_n)^m.$$

Given an $\epsilon>0$, we can choose m_ϵ so large that

$$KR^{m_\epsilon}<\frac{\epsilon L}{4},$$
$$\frac{\Theta}{m_\epsilon}\left(1+\frac{1}{m_\epsilon}\right)^{m_\epsilon}<\frac{\Theta e}{m_\epsilon}<\frac{\epsilon L}{4}, \tag{5.9}$$

where

$$0<L:=L(z)=(1-|z|)|\gamma u_2|.$$

Note that we have

$$\left|1-\frac{w_2}{g_{n-1}}\right| \geq 1-\left|\frac{w_2}{g_{n-1}}\right| \geq 1-|z|.$$

For a fixed m_ϵ we then choose n_ϵ so that for $n \geq N_\epsilon - m_\epsilon$

$$c_1 d_n < \frac{1}{m_\epsilon}, \qquad m_\epsilon d_n < \frac{1}{m_\epsilon}$$

$$\begin{aligned}\frac{d_n(|zw_2|+1)}{1-R} + |\sigma^{(1)}_{n,m_\epsilon}| + \sigma^{(3)}_{n,m_\epsilon}| &< \frac{d_n(|zw_2|+1)}{1-R} \\ &\quad + K^* R^{m_\epsilon} + \Theta^* d_n m_\epsilon (1+c_1 d_n)^{m_\epsilon} \\ &< \frac{L}{2}.\end{aligned}$$

We then have

$$\begin{aligned}&\left|1-\frac{w_2}{g_{n-1}}\right||\gamma u_2| - \left(\frac{d_n(|zw_2|+1)}{1-R} + |\sigma^{(1)}_{n,m}| + |\sigma^{(3)}_{n,m}|\right) \\ &\quad > (1-|z|)|\gamma u_2| - \left(\frac{d_n(|zw_2|+1)}{1-R} + |\sigma^{(1)}_{n,m}| + |\sigma^{(3)}_{n,m}|\right) \\ &\quad > L - \frac{L}{2} = \frac{L}{2}.\end{aligned}$$

It follows that

$$|g_n - w_1| < \epsilon \quad \text{for } N > N_\epsilon.$$

Hence g_n converges to w_1, at least for $|z| < 1$.

6. Convergence of $\{T^{(n)}\}$. We have

$$T^{(n)}(w) = \lim_{m\to\infty} T^{(n)}_{n+m}(w), \quad T^{(n)} = T^{(n)}(0)$$

(see also (4.5)). We shall prove that $T^{(n)} \to w_2$. The proof is an adaptation of a proof of Perron [7, pp. 93–94] for the tails of a limit periodic continued fraction to lpS.

For periodic sequences, that is, when $\gamma_m = \gamma$ for all m, we use the notation

$$\overset{P}{T}{}^{(n)}_{n+m}(w). \tag{6.1}$$

From formulas (3.6)–(3.9) with $\delta_m = 0$ for all m one deduces easily that

$$\overset{P}{T}{}^{(n)}_{w} = \lim_{m\to\infty} \overset{P}{T}{}^{(n)}_{n+m}(w) = w_2, \qquad w \neq w_1, \quad |z| < 1. \tag{6.2}$$

Hence

(6.3)

$$\begin{aligned}T^{(n)}(w) - w_2 &= T^{(n)}(w) - \overset{P}{T}{}^{(n)}_{n+m}(w) \\ &= (T^{(n)}(w) - T^{(n)}_{n+m}(w)) + (T^{(n)}_{n+m}(w) - \overset{P}{T}{}^{(n)}_{n+m}(w)) \\ &\quad + \left(\overset{P}{T}{}^{(n)}_{n+m}(w) - \overset{P}{T}{}^{(n)}(w)\right).\end{aligned}$$

According to Schur [8, p. 211], one has for all $|w| < 1$, $|z| < a < 1$

$$|T_m(w) - T| < \frac{2|a|^m}{1-a}$$

provided that $|\gamma_n| < 1$, $n \geq 1$. The convergence of $\{T_{n+m}^{(n)}(w)\}$ is thus independent of the choice of the sequence $\{\gamma_{n+m}\}_{m=0}^{\infty}$. It is also independent of the values of z and w provided that $|z| < a < 1$, $|w| < 1$ (and hence $w \neq w_1$).

We now choose m so large that the terms in the first and third sets of parentheses on the right-hand side of (6.3) are both, in absolute value, less than $\epsilon/3$ for a given $\epsilon > 0$. The m found in the preceding is independent of n. It is then clear that we can find an N such that

$$\left| T_{n+m}^{(n)}(w) - \overset{P}{T}{}_{n+m}^{(n)}(w) \right| < \epsilon/3 \quad \text{for } n > N.$$

This follows from the facts that $T_{n+m}^{(n)}(w)$ is a continuous function of the $m+1$ variables $\gamma_n, \ldots, \gamma_{n+m}$ and that $\gamma_{n+m} \to \gamma$ as $n \to \infty$. Hence

$$\lim_{n\to\infty} T^{(n)}(w) = w_2 \tag{6.4}$$

provided that $|z| < 1$, $|w| < 1$.

To find a truncation error for $T^{(n)} - w_2$ we proceed as follows. Let n be fixed and $m > n$. Then

$$T^{(m)} = T^{(m)}(0) = t_m \circ T^{(m+1)}(0) = \frac{\gamma_m + zT^{(m+1)}}{1 + \bar{\gamma}_m z T^{(m+1)}}.$$

Recall (see (1.6)) that $\gamma_m = \gamma + \delta_m$. Let

$$T^{(m)} =: w_2 + \omega_m. \tag{6.5}$$

In terms of these expressions we have, since $-w_2^2\bar{\gamma}z + w_2(z-1) + \gamma = 0$,

$$\begin{aligned}
\omega_m &= T^{(m)} - w_2 \\
&= \frac{\gamma + \delta_m + zw_2 + z\omega_{m+1} - w_2(1 + \bar{\gamma} - zw_2 + \bar{\delta}_m z w_2 + (\bar{\gamma} + \bar{\delta}_m) z \omega_{m+1})}{1 + \bar{\gamma} z w_2 + \bar{\delta}_m z w_2 + (\bar{\gamma} + \bar{\delta}_m) z \omega_{m+1}} \\
&= \frac{\bar{\delta}_m - \bar{\delta}_m z w_2^2 + z\omega_{m+1}(1 - w_2(\bar{\gamma} + \bar{\delta}_m))}{1 + \bar{\gamma} z w_2 + \bar{\delta}_m z w_2 + (\bar{\gamma} + \bar{\delta}_m) z \omega_{m+1}}.
\end{aligned}$$

Since $m > n$ implies $d_m \leq d_n$, we have

$$|\omega_m| \leq \frac{d_n(1 + |zw_2^2|) + |\omega_{m+1}|(|z - w_2\bar{\gamma}z| + |w_2 z| d_n)}{|1 + \bar{\gamma} z w_2| - \delta_n |z w_2| - |\omega_{m+1}|(|\gamma| + d_n)|z|}. \tag{6.6}$$

We would like to prove that there exists a $P > 0$ so that

$$|\omega_m| < P d_n, \qquad m > n > n_0. \tag{6.7}$$

Our proof is similar to that given in [12].

We consider the inequality

$$P d_n \geq \frac{d_n(1 + |zw_2^2|) + P d_n(|z - w_2\bar{\gamma}z| + |w_2 z| d_n)}{|1 + \bar{\gamma} z w_2| - d_n |z w_2| - d_n P(|\gamma| + d_n)|z|}. \tag{6.8}$$

If $d_n \neq 0$, which is the only interesting case, then (6.8) is equivalent to

$$P \geq \frac{1 + |zw_2^2| + P(|z - w_2\bar{\gamma}z| + |w_2 z|d_n)}{|1 + \bar{\gamma}zw_2| - d_n|zw_2| - d_n P(|\gamma| + d_n)|z|}. \tag{6.9}$$

Now set

$$D := |1 + \bar{\gamma}zw_2| - |z - w_2\bar{\gamma}z|. \tag{6.10}$$

Then, since $\bar{\gamma}z(w_1 + w_2) = z - 1$ for $u_1 = |1 + \bar{\gamma}zw_1| > 0$,

$$\begin{aligned} D &= \left(\left|\frac{1 + \bar{\gamma}zw_2}{z - w_2\bar{\gamma}z}\right| - 1\right)|z - w_2\bar{\gamma}z| \\ &= \left(\left|\frac{1 + \bar{\gamma}zw_2}{1 + \bar{\gamma}zw_1}\right| - 1\right)|1 + \bar{\gamma}zw_1| = \left(\frac{1}{R} - 1\right)|1 + \bar{\gamma}zw_1| > 0. \end{aligned}$$

For d_n sufficiently small (so that the denominator in (6.9) is positive) the inequality (6.9) is equivalent to

$$P(D - 2d_n|zw_2|) - P^2 d_n(|\gamma| + d_n)|z| \geq 1 + |zw_2^2|. \tag{6.11}$$

Choose n_0 such that for $n > n_0(P)$

$$d_n < 1, \qquad d_n < \frac{|1 + \bar{\gamma}zw_2|}{|zw_2| + P(|\gamma| + 1)|z|} < \frac{|1 + \bar{\gamma}zw_2|}{|zw_2| + P(|\gamma| + d_n)|z|},$$

$$d_n < \frac{D}{4|zw_2|}, \qquad d_n^{1/2} < \frac{D^2}{(|\gamma| + 1)|z|} < \frac{D^2}{(|\gamma| + d_n)|z|},$$

$$d_n^{1/2} < \frac{1}{16(1 + |zw_2^2|)}.$$

Then

$$P(D - 2d_n|zw_2|) - P^2 d_n(|\gamma| + d_n)|z| \geq \frac{PD}{2} - P^2 d_n^{1/2} D^2.$$

For

$$P = \frac{4(1 + |zw_2^2|)}{D} \tag{6.12}$$

one obtains

$$\frac{PD}{2} - P^2 d_n^{1/2} D^2 = 2(1 + |zw_2^2|) - 16 d_n^{1/2}(1 + |zw_2^2|)^2 > 1 + |zw_2^2|.$$

Hence for this value of P the inequality (6.11) and thus also (6.8) are satisfied for $n > n_0$.

Now hold both m and n fixed, $m > n > n_0$. Since we know that $T^{(m+k)} \to w_2$ as $k \to \infty$, we can choose k so large that

$$|\omega_{m+k}| = |T^{(m+k)} - w_2| < P d_n.$$

From (6.6), with m replaced by $m+k-1$, and from (6.8) we have

$$|\omega_{m+k-1}| \leq \frac{d_n(1+|zw_2^2|)+Pd_n(|z0w_2\bar{\gamma}z|+|w_2z|d_n)}{|1+\bar{\gamma}zw_2|-d_n|zw_2|-d_nP(|\gamma|+d_n)|z|} \leq Pd_n.$$

Repeating this argument k times, we obtain (6.7) or, writing it explicitly,

$$|T^{(m)}-w_2| \leq \frac{4R(1+|zw_2^2|)}{|i_1|(1-R)}d_n \quad \text{for } m>n>n_0 \tag{6.13}$$

for $u_1 \neq 0$. Note that $u_1 = 0$ if and only if $|\gamma| = 1$. In that case $D = |1+z|$, $P = 4(1+|z|)/|1+z|$, and hence we have

$$|T^{(m)}-w_2| \leq \frac{4(1+|z|)}{|1+z|}d_n, \qquad m>n>n_0, \quad |\gamma|=1. \tag{6.13}'$$

REFERENCES

[1] W. B. JONES, *Schur's algorithm extended and Schur continued fractions*, in Nonlinear Numerical Methods and Rational Approximation, A. Cuyt, ed., Reidel, Dordrecht, the Netherlands, 1988, pp. 281–298.

[2] W. B. JONES, O. NJÅSTAD, AND W. J. THRON, *Schur fractions, Perron Caratheodory fractions and Szegö polynomials: A Survey*, in Analytic Theory of Continued Fractions II, W. J. Thron, ed., Lecture Notes in Math. 1199, Springer-Verlag, New York, 1986, pp. 127–158.

[3] W. B. JONES AND W. J. THRON, *Sequences of meromorphic functions corresponding to a formal Laurent series*, SIAM J. Math. Anal., 10 (1979), pp. 1–17.

[4] ———, *Continued Fractions, Analytic Theory and Applications*, Encyclopedia of Mathematics and its Applications, Vol. 11, Addison-Wesley, Reading, MA, 1980; now distributed by Cambridge University Press, London.

[5] ———, *Contractions of the Schur algorithm for functions bounded in the unit circle*, Rocky Mountain J. Math., 19 (1989), pp. 211–222.

[6] ———, *A constructive proof of convergence of the even part of positive* PC*-fractions*, Rocky Mountain J. Math., 19 (1989), pp. 199–210.

[7] O. PERRON, *Die Lehre von den Kettenbrüchen*, 3rd ed., Vol. 2, Teubner Verlag, Stuttgart, Germany, 1957.

[8] J. SCHUR, *Über Potenzreihen die im Inneren des Einheitskreises beschränkt sind*, J. Reine Angew. Math., 147 (1917), pp. 205–232; 148 (1918/19), pp. 122–145.

[9] W. J. THRON, *Two point Padé tables,* T*-fractions and sequences of Schur*, in Padé and Rational Approximation, E. B. Saff and R. S. Varga, eds., Academic Press, New York, 1977, pp. 215–226.

[10] ———, *Continued fraction identities derived from the invariance of the cross ratio under linear fractional transformations*, in Analytic Theory of Continued Fractions III, L. Jacobsen, ed., Lecture Notes in Math. 1406, Springer-Verlag, New York, 1989, pp. 124–134.

[11] ———, *Some results on separate convergence of continued fractions*, in Computational Methods and Function Theory, S. Ruscheweyh, E. B. Saff, L. C. Salinas, and R. S. Varga, eds., Lecture Notes in Math. 1435, Springer-Verlag, New York, 1990, pp. 191–200.

[12] W. J. THRON AND H. WAADELAND, *Accelerating convergence of limit periodic continued fractions* $K(a_n/1)$, Numer. Math., 34 (1980), pp. 155–170.

[13] ———, *Convergence questions for limit periodic continued fractions*, Rocky Mountain J. Math., 11 (1981), pp. 641–657.

[14] ———, *Truncation error bounds for limit periodic continued fractions*, Math. Comput., 40 (1983), pp. 589–597.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 749–775, March 1994

032

ASYMPTOTIC ANALYSIS OF SOME ASSOCIATED ORTHOGONAL POLYNOMIALS CONNECTED WITH ELLIPTIC FUNCTIONS*

GALLIANO VALENT†

Abstract. For the two families of associated Stieltjes–Carlitz polynomials a generating function is derived that allows the asymptotic analysis to be worked out and leads to the continued fraction. The spectrum and the orthogonality measure do not seem amenable to a closed form except for special values of the association parameters.

A similar analysis is developed for a particular class of associated polynomials related to a quartic birth and death process; however, in this case the Stieltjes moment problem is indeterminate. The author obtains a generating function that yields the asymptotic behavior of the related polynomials. This result gives the Stieltjes transform of a Nevanlinna extremal orthogonality measure. For two particular choices of the association parameters the reader is led to closed form results.

Key words. orthogonal polynomials, elliptic functions, asymptotic analysis

AMS subject classifications. 33C45, 33E05, 33E30, 34A05

1. Introduction.

In this article we deal with the spectral properties of the polynomials $F_n(x)$ defined by the second-order recurrence

$$(\lambda_n + \mu_n - x)F_n(x) = \mu_{n+1}F_{n+1} + \lambda_{n-1}F_{n-1}, \qquad n \geq 0, \tag{1.1}$$

$$F_{-1}(x) = 0, \qquad F_0(x) = 1, \tag{1.2}$$

satisfying the orthogonality relation

$$\frac{1}{\pi_m}\int_0^\infty d\Psi(x)F_m(x)F_n(x) = \delta_{mn} \tag{1.3}$$

where

$$\pi_0 = 1, \quad \pi_m = \frac{\lambda_0 \dots \lambda_{m-1}}{\mu_1 \dots \mu_m}, \quad m = 1, 2, \dots.$$

One of the most important problems in orthogonal polynomial theory is, therefore, given an explicit form of $\{\lambda_n, \mu_n\}$, to determine the corresponding orthogonality measure.

The knowledge of Ψ has an immediate application to birth and death processes. As explained in [18] these are special stationary Markov processes whose state space is the nonnegative integers. The probability $\mathcal{P}_{m,n}(t)$ for the system to evolve from state m at the time $t = 0$ to the state n at time t is the solution of the forward Kolmogorov equation

$$\frac{d}{dt}\mathcal{P}_{m,n}(t) = \lambda_{n-1}\mathcal{P}_{m,n-1}(t) + \mu_{n+1}\mathcal{P}_{m,n+1}(t) - (\lambda_n + \mu_n)\mathcal{P}_{m,n}(t)$$

with the boundary condition

$$\mathcal{P}_{m,n}(0) = \delta_{mn}.$$

* Received by the editors March 19, 1992; accepted for publication (in revised form) February 16, 1993.

† Laboratoire de Physique Théorique et Hautes Energies, Unité Associée au Centre National de la Recherche Scientifique UA 280, Université Paris 7, 2 Place Jussieu, F-75251 Cedex 05, Paris, France.

The coefficients λ_n (respectively, μ_n) are related to the probabilities of birth $n \to n+1$ (respectively, death $n \to n-1$) by the relations

$$\mathcal{P}_{n-1,n}(t) = \lambda_{n-1}t, \quad \mathcal{P}_{n+1,n}(t) = \mu_{n+1}t, \quad t \to 0+,$$

and will be restricted by positivity

$$\lambda_n > 0, \quad n = 0, 1, \ldots, \quad \mu_0 \geq 0, \quad \mu_n > 0, \quad n = 1, 2, \ldots.$$

We say that a process is asymptotically symmetric if we have $\lim_{n\to\infty}(\lambda_n/\mu_n) = 1$.

The link with orthogonal polynomial theory is provided by the representation theorem of Karlin and MacGregor [21], [22], which states that

$$\mathcal{P}_{m,n}(t) = \frac{1}{\pi_m}\int_0^\infty d\Psi(x)F_m(x)F_n(x)e^{-tx}.$$

In general, we have

$$\sum_{n\geq 0}\mathcal{P}_{mn}(t) \leq 1, \qquad t \geq 0,$$

and the particular solution $\mathcal{P}^*_{mn}(t)$ for which the equality holds is called the "honest" solution with measure Ψ^*.

As emphasized in [18], for most applications the transition rates $\{\lambda_n, \mu_n\}$ are *polynomials* and, unfortunately, only for a few cases are we able to get their orthogonality measure:

1. The asymptotically symmetric linear case

$$\lambda_n = n + a, \qquad \mu_n = n + b$$

was worked out by Askey and Wimp [5]; a simpler derivation was given in [20]. The polynomials involved are the associated Laguerre polynomials.

2. The nonsymmetric linear case

$$\lambda_n = c(n + a), \quad \mu_n = n + b, \quad 0 < c < 1$$

has been analyzed by Ismail, Letessier, and Valent [20], who obtained the Stieltjes function. The corresponding polynomials are the associated Meixner polynomials.

3. The asymptotically symmetric quadratic case

$$\lambda_n = n^2 + an + b, \qquad \mu_n = n^2 + cn + d$$

was worked out by Ismail, Letessier, and Valent [16], [17]. The relevant polynomials are the associated continuous dual Hahn polynomials.

Very little is known for nonsymmetric quadratic rates, for a simple reason which has to do with the generating function

$$F(x, w) = \sum_{n\geq 0} F_n(x)w^n.$$

In the asymptotically symmetric case, $F(x, w)$ is the solution of a second-order differential equation leading to hypergeometric functions, while in the nonsymmetric case we are led to a differential equation of Heun's type, which can be solved explicitly only in some very particular cases [8], [31].

Despite these difficulties two exact solutions are known which are due to Stieltjes [29]:

1. For the rates

$$\lambda_n = k^2(2n+1)^2, \quad \mu_n = (2n)^2, \quad 0 < k^2 < 1, \tag{1.4}$$

the Laplace transform of the orthogonality measure may be concisely written

$$\mathcal{P}_{00}(t) = \int_0^\infty d\Psi(x)\, e^{-tx} = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} d\theta\, \mathrm{dn}\,(\theta, k^2)\, e^{-\theta^2/4t} \tag{1.5}$$

where dn is one of Jacobi's elliptic functions of parameter k^2. For all that concerns elliptic functions we follow [36, pp. 491–528].

In order to extract the explicit form of $d\Psi$ it is sufficient to use the Fourier series for dn given in [36, p. 511]

$$\mathrm{dn}\,(\theta, k^2) = \sum_{l\geq 0} \psi_l \cos(\sqrt{x_l}\,\theta)$$

with

$$\sqrt{x_l} = \frac{l\pi}{K}, \quad l = 0, 1, \ldots, \quad \psi_0 = \frac{\pi}{2K}, \quad \psi_l = \frac{2\pi}{K} \frac{q^l}{1+q^{2l}}, \quad l = 1, 2, \ldots.$$

Substituting this series into (1.5) and integrating term by term gives

$$\int_0^\infty d\Psi(x)\, e^{-tx} = \sum_{l\geq 0} \Psi_l\, e^{-tx_l},$$

which shows that this measure has for support the points x_l with the masses Ψ_l, in agreement with [10, p. 194].

2. For the rates

$$\lambda_n = (2n+1)^2, \quad \mu_n = k^2(2n)^2, \quad 0 < k^2 < 1, \tag{1.6}$$

one has

$$\mathcal{P}_{00}(t) = \int_0^\infty d\Psi(x)\, e^{-tx} = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} d\theta\, \mathrm{cn}\,(\theta, k^2)\, e^{-\theta^2/4t}.$$

Using the Fourier series of cn in [36, p. 511] gives

$$\int_0^\infty d\Psi(x)\, e^{-tx} = \sum_{l\geq 0} \psi_l\, e^{-tx_l}$$

with

$$\sqrt{x_l} = \frac{(l+1/2)\pi}{K}, \quad \psi_l = \frac{2\pi}{kK} \frac{x_l\, q^{l+1/2}}{1+q^{2l+1}}, \quad l = 0, 1, \ldots$$

in agreement with [10, p. 194].

Stieltjes derived these results for the first time using continued fraction theory [29]. Later, Carlitz [8] was able to obtain several explicit solutions of Heun's differential equation, which led him to the orthogonality measures. More recently, we gave a new

derivation [33] of these results by solving linear partial differential equations for the transition probabilities.

However, until now, the derivation of these deep results using asymptotic analysis remained undone. It is clear that to fill this gap we need generating functions for the associated polynomials and this is the main motivation of our work.

In §2 we consider the rates

$$\lambda_n = k^2(2n+2c+1)^2, \qquad \mu_n = 4(n+c)^2 + \mu\delta_{n0}, \tag{1.7}$$

which reduce to the form (1.4) for $(c = 0, \mu = 0)$. The differential equation for $F(x,w)$ remains of Heun's type but is now *inhomogeneous*. A suitable generalization of Carlitz's factorization technique [8] enables us to get a generating function for the polynomials with rates (1.7). From this result their asymptotics and the Stieltjes transform of the orthogonality measure follow.

For $(c = 0, \mu = 0)$ we recover the results of Stieltjes and Carlitz. Another case with $(c = 1/2, \mu = 0)$ appears for which the measure can be given in closed form while for other values of (c, μ) this does not seem to happen; nevertheless, we can give an approximate formula for the spectrum x_n valid for sufficiently large n. On the way to this result we obtain two elliptic generalizations of Euler's beta integral that are markedly different from their q generalization [14, p. 18].

In §3 we deal, along the same lines as in §2, with the case

$$\lambda_n = (2n+2c+1)^2, \quad \mu_n = 4k^2(n+c)^2 + k^2\mu\delta_{n0}, \quad n \geq 0.$$

These rates reduce for $(c = 0, \mu = 0)$ to (1.6). Here too there is a new closed form orthogonality measure for $(c = 1/2, \mu = 0)$.

The two new cases with closed form measures obtained in §§2 and 3,

$$\lambda_n = k^2(2n+2)^2, \qquad \mu_n = (2n+1)^2,$$

and

$$\lambda_n = (2n+2)^2, \qquad \mu_n = k^2(2n+1)^2,$$

do not appear in either Stieltjes' or Carlitz's work. We show at the end of §3 how the corresponding measures can also be deduced by a completely different technique which makes use of the duality transformation of Karlin and MacGregor [21].

In §4 we consider the rates

$$\begin{cases} \lambda_n = (4n+4c+1)(4n+4c+2)^2(4n+4c+3), \\ \mu_n = (4n+4c-1)(4n+4c)^2(4n+4c+1) + \mu\delta_{n0} \end{cases} \tag{1.8}$$

whose Stieltjes moment problem is indeterminate for all values of (c, μ) that ensure the positivity of the rates. We are, therefore, in a very particular situation where the polynomials $F_n(x)$ are orthogonal with respect to infinitely many different measures.

Such a phenomenon is not new: the Stieltjes–Wigert polynomials [10, p. 174] for which

$$\lambda_n = q^{-2n-1}, \quad \mu_n = q^{-2n}(1-q^n), \quad 0 < q < 1$$

give the simplest and oldest example of an indeterminate moment problem for which large classes of orthogonality measures are known [9], [11], [3].

Among all of these measures a one parameter family Ψ_α, labeled by $\alpha \in \mathbb{R} \cup \{\infty\}$, is of greatest importance. It is the family of Nevanlinna extremal measures [1, p. 45], [28, p. 60] whose Stieltjes transform is given by

$$\int_{-\infty}^{+\infty} \frac{d\Psi_\alpha(s)}{x-s} = \frac{\alpha A(x) - C(x)}{\alpha B(x) - D(x)}$$

where A, B, C, and D are entire functions of x constrained by $AD - BC = 1$. A theorem of Riesz ensures that for these and only these measures the polynomials $F_n(x)$ are dense in $L^2(d\Psi_\alpha)$ [1, p. 45].

The computation of the entire functions A, B, C, and D and of the Nevanlinna extremal measures for $(c = \mu = 0)$ will be done elsewhere [7]. Let us point out that such an analysis for the Al-Salam–Chihara polynomials in the indeterminate case has been worked out by Chihara and Ismail [12].

Section 4 is devoted to the asymptotic analysis of the indeterminate Stieltjes moment problem with rates (1.8). We first derive a generating function for the corresponding polynomials. This result is then used to deduce the asymptotics which, according to a result of Berg [6], lead to the Stieltjes transform of a Nevanlinna extremal measure for a definite value of the parameter α.

Closed form measures appear for $(c = \mu = 0)$ and $(c = 1/2, \mu = 0)$. The results in the former case are in agreement with those obtained either in [32] and [33] or in [7], while the latter case appears to be new.

2. First family of associated polynomials. We will denote by $F_n(c, \mu; x)$ the polynomials defined by the second-order recurrence (1.1) with the rates

$$\lambda_n = k^2(2n + 2c + 1)^2, \quad \mu_n = 4(n + c)^2 + \mu\delta_{n0}, \quad n \geq 0, \tag{2.1}$$

where $0 < k^2 < 1$. The new parameters c and μ have the following meanings:

(i) c is an association parameter with $c \geq 0$.

(ii) μ is a co-recursivity parameter restricted by the positivity of μ_0 to $\mu + 4c^2 \geq 0$. The polynomials for which $\mu_0 = 0$ are referred to as zero-related polynomials [19, p. 676].

Let us first prove that for the allowed range of (c, μ) the Hamburger moment problem, and therefore the Stieltjes moment problem, is determined. This is the case if and only if [1, p. 240] the series

$$\sum_{n \geq 1} \pi_n \left(\frac{1}{\mu_1 \pi_1} + \cdots + \frac{1}{\mu_n \pi_n} \right)^2$$

diverges. Due to the positivity of the rates λ_n and μ_n this series is greater than

$$\sum_{n \geq 1} \frac{1}{\mu_n^2 \pi_n},$$

whose divergence is easily ascertained. This proof works also for the general quadratic rates

$$\lambda_n = k^2(n + a)(n + b), \quad \mu_n = (n + c)(n + d) + \mu\delta_{n0}, \quad 0 < k^2 < 1,$$

provided that the rates' positivity is ensured.

It follows that the orthogonality measure is unique and that the Markov theorem can be used to deduce the Stieltjes transform

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = -\lim_{n\to\infty} \frac{1}{\mu_1} \frac{F_{n-1}(c+1,0;x)}{F_n(c,\mu;x)}. \tag{2.2}$$

The convergence is uniform for x in any compact subset of $\mathbb{C}\backslash \mathrm{I}$, where I is the smallest closed interval containing the support of Ψ, provided that the Stieltjes moment problem is *determined*. In Szegö [30, p. 57] and in Chihara [10, p. 90] the technical assumption that the support of Ψ is compact is added, but several proofs are available without such an additional hypothesis [28, p. 110], [34, p. 243], [6].

It is precisely to work out the asymptotic analysis of the polynomials $F_n(c,\mu;x)$ that we need a generating function. To achieve this goal we will generalize the factorization technique of Carlitz [8] to the case under consideration.

Let us first switch from the $F_n(x)$ to the $G_n(x)$ defined by

$$F_0 = G_0, \quad F_n = \frac{G_n}{\mu_1 \cdots \mu_n}, \quad n = 1,2,\ldots. \tag{2.3}$$

This gives for recurrence

$$(\lambda_n + \mu_n - x)G_n(x) = G_{n+1}(x) + \lambda_{n-1}\mu_n G_{n-1}, \qquad n = 1,2,\ldots, \tag{2.4}$$

with the boundary conditions

$$G_0(x) = 1, \qquad G_1(x) = \lambda_0 + \mu_0 - x.$$

This recurrence can be factorized into

$$d_{2n+1} = i\sqrt{x}d_{2n} + \mu_n d_{2n-1}, \tag{2.5}$$

$$d_{2n+2} = i\sqrt{x}d_{2n+1} + \lambda_n d_{2n}, \qquad n = 0,1,\ldots, \tag{2.6}$$

provided that we take

$$G_n(x) = d_{2n}(i\sqrt{x}), \qquad n = 0,1,\ldots. \tag{2.7}$$

In order to secure the boundary conditions we take

$$d_{-1} = \frac{1}{i\sqrt{x}}, \qquad d_0 = 1. \tag{2.8}$$

We take the branch of $\sqrt{x}$, which is positive for positive x, but let us observe that $d_{2n}(i\sqrt{x})$ and $d_{2n+1}(i\sqrt{x})/i\sqrt{x}$ are polynomials in the complex variable x.

It is then easy to show that relations (2.5), (2.6), and (2.7) imply the recurrence (2.4):

$$\begin{aligned} G_{n+1} = d_{2n+2} &= i\sqrt{x}d_{2n+1} + \lambda_n d_{2n} \\ &= i\sqrt{x}(i\sqrt{x}d_{2n} + \mu_n d_{2n-1}) + \lambda_n d_{2n} \\ &= (\lambda_n + \mu_n - x)G_n - \mu_n(d_{2n} - i\sqrt{x}d_{2n-1}) \\ &= (\lambda_n + \mu_n - x)G_n - \lambda_{n-1}\mu_n G_{n-1}, \qquad n = 1,2,\ldots. \end{aligned}$$

Using (2.5), (2.6) and (2.8) one can ascertain that G_0 and G_1 are correctly reproduced.

Let us define two generating functions

$$D_0(x,t) = \sum_{n\geq 0} \frac{t^{2n+2c}}{(2n+2c)!} d_{2n}(i\sqrt{x}), \tag{2.9}$$

$$D_1(x,t) = \sum_{n\geq 0} \frac{t^{2n+2c+1}}{(2n+2c+1)!} d_{2n+1}(i\sqrt{x}) \tag{2.10}$$

with the shorthand notation $\alpha! \equiv \Gamma(\alpha+1)$.

Standard techniques give the differential system

$$(1-k^2t^2)\partial_t D_0 - k^2 t D_0 = i\sqrt{x} D_1 + \frac{t^{2c-1}}{(2c-1)!},$$

$$(1-t^2)\partial_t D_1 - t D_1 = i\sqrt{x} D_0 + \frac{\mu_0}{i\sqrt{x}} \frac{t^{2c}}{2c!}, \qquad \mu_0 = 4c^2 + \mu,$$

and the change of functions

$$D_0 = \frac{\mathcal{D}_0}{\sqrt{1-k^2t^2}}, \qquad D_1 = \frac{\mathcal{D}_1}{\sqrt{1-t^2}}$$

leads to the more symmetric form

$$\sqrt{(1-t^2)(1-k^2t^2)}\partial_t \mathcal{D}_0 = i\sqrt{x}\mathcal{D}_1 + \frac{t^{2c-1}\sqrt{1-t^2}}{(2c-1)!}, \tag{2.11}$$

$$\sqrt{(1-t^2)(1-k^2t^2)}\partial_t \mathcal{D}_1 = i\sqrt{x}\mathcal{D}_0 + \frac{\mu_0}{i\sqrt{x}} \frac{t^{2c-1}\sqrt{1-k^2t^2}}{2c!}. \tag{2.12}$$

At that point the elliptic functions come in since we look for a new variable θ such that

$$d\theta = \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}.$$

Using Jacobian elliptic functions we take

$$t = \operatorname{sn}(\theta, k^2) \quad \Longleftrightarrow \quad \theta(t) = \int_0^t \frac{du}{\sqrt{(1-u^2)(1-k^2u^2)}} \tag{2.13}$$

to simplify (2.11), (2.12) to

$$\partial_\theta \mathcal{D}_0 - i\sqrt{x}\mathcal{D}_1 = \frac{(\operatorname{sn}\theta)^{2c-1}}{(2c-1)!}\operatorname{cn}\theta,$$

$$\partial_\theta \mathcal{D}_1 - i\sqrt{x}\mathcal{D}_0 = \frac{\mu_0}{i\sqrt{x}} \frac{(\operatorname{sn}\theta)^{2c}}{2c!}\operatorname{dn}\theta.$$

The mapping $\theta(t)$ is analytic for $|t| < 1$. Its analytic continuation to the whole complex t plane is given in [2, p. 119]: it has branch points for $t = \pm 1, \pm 1/k$ and it maps conformally the complex t plane into a rectangle with vertices $\pm K, \pm K + iK'$ in the θ plane.

The system is easily integrated to

$$\mathcal{D}_0(\theta) = \int_0^\theta du \cos(\sqrt{x}(\theta-u)) \frac{(\operatorname{sn} u)^{2c-1}}{(2c-1)!} \operatorname{cn} u + \mu_0 \int_0^\theta du \frac{\sin(\sqrt{x}(\theta-u))}{\sqrt{x}} \frac{(\operatorname{sn} u)^{2c}}{2c!} \operatorname{dn} u. \tag{2.14}$$

Using (2.3), (2.7), and (2.9) we get the generating function

$$\sum_{n\geq 0} \frac{(1+c)_n}{(1/2+c)_n} \frac{t^{2n+2c}}{2c!} F_n(c,\mu;x) = \frac{\mathcal{D}_0(\theta(t))}{\sqrt{1-k^2t^2}}, \qquad c>0, \qquad |t|<1. \tag{2.15}$$

For complex t, and therefore complex θ, one should take for $(\operatorname{sn} u)^{2c}$ and t^{2c} the same principal determination in order to ensure that $t^{-2c}\mathcal{D}_0(\theta(t))$ is indeed analytic for $|t|<1$.

The value of $\mathcal{D}_0$ when $c=0$ is well defined and can be reached by a limiting procedure which gives

$$\lim_{c\to 0} \mathcal{D}_0(\theta) = \cos(\sqrt{x}\theta) + \mu \int_0^\theta du \frac{\sin(\sqrt{x}(\theta-u))}{\sqrt{x}} \operatorname{dn} u$$

in agreement with Carlitz's result [8] for $\mu=0$. Technically speaking, D_0 and D_1 are solutions of an *inhomogeneous* Heun differential equation in the variable $w=\operatorname{sn}^2\theta$ which reduces for $c=\mu=0$ to a homogeneous one.

As an application of the generating function (2.15) we extract the asymptotic behavior of the $F_n(c,\mu;x)$ using the Darboux theorem [4, p. 12], [27, p. 309].

We first apply the operator $t\partial_t$ to both sides of (2.15). We get

$$\sum_{n\geq 0} \frac{(1+c)_n}{(1/2+c)_n}(2n+2c)\frac{t^{2n}}{2c!}F_n(c,\mu;x) = \frac{k^2t^{-2c}}{(1-k^2t^2)^{3/2}}\mathcal{D}_0 + \frac{t^{-2c}}{(1-k^2t^2)\sqrt{1-t^2}}\frac{d\mathcal{D}_0}{d\theta},$$

from which we realize that the asymptotic behavior is controlled by the square root singularity at $t^2=1$ of the second term in the right-hand side ($t=0$ is only an apparent singularity). The binomial expansion of the square root gives the asymptotic behavior

$$F_n(c,\mu;x) \sim \frac{2c!}{2k'^2}\frac{1}{n+c}\frac{(1/2)_n(1/2+c)_n}{n!(1+c)_n}\frac{d\mathcal{D}_0}{d\theta} \qquad (\theta=K).$$

In what follows $F_n \sim G_n$ has the precise meaning $\lim_{n\to\infty}(F_n/G_n)=1$. Plugging the asymptotics into (2.2) gives

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = -\frac{H(c+1,0;x)}{H(c,\mu;x)}, \qquad c\geq 0, \qquad \mu+4c^2\geq 0, \tag{2.16}$$

with

$$H(c,\mu;x) = -\sqrt{x}\int_0^K du \sin(\sqrt{x}(K-u))\frac{(\operatorname{sn} u)^{2c-1}}{(2c-1)!}\operatorname{cn} u + (4c^2+\mu)\int_0^K du\cos(\sqrt{x}(K-u))\frac{(\operatorname{sn} u)^{2c}}{2c!}\operatorname{dn} u.$$

If $c>1/2$ an integration by parts gives the simpler form

$$H(c,\mu;x) = \int_0^K du\cos(\sqrt{x}(K-u))\left[\frac{(\operatorname{sn} u)^{2c-2}}{(2c-2)!} + \mu\frac{(\operatorname{sn} u)^{2c}}{2c!}\right]\operatorname{dn} u. \tag{2.17}$$

2.1. Ergodicity and the measure jump at $x = 0$. If $\mu_0 = 0$ we see from (2.16) that $x = 0$ is a simple pole with residue Ψ_0 given by

$$\Psi_0 = \frac{c+1}{c(2c+1)} \frac{\int_0^K du(\mathrm{sn}\, u)^{2c+2}\mathrm{dn}\, u}{\int_0^K du(K-u)(\mathrm{sn}\, u)^{2c-1}\mathrm{cn}\, u}.$$

The change of variable $t = \mathrm{sn}^2 u$ in the numerator gives a beta integral and we get

$$1/\Psi_0 = \frac{4c}{B(1/2, 1/2+c)} \int_0^K du(K-u)(\mathrm{sn}\, u)^{2c-1}\mathrm{cn}\, u. \tag{2.18}$$

The remaining integral is nothing but

$$\int_0^K d\theta \int_0^\theta d\varphi(\mathrm{sn}\, \varphi)^{2c-1}\mathrm{cn}\, \varphi\ ;$$

the change of variables $u = \mathrm{sn}^2\theta$ and $v = u\ \mathrm{sn}^2\varphi$ and use of Euler's integral representation [13, formula 10, p. 59] for the hypergeometric function brings (2.18) to

$$1/\Psi_0 = \frac{1}{B(1/2, 1/2+c)} \int_0^1 du \frac{u^{c-1/2}}{\sqrt{(1-u)(1-k^2u)}}\, {}_2F_1\left(\begin{matrix} 1/2, c \\ 1+c \end{matrix} ; k^2u\right).$$

Using [13, relation 2, p. 105]

$${}_2F_1\left(\begin{matrix} 1/2, c \\ 1+c \end{matrix} ; k^2u\right) = (1-k^2u)^{1/2}\, {}_2F_1\left(\begin{matrix} 1, 1/2+c \\ 1+c \end{matrix} ; k^2u\right)$$

and [15, relation 12, p. 850] eventually gives

$$1/\Psi_0 = {}_3F_2\left(\begin{matrix} 1, 1/2+c, 1/2+c \\ 1+c, 1+c \end{matrix} ; k^2\right), \quad c \geq 0, \quad 0 \leq k^2 < 1. \tag{2.19}$$

This result checks the ergodicity of the process with $\mu_0 = 0$. It has been proved in [21] that if

$$\mu_0 = 0, \quad \sum_{n\geq 0} \pi_n < +\infty, \quad \sum_{n\geq 0} \frac{1}{\lambda_n \pi_n} = +\infty,$$

the corresponding birth and death process is ergodic i.e., we have a discrete mass at $x = 0$ and furthermore

$$\lim_{t\to+\infty} \frac{1}{\pi_m} \int_0^\infty d\Psi(x) e^{-xt} F_m(x) F_n(x) = \pi_n \left(\sum_{l\geq 0} \pi_l\right)^{-1}. \tag{2.20}$$

Now the large time behavior is controlled by the mass at $x = 0$ whose jump is Ψ_0. From (2.20) we conclude with

$$1/\Psi_0 = \sum_{l\geq 0} \pi_l.$$

Since we have

$$\pi_n = (k^2)^n \left[\frac{(1/2+c)_n}{(1+c)_n}\right]^2$$

this gives a check of the relation (2.19).

2.2. The case $c < 0$ and a finite process. Until now we have supposed $c > 0$. It happens that, for negative values of c, relation (2.16) remains valid provided that we use Hadamard regularization of these integrals at $u = 0$ (see, for instance, [4, p. 45]). Such a technicality is necessary if one is willing to use the continued fraction to study the finite population case with rates

$$\lambda_n = 4k^2(n-p)^2, \quad \mu_n = (2n-2p-1)^2 - (2p+1)^2\delta_{n0}, \quad p = 1, 2, \ldots.$$

Since $\mu_0 = 0$ and $\lambda_p = 0$ we have a *finite* birth and death process whose population n evolves between 0 and p. It follows [35, p. 87] that the orthogonality measure has for support finitely many points on the positive real axis, including $x = 0$.

Using Hadamard regularization we have

$$\lim_{c \to -p} \int_0^a du f(u) \frac{(\operatorname{sn} u)^c}{c!} = (-1)^{p-1} f^{(p-1)}(0)$$

valid for $a > 0$ and provided that $f(u)$ is C^∞ for $u \in [0, a]$. (The superscript indicates the order of derivation with respect to the variable u.)

It follows that for this finite process we have

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = \frac{-\sqrt{x}\left[\sin(\sqrt{x}u)\operatorname{cn} u\right]^{(2p-1)}(0) + (2p-1)^2\left[\cos(\sqrt{x}u)\operatorname{dn} u\right]^{(2p-2)}(0)}{\sqrt{x}\left[\sin(\sqrt{x}u)\operatorname{cn} u\right]^{(2p+1)}(0)}.$$

Defining the polynomials $\rho_l(k^2)$ by

$$\operatorname{cn} u = \sum_{l \geq 0} (-1)^l \rho_l(k^2) \frac{u^{2l}}{2l!}$$

we obtain an explicit form of the algebraic equation for the support of the measure

$$\sqrt{x}\left[\sin(\sqrt{x}u)\operatorname{cn} u\right]^{(2p+1)}(0) \equiv (-1)^p(2p+1)!\, x \sum_{l=0}^{p} \frac{\rho_l(k^2)}{2l!} \frac{x^{p-l}}{[2(p-l)+1]!} = 0,$$

which gives $x = 0$ as expected. The remaining roots will be real and positive but it seems rather difficult to get any explicit form for them.

2.3. Two processes with closed form orthogonality measure. There are nevertheless two cases for which everything simplifies to a closed result; the first one corresponds to $c = \mu = 0$. Relation (2.16) gives for the Stieltjes transform

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = \frac{1}{\sqrt{x}\sin(K\sqrt{x})} \int_0^K du \cos(\sqrt{x}(K-u))\operatorname{dn} u.$$

Both numerator and denominator of this fraction are entire functions of x. Its singularities can only be simple poles given by the zeros of the denominator

$$\sqrt{x}\sin(K\sqrt{x}) = 0.$$

The measure Ψ has therefore the discrete support

$$x_n = \left(\frac{n\pi}{K}\right)^2, \qquad n = 0, 1, \ldots,$$

and the masses Ψ_n are the residues at the poles. An easy computation gives

$$\psi_n = \frac{1}{K}\int_0^{2K} du \exp\left(-i\frac{n\pi u}{K}\right) \mathrm{dn}\, u, \qquad n = 0, 1, \ldots,$$

which are the Fourier coefficients of $\mathrm{dn}\, u$ already given in the introduction. We conclude with the following theorem.

THEOREM 1 (Stieltjes–Carlitz). *For $\lambda_n = k^2(2n+1)^2$ and $\mu_n = 4n^2$ the corresponding polynomials $F_n(0,0;x)$ are orthogonal:*

$$\sum_{l\geq 0} \psi_l F_m(0,0;x_l) F_n(0,0;x_l) = k^{2n}\left(\frac{(1/2)_n}{n!}\right)^2 \delta_{mn}$$

with

$$x_l = \left(\frac{l\pi}{K}\right)^2, \quad l = 0, 1, \ldots, \quad \psi_0 = \frac{\pi}{2K}, \quad \psi_l = \frac{2\pi}{K}\frac{q^l}{1+q^{2l}}, \quad l = 1, 2, \ldots.$$

The second case where everything simplifies to a closed form result is ($c = 1/2, \mu = 0$). Integrations by parts lead to the final form of the continued fraction

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = -1 - \frac{\sqrt{x}}{\cos(\sqrt{x}K)}\int_0^K du \sin(\sqrt{x}(K-u))\mathrm{cn}\, u.$$

Here too the measure is discretely supported:

$$\cos(K\sqrt{x}) = 0 \quad \Longrightarrow \quad x_n = \left(\frac{(n+1/2)\pi}{K}\right)^2, \qquad n = 0, 1, \ldots,$$

and the masses

$$\psi_n = \frac{x_n}{K}\int_0^{2K} du \exp\left(-i(n+1/2)\frac{\pi u}{K}\right)\mathrm{cn}\, u$$

are proportional to the Fourier coefficients of $\mathrm{cn}\, u$ given in the introduction. We conclude with the following theorem.

THEOREM 2. *For $\lambda_n = 4k^2(n+1)^2$ and $\mu_n = (2n+1)^2$ the corresponding polynomials $F_n(1/2,0;x)$ are orthogonal:*

$$\sum_{l\geq 0} \psi_l F_m(1/2,0;x_l) F_n(1/2,0;x_l) = k^{2n}\left(\frac{n!}{(3/2)_n}\right)^2 \delta_{mn}$$

with

$$x_l = \left((l+1/2)\frac{\pi}{K}\right)^2, \quad \psi_l = \frac{2\pi}{kK}\frac{x_l\, q^{l+1/2}}{1+q^{2l+1}}, \quad l = 0, 1, \ldots.$$

2.4. Elliptic generalizations of the beta integral. We come back to the relation (2.17), which can be written

$$H(c,\mu;x)=\tfrac{1}{2}\left(B(c-1,\sqrt{x};k^2)+\mu B(c,\sqrt{x};k^2)\right),\quad c>\tfrac{1}{2}$$

using the notation

$$B(c,\sqrt{x};k^2)=\int_0^{2K}du\,\frac{(\operatorname{sn}u)^{2c}}{2c!}\operatorname{dn}u\;e^{i\sqrt{x}(u-K)}. \tag{2.21}$$

The change of variable $\theta=\operatorname{am}u$ brings it to the form

$$B(c,\sqrt{x};k^2)=\int_0^{\pi}d\theta\frac{(\sin\theta)^{2c}}{2c!}\exp\left[i\sqrt{x}\left(\int_0^{\theta}\frac{du}{\sqrt{1-k^2\sin^2u}}-K\right)\right],$$

on which we recognize an *elliptic generalization* of Euler's beta integral, which is recovered in the limit $k^2=0$. Indeed, for $k^2=0$, we have

$$B(c,\sqrt{x};0)=\int_0^{\pi}d\theta\frac{(\sin\theta)^{2c}}{2c!}\exp\left(i\sqrt{x}(\theta-\pi/2)\right),$$

which is computed using [26, p. 8]

$$B(c,\sqrt{x};0)=\frac{\pi}{2^{2c}}\frac{1}{\Gamma(1+c+\sqrt{x}/2)\Gamma(1+c-\sqrt{x}/2)}. \tag{2.22}$$

Therefore, when $k^2=0$, we have the functional relation

$$[4c^2-x]B(c,\sqrt{x};0)=B(c-1,\sqrt{x};0),\qquad c>0,$$

which is a first-order recurrence relation in c. For $k^2>0$ this relation becomes a second-order recurrence which can be derived from (2.21) upon integrations by parts for $c>0$:

$$\begin{aligned}[k^2(2c+1)^2+4c^2-x]B(c,\sqrt{x};k^2)&=B(c-1,\sqrt{x};k^2)\\&\quad+k^2(2c+1)^2(2c+2)^2B(c+1,\sqrt{x};k^2).\end{aligned} \tag{2.23}$$

Another integral of interest related to B, which will be useful in the next section, is

$$\mathcal{B}(c,\sqrt{x};k^2)=\int_0^{2K}du\frac{(\operatorname{sn}u)^{2c}}{2c!}\operatorname{cn}u\;e^{i\sqrt{x}(u-K)}. \tag{2.24}$$

These two integrals are connected by the relations

$$i\sqrt{x}B(c,\sqrt{x};k^2)=k^2(2c+1)^2\mathcal{B}(c+1/2,\sqrt{x};k^2)-\mathcal{B}(c-1/2,\sqrt{x};k^2), \tag{2.25}$$

$$i\sqrt{x}\mathcal{B}(c,\sqrt{x};k^2)=(2c+1)^2B(c+1/2,\sqrt{x};k^2)-B(c-1/2,\sqrt{x};k^2), \tag{2.26}$$

valid for $c>0$, and from which (2.23) follows as well as

$$\begin{aligned}[(2c+1)^2+4k^2c^2-x]\mathcal{B}(c,\sqrt{x};k^2)&=\mathcal{B}(c-1,\sqrt{x};k^2)\\&\quad+k^2(2c+1)^2(2c+2)^2\mathcal{B}(c+1,\sqrt{x};k^2),\end{aligned}$$

valid if $c > 1/2$.

Let us use the preceding analysis to get an approximate form of the measure support x_n for $c > 1/2$. It is given by the zeros of $H(c, \mu; x)$.

For $k^2 = 0$ relation (2.22) gives

$$x_n = 4(n+c)^2 + \mu\delta_{n0} = \mu_n, \qquad n = 0, 1, \ldots.$$

To improve this result we use the Fourier series [25, relation 2, p. 35]

$$\int_0^\theta \frac{du}{\sqrt{1-k^2\sin^2 u}} = \frac{2K}{\pi}\theta + \sum_{l\geq 1} \frac{(1/2)_l}{l\,l!}(-k^2/4)^l {}_2F_1\left(\begin{matrix} l+1/2, l+1/2 \\ 2l+1 \end{matrix}; k^2\right)\sin(2l\theta) \tag{2.27}$$

and we keep only the first term. It follows that $\sqrt{x}$ is rescaled to $\frac{2K}{\pi}\sqrt{x}$ and the spectrum becomes

$$x_n = \left(\frac{\pi}{2K}\right)^2 \mu_n, \qquad n = 0, 1, \ldots.$$

This approximation becomes exact only for the two cases covered by Theorems 1 and 2. Taking further terms in (2.27) leads to complicated formulas of little interest. One can ascertain the relation

$$x_n - \left(\frac{\pi}{2K}\right)^2 \mu_n = \mathcal{O}(k^{2n}),$$

which indicates that we have obtained a good approximation to the spectrum for n sufficiently large.

Let us observe, to conclude, that if we consider the zero-related polynomials, i.e., those with $\mu = -4c^2$, the continued fraction

$$S(x) = \int_0^\infty \frac{d\Psi(s)}{x-s} = -\frac{B(c, \sqrt{x}; k^2)}{B(c-1, \sqrt{x}; k^2) + \mu B(c, \sqrt{x}; k^2)} \tag{2.28}$$

(recall that here $c > 1/2$) can be reduced, using (2.25), (2.26), to the form

$$xS(x) = 1 - k^2(2c+1)^2 \frac{\mathcal{B}(c+1/2, \sqrt{x}; k^2)}{\mathcal{B}(c-1/2, \sqrt{x}; k^2)}, \tag{2.29}$$

which implies an exact relation between some spectra of §§2 and 3, to be discussed below.

3. Second family of associated polynomials. We shall denote by $\mathcal{F}_n(c, \mu; x)$ (respectively, $\mathcal{G}_n(c, \mu; x)$) the polynomials defined by the recurrence (1.1) (respectively, (2.4)) with the rates

$$\lambda_n = (2n+2c+1)^2, \quad \mu_n = 4k^2(n+c)^2 + k^2\mu\delta_{n0}, \quad n \geq 0,$$

with $0 < k^2 < 1$.

Let us first prove that both Hamburger and Stieltjes moment problems are determined. This follows from

$$\sum_{n\geq 1} \pi_n \left(\frac{1}{\mu_1\pi_1} + \cdots + \frac{1}{\mu_n\pi_n}\right)^2 \geq \frac{1}{\mu_1^2\pi_1^2} \sum_{n\geq 1} \pi_n$$

and from the large n behavior of π_n. This proof remains valid for the more general quadratic case

$$\lambda_n = (n+a)(n+b), \quad \mu_n = k^2(n+c)(n+d) + \mu\delta_{n0}, \quad 0 < k^2 < 1$$

if positivity is ensured.

Using the same factorization technique as in §2 and the definitions

$$C_0(x,t) = \sum_{n\geq 0} \frac{t^{2n+2c}}{(2n+2c)!} d_{2n}(i\sqrt{x}),$$
$$C_1(x,t) = \sum_{n\geq 0} \frac{t^{2n+2c+1}}{(2n+2c+1)!} d_{2n+1}(i\sqrt{x}),$$

we get the differential system

$$(1-t^2)\partial_t C_0 - k^2 t C_0 = i\sqrt{x} C_1 + \frac{t^{2c-1}}{(2c-1)!},$$
$$(1-k^2t^2)\partial_t C_1 - tC_1 = i\sqrt{x} C_0 + \frac{\mu_0}{i\sqrt{x}} \frac{t^{2c}}{2c!}, \qquad \mu_0 = k^2(4c^2+\mu).$$

We define

$$C_0 = \frac{\mathcal{C}_0}{\sqrt{1-t^2}}, \qquad C_1 = \frac{\mathcal{C}_1}{\sqrt{1-k^2t^2}},$$

and switch to the variable θ defined by (2.13):

$$\partial_\theta \mathcal{C}_0 - i\sqrt{x}\mathcal{C}_1 = \frac{(\operatorname{sn}\theta)^{2c-1}}{(2c-1)!} \operatorname{dn}\theta,$$
$$\partial_\theta \mathcal{C}_1 - i\sqrt{x}\mathcal{C}_0 = \frac{\mu_0}{i\sqrt{x}} \frac{(\operatorname{sn}\theta)^{2c}}{2c!} \operatorname{cn}\theta.$$

This leads eventually to

$$\mathcal{C}_0(\theta) = \int_0^\theta du \cos(\sqrt{x}(\theta-u)) \frac{(\operatorname{sn} u)^{2c-1}}{(2c-1)!} \operatorname{dn} u + \mu_0 \int_0^\theta du \frac{\sin(\sqrt{x}(\theta-u))}{\sqrt{x}} \frac{(\operatorname{sn} u)^{2c}}{2c!} \operatorname{cn} u \tag{3.1}$$

and therefore to the generating function

$$\sum_{n\geq 0} k^{2n} \frac{(1+c)_n}{(1/2+c)_n} \frac{t^{2n+2c}}{2c!} \mathcal{F}_n(c,\mu;x) = \frac{\mathcal{C}_0(\theta(t))}{\sqrt{1-t^2}}, \quad c>0, \quad |t|<1.$$

The limit $c \to 0$ gives here

$$\lim_{c\to 0} \mathcal{C}_0(\theta) = \cos(\sqrt{x}\theta) + k^2\mu \int_0^\theta du \frac{\sin(\sqrt{x}(\theta-u))}{\sqrt{x}} \operatorname{cn} u$$

in agreement with Carlitz's result for $\mu = 0$ [8].

These results can be checked in a different way. Indeed, using the recurrence relations, one can prove

$$\mathcal{G}_n(c,\mu;x;k^2) = (k^2)^n\, G_n\left(c,\mu;\frac{x}{k^2};\frac{1}{k^2}\right).$$

From this we deduce

$$C_0(x,t;k^2) = \frac{1}{(k^2)^c} D_0\left(\frac{x}{k^2}, kt; \frac{1}{k^2}\right).$$

Using the transformation theory of elliptic functions

$$\begin{aligned}
\operatorname{sn}(kx;1/k^2) &= k\,\operatorname{sn}(x;k^2),\\
\operatorname{cn}(kx;1/k^2) &= \operatorname{dn}(x;k^2),\\
\operatorname{dn}(kx;1/k^2) &= \operatorname{cn}(x;k^2),
\end{aligned}$$

one can see that (2.14) implies (3.1).

The asymptotics follow from Darboux's theorem

$$\mathcal{F}_n(c,\mu;x) \sim \frac{1}{k^{2n}} \frac{(1/2)_n (1/2+c)_n}{n!(1+c)_n} \mathcal{H}(c,\mu;x)$$

with

$$\begin{aligned}
\mathcal{H}(c,\mu;x) = \int_0^K du \cos(\sqrt{x}(K-u)) \frac{(\operatorname{sn} u)^{2c-1}}{(2c-1)!} \operatorname{dn} u \\
+k^2(4c^2+\mu) \int_0^K du \frac{\sin(\sqrt{x}(K-u))}{\sqrt{x}} \frac{(\operatorname{sn} u)^{2c}}{2c!} \operatorname{cn} u.
\end{aligned}$$

Since the Stieltjes moment problem is determined, the same arguments as in §2 give

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = -\frac{\mathcal{H}(c+1,0;x)}{\mathcal{H}(c,\mu;x)}$$

uniformly for $x \in \mathbb{C}\backslash \mathrm{I}$. This reduces to

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = -\frac{\mathcal{H}(c+1,0;x)}{\mathcal{H}(c,\mu;x)}, \tag{3.2}$$

valid for $c \geq 0$ and $\mu + 4c^2 \geq 0$. The extension of these results to arbitrary negative values of c should be done along the same lines as in the previous section.

For $c \geq 1/2$ this can be reduced to

$$\mathcal{H}(c,\mu;x) = \frac{1}{2i\sqrt{x}} \int_0^{2K} du \left[\frac{(\operatorname{sn} u)^{2c-2}}{(2c-2)!} + k^2\mu \frac{(\operatorname{sn} u)^{2c}}{2c!}\right] \operatorname{cn} u\; e^{i\sqrt{x}(u-K)},$$

i.e., in terms of the elliptic generalization of the beta integral (2.24)

$$\mathcal{H}(c,\mu;x) = \frac{1}{2i\sqrt{x}} \left[\mathcal{B}(c-1,\sqrt{x};k^2) + k^2\mu \mathcal{B}(c,\sqrt{x};k^2)\right].$$

It follows that the continued fraction is

$$\mathcal{S}(x) = \int_0^\infty \frac{d\Psi(s)}{x-s} = -\frac{\mathcal{B}(c,\sqrt{x};k^2)}{\mathcal{B}(c-1,\sqrt{x};k^2) + k^2\mu \mathcal{B}(c,\sqrt{x};k^2)}. \tag{3.3}$$

Here too, for the zero-related polynomials (i.e., $\mu = -4c^2$), use of relations (2.25), (2.26) gives

$$x\mathcal{S}(x) = 1 - (2c+1)^2 \frac{B(c+1/2, \sqrt{x}; k^2)}{B(c-1/2, \sqrt{x}; k^2)}, \tag{3.4}$$

which will be discussed later on.

3.1. Two processes with closed form orthogonality measure. For $c = \mu = 0$ the continued fraction (3.2) reduces to

$$-\frac{1}{\cos(K\sqrt{x})} \int_0^K du \frac{\sin(\sqrt{x}(K-u))}{\sqrt{x}} \operatorname{cn} u,$$

and by the same token which led to Theorem 2 we get the following theorem.

THEOREM 3 (Stieltjes–Carlitz). *For $\lambda_n = (2n+1)^2$ and $\mu_n = 4k^2n^2$ the corresponding polynomials $\mathcal{F}_n(0,0;x)$ are orthogonal:*

$$\sum_{l\geq 0} \psi_l \mathcal{F}_m(0,0;x_l)\mathcal{F}_n(0,0;x_l) = \frac{1}{k^{2n}} \left(\frac{(1/2)_n}{n!}\right)^2 \delta_{mn}$$

with

$$x_l = \left((l+1/2)\frac{\pi}{K}\right)^2, \quad \psi_l = \frac{2\pi}{kK} \frac{q^{l+1/2}}{1+q^{2l+1}}, \quad l = 0, 1, \ldots.$$

The remaining case where everything simplifies corresponds to $(c = 1/2, \mu = 0)$. The continued fraction is then

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = \frac{k'}{k^2} - \frac{\sqrt{x}}{k^2 \sin(\sqrt{x}K)} \int_0^K du \cos(\sqrt{x}(K-u)) \operatorname{dn} u.$$

The technique already used for Theorem 1 leads to Theorem 4.

THEOREM 4. *For $\lambda_n = 4(n+1)^2$ and $\mu_n = k^2(2n+1)^2$ the corresponding polynomials $\mathcal{F}_n(1/2,0;x)$ are orthogonal:*

$$\sum_{l\geq 0} \psi_l \mathcal{F}_m(1/2,0;x_l)\mathcal{F}_n(1/2,0;x_l) = \frac{1}{k^{2n}} \left(\frac{n!}{(3/2)_n}\right)^2 \delta_{mn}$$

with

$$x_l = \left(\frac{l\pi}{K}\right)^2, \quad \psi_l = \frac{2\pi}{k^2K} \frac{x_l\, q^l}{1+q^{2l+1}}, \quad l = 1, 2, \ldots.$$

3.2. Duality transformation. The duality transformation has been defined by Karlin and MacGregor [21, p. 384] and gives a useful tool to relate processes with different rates. Starting from

$$\lambda_n, \quad \mu_n, \quad n \geq 0, \quad \mu_0 = 0$$

with orthogonality measure $\Psi(x)$, its dual process has the rates

$$\lambda_n^* = \mu_{n+1}, \qquad \mu_n^* = \lambda_n, \qquad n \geq 0$$

with orthogonality measure $\Psi^*(x)$. Clearly, for the dual process we cannot have $\mu_0^* = 0$.

It has been proved in [22, Lemma 3, p. 504] that, provided we deal with *determined processes* (for which the Stieltjes moment problem is determined), the measures are related by

$$d\Psi^*(x) = \frac{x}{\lambda_0}\, d\Psi(x). \tag{3.5}$$

This result gives us the opportunity of checking the results of §2 against those of §3.

Let us first consider the rates of §2. In order to have $\mu_0 = 0$ we restrict ourselves to the zero-related polynomials with rates

$$\lambda_n = k^2(2n+2c+1)^2, \quad \mu_n = 4(n+c)^2 - 4c^2\delta_{n0}, \quad n \geq 0,$$

whose Stieltjes function given by (2.29) is

$$\frac{x}{\lambda_0} S(x) = \frac{1}{k^2(2c+1)^2} - \frac{\mathcal{B}(c+1/2, \sqrt{x}; k^2)}{\mathcal{B}(c-1/2, \sqrt{x}; k^2)}. \tag{3.6}$$

Its dual process has for rates

$$\lambda_n^* = 4(n+c+1)^2, \quad \mu_n^* = k^2(2n+2c+1)^2, \quad n \geq 0,$$

whose Stieltjes function is given by (3.3), in which we substitute

$$\mu \to 0, \qquad c \to c + 1/2.$$

and reads

$$\mathcal{S}^*(x) = -\frac{\mathcal{B}(c+1/2, \sqrt{x}; k^2)}{\mathcal{B}(c-1/2, \sqrt{x}; k^2)}. \tag{3.7}$$

Comparing (3.6) and (3.7) shows that relation (3.5) holds. In particular, for $c = 0$ we see that the rates of Theorem 1,

$$\lambda_n = k^2(2n+1)^2, \qquad \mu_n = 4n^2,$$

are dual to those of Theorem 4,

$$\lambda_n = 4(n+1)^2, \qquad \mu_n = k^2(2n+1)^2.$$

Similarly, if we start from the rates of §3 corresponding to the zero-related polynomials

$$\lambda_n = (2n+2c+1)^2, \quad \mu_n = 4k^2(n+c)^2 - 4k^2c^2\delta_{n0}, \quad n \geq 0,$$

their Stieltjes function (see (3.4)) is

$$\frac{x}{\lambda_0} \mathcal{S}(x) = \frac{1}{(2c+1)^2} - \frac{B(c+1/2, \sqrt{x}; k^2)}{B(c-1/2, \sqrt{x}; k^2)}. \tag{3.8}$$

Its dual process is

$$\lambda_n^* = 4k^2(n+c+1)^2, \quad \mu_n^* = (2n+2c+1)^2, \quad n \geq 0,$$

whose Stieltjes function follows from (2.28) with the substitutions

$$\mu \to 0, \qquad c \to c + 1/2$$

and reads

$$S^*(x) = -\frac{B(c+1/2, \sqrt{x}; k^2)}{B(c-1/2, \sqrt{x}; k^2)}. \tag{3.9}$$

Comparing (3.8) and (3.9) checks therefore with relation (3.5). Here too, for $c = 0$, we observe that the rates of Theorem 3,

$$\lambda_n = (2n+1)^2, \qquad \mu_n = 4k^2n^2,$$

are dual to those of Theorem 2,

$$\lambda_n = 4k^2(n+1)^2, \qquad \mu_n = (2n+1)^2.$$

4. Associated polynomials for a quartic birth and death process. Let us first consider the transition rates

$$\lambda_n = (n+a_1)\cdots(n+a_4), \qquad \mu_n = (n+b_1)\cdots(n+b_4) + \mu\delta_{n0}.$$

We suppose that the positivity constraints are fulfilled. The corresponding Stieltjes moment problem is determined if and only if [1, p. 237]

$$\sum_{n\geq 1}\left(\pi_n + \frac{1}{\mu_n\pi_n}\right) = \infty,$$

with here

$$\pi_n = \frac{(a_1)_n\cdots(a_4)_n}{(1+b_1)_n\cdots(1+b_4)_n} \underset{n\to\infty}{\sim} n^{\Delta-4}, \qquad \Delta = \sum_{i=1}^{4}(a_i - b_i).$$

For the Hamburger moment problem we use the necessary and sufficient condition stated in §2. The analysis of the series involved is elementary and leads to the following possibilities:

1. det H $\Rightarrow$ det S: $\Delta \geq 3$ or $\Delta \leq -1$,
2. indet S $\Rightarrow$ indet H: $1 < \Delta < 3$,
3. indet H, det S: $-1 < \Delta \leq 1$.

In this section we deal with the polynomials $\mathfrak{F}_n(c, \mu; x)$ defined by the second-order recurrence (1.1) with the rates

$$\begin{cases} \lambda_n = (4n+4c+1)(4n+4c+2)^2(4n+4c+3), \\ \mu_n = (4n+4c-1)(4n+4c)^2(4n+4c+1) + \mu\delta_{n0} \end{cases} \tag{4.1}$$

under the positivity constraints $c \geq 0$ and $\mu + 4c^2(4c^2-1) \geq 0$. Since we have for these rates $\Delta = 2$ the Stieltjes and the Hamburger moment problems are always indeterminate. Therefore, as opposed to the quadratic cases already encountered, here we have to deal with a nonunique orthogonality measure.

In the particular case where $c = \mu = 0$ a one parameter family of orthogonality measures has already been obtained in [32] and [33] and we quote it for the reader's convenience.

THEOREM 5. *For the transition rates* (4.1) *with* $c = \mu = 0$, *one has the orthogonality relation*

$$\sum_{l\geq 0} \psi_l \mathfrak{F}_m(0,0;x_l)\mathfrak{F}_n(0,0;x_l) = \frac{1}{4m+1}\left(\frac{(1/2)_m}{m!}\right)^2 \delta_{mn}$$

with

$$x_l = \left(\frac{l\pi}{K_0}\right)^4,\ l = 0,1,\ldots,\ \Psi_0 = \frac{(1+a)\pi}{2K_0^2},\ \psi_l = \frac{2\pi^2}{K_0^2}\frac{l(1+a(-1)^l)}{\sinh(l\pi)},\ l = 1,2,\ldots.$$

The parameter a *is restricted by* $-1 \leq a \leq +1$. *The honest measure corresponds to* $a = +1$.

The constant K_0 is the period of the lemniscate functions (elliptic functions with modulus $k^2 = 1/2$):

$$K_0 = K(k^2 = 1/2) = \frac{\Gamma^2(1/4)}{4\sqrt{\pi}}.$$

It is therefore of interest to compare this result with the one given by asymptotic analysis. Since there are infinitely many orthogonality measures and since, as we shall see below, asymptotic analysis gives just one of them, how is this particular measure characterized?

The answer can be found in [6]; for an indeterminate Hamburger and Stieltjes moment problem the orthogonality measure given by

$$-\lim_{n\to\infty}\frac{1}{\mu_1}\frac{\mathfrak{F}_{n-1}(c+1,0;x)}{\mathfrak{F}_n(c,\mu;x)} = \int_{-\infty}^{\infty}\frac{d\Psi(s)}{x-s}$$

is a Nevanlinna extremal measure corresponding to a particular value of the parameter α defined in the introduction; this value is

$$\frac{1}{\alpha} = -\sum_{n\geq 1}\frac{1}{\mu_n\pi_n}$$

and we have $\mathrm{supp}(\Psi)\subseteq [0,+\infty[$.

In order to work out all the details we need a generating function for the polynomials $\mathfrak{F}_n(c,\mu;x)$ defined by the second-order recurrence (1.1) with the rates (4.1).

The basic idea to solve this recurrence for quadratic rates was a two-stepped factorization; this generalizes to quartic rates with a four-stepped factorization of the form

$$\begin{aligned}
d_{4n+1} &= j\rho d_{4n},\\
d_{4n+2} &= j\rho d_{4n+1} + \mu_n\, d_{4n-2},\\
d_{4n+3} &= j\rho d_{4n+2},\\
d_{4n+4} &= j\rho d_{4n+3} + \lambda_n\, d_{4n}, \qquad n = 0,1,\ldots,
\end{aligned}$$

with

$$j = \exp(i\pi/4), \qquad \rho = x^{1/4}.$$

For complex x we take for ρ the determination which is real positive when x is real positive; but it should be observed that d_{4n} is a polynomial in x. We impose

$$G_n(x) = d_{4n}(jx^{1/4}), \qquad n = 0, 1, \ldots, \tag{4.2}$$

and the boundary conditions

$$d_{-2} = 1/(j\rho)^2, \qquad d_0 = 1. \tag{4.3}$$

Let us check that the relation (4.2) gives the solution of the recurrence (2.4). We have

$$\begin{aligned} G_{n+1} = d_{4n+4} &= j\rho\, d_{4n+3} + \lambda_n\, d_{4n} = (j\rho)^2 d_{4n+2} + \lambda_n G_n \\ &= (j\rho)^3 d_{4n+1} + (j\rho)^2 \mu_n d_{4n-2} + \lambda_n G_n \\ &= (\lambda_n - x) G_n + j\rho\mu_n d_{4n-1} \\ &= (\lambda_n + \mu_n - x) G_n + \mu_n (j\rho d_{4n-1} - d_{4n}) \\ &= (\lambda_n + \mu_n - x) G_n - \lambda_{n-1}\mu_n\, G_{n-1}, \qquad n = 0, 1, \ldots, \end{aligned}$$

and (4.3) implies

$$G_0 = 1, \qquad G_1 = \lambda_0 + \mu_0 - x.$$

Now we define four generating functions

$$D_l(x,t) = \sum_{n \geq 0} \frac{t^{4n+4c+l}}{(4n+4c+l)!} d_{4n+l}, \qquad l = 0, 1, 2, 3. \tag{4.4}$$

Standard techniques give the differential system

$$\begin{aligned} (1-t^4)\partial_t D_0 - 2t^3 D_0 - j\rho D_3 &= \frac{t^{4c-1}}{(4c-1)!}, \\ \partial_t D_1 - j\rho D_0 &= 0, \\ (1-t^4)\partial_t D_2 - 2t^3 D_2 - j\rho D_1 &= \frac{\mu_0}{i\sqrt{x}} \frac{t^{4c+1}}{(4c+1)!}, \\ \partial_t D_3 - j\rho D_2 &= 0, \qquad \mu_0 = 4c^2(4c^2-1) + \mu, \end{aligned}$$

which upon the change of functions

$$D_0 = \frac{\mathcal{D}_0}{\sqrt{1-t^4}}, \qquad D_1 = \mathcal{D}_1, \qquad D_2 = \frac{\mathcal{D}_2}{\sqrt{1-t^4}}, \qquad D_3 = \mathcal{D}_3 \tag{4.5}$$

simplifies to

$$\begin{aligned} \sqrt{1-t^4}\, \partial_t \mathcal{D}_0 - j\rho \mathcal{D}_3 &= \frac{t^{4c-1}}{(4c-1)!}, \\ \sqrt{1-t^4}\, \partial_t \mathcal{D}_1 - j\rho \mathcal{D}_0 &= 0, \\ \sqrt{1-t^4}\, \partial_t \mathcal{D}_2 - j\rho \mathcal{D}_1 &= \frac{\mu_0}{i\sqrt{x}} \frac{t^{4c+1}}{(4c+1)!}, \\ \sqrt{1-t^4}\, \partial_t \mathcal{D}_3 - j\rho \mathcal{D}_2 &= 0. \end{aligned}$$

At this point lemniscate elliptic functions appear when we define [36, p. 524]

$$t(\theta) = \frac{1}{\sqrt{2}} \operatorname{sd}(\sqrt{2}\theta) \qquad \longleftrightarrow \qquad \theta(t) = \int_0^t \frac{du}{\sqrt{1-u^4}}$$

(in what follows we deal exclusively with lemniscate functions). For complex t the function θ is analytic, but for the branch points, $t = \pm 1, \pm i$.

This change of variable trivializes the system to

$$\begin{aligned}
&\partial_\theta \mathcal{D}_0 - j\rho \mathcal{D}_3 = a(\theta), && a(\theta) = \frac{t(\theta)^{4c-1}}{(4c-1)!},\\
&\partial_\theta \mathcal{D}_1 - j\rho \mathcal{D}_0 = 0,\\
&\partial_\theta \mathcal{D}_2 - j\rho \mathcal{D}_1 = b(\theta), && b(\theta) = \frac{\mu_0}{i\sqrt{x}} \frac{t(\theta)^{4c+1}}{(4c+1)!},\\
&\partial_\theta \mathcal{D}_3 - j\rho \mathcal{D}_2 = 0.
\end{aligned}$$

We define four solutions of the homogeneous system by

$$\delta_l(\rho\theta) = \sum_{n\geq 0} \frac{(j\rho\theta)^{4n+l}}{(4n+l)!}, \qquad l = 0, 1, 2, 3,$$

from which we deduce

$$\mathcal{D}_1(\theta) = \int_0^\theta du \delta_1(\rho(\theta-u)) a(u) + \int_0^\theta du \delta_3(\rho(\theta-u)) b(u).$$

Observing that

$$\frac{\mu_1 \cdots \mu_n}{(4n+4c+1)!} = \frac{1}{(4c+1)!} \frac{(1+c)_n}{(1/2+c)_n} \tag{4.6}$$

and using relations (4.4), (4.5), and (4.6) we obtain the generating function

$$\mathcal{D}_1(\theta(t)) = \sum_{n\geq 0} \frac{(1+c)_n}{(1/2+c)_n} \frac{t^{4n+4c+1}}{(4c+1)!} \mathfrak{F}_n(c, \mu; x), \quad c \geq 0, \quad |t| < 1, \tag{4.7}$$

where

$$\mathcal{D}_1(\theta(t)) = \int_0^{\theta(t)} du \frac{\delta_1(\rho(\theta(t)-u))}{j\rho} \frac{t(u)^{4c-1}}{(4c-1)!} + \mu_0 \int_0^{\theta(t)} du \frac{\delta_3(\rho(\theta(t)-u))}{(j\rho)^3} \frac{t(u)^{4c+1}}{(4c+1)!}$$

with $t(u) = \frac{1}{\sqrt{2}} \operatorname{sd}(\sqrt{2}u)$ and $\rho = x^{1/4}$.

The remarks that follow relation (2.15) in §2 should be kept in mind: the determinations of $\mathcal{D}_1(\theta(t))$ and t^{4c} must be chosen in order to have

$$t^{-4c} \mathcal{D}_1(\theta(t))$$

analytic for $|t| < 1$.

Taking the limit $c \to 0$ gives

$$\mathcal{D}_1(\theta) = \frac{\delta_1(\rho\theta)}{j\rho} + \mu \int_0^\theta du \frac{\delta_3(\rho(\theta-u))}{(j\rho)^3} \frac{1}{\sqrt{2}} \operatorname{sd}(\sqrt{2}u),$$

which agrees, for $\mu = 0$, with one of the generating functions already obtained in [24], process $P5$ in this reference.

As an interesting application of the generating function (4.7) let us work out the asymptotic analysis of the polynomials $\mathfrak{F}_n(c, \mu; x)$. We apply the operator $t\partial_t$ to the relation (4.7) to get

$$\sum_{n\geq 0} \frac{(1+c)_n}{(1/2+c)_n}(4n+4c+1)\frac{t^{4n+4c+1}}{(4c+1)!}\,\mathfrak{F}_n(c,\mu;x) = \frac{t^{-4c}}{\sqrt{1-t^4}}\frac{d\mathcal{D}_1}{d\theta}.$$

Since $t = 0$ is an apparent singularity, the nearest singularity is that of the inverse square root. From Darboux's theorem the leading behavior of the polynomials $\mathfrak{F}_n$ is given by the binomial expansion of this square root:

$$\mathfrak{F}_n(c,\mu;x) \sim \frac{(4c+1)!}{4n+4c+1}\frac{(1/2)_n(1/2+c)_n}{n!(1+c)_n}\mathfrak{H}(c,\mu;x) \tag{4.8}$$

with

$$\mathfrak{H}(c,\mu;x) = \int_0^{\theta(1)} du\,\delta_0\left(\rho(\theta(1)-u)\right)\frac{t(u)^{4c-1}}{(4c-1)!} + \mu_0\int_0^{\theta(1)} du\frac{\delta_2(\rho(\theta(1)-u))}{(j\rho)^2}\frac{t(u)^{4c+1}}{(4c+1)!},$$

where $\rho = x^{1/4}$ and the constant $\theta(1)$ is just $K_0/\sqrt{2}$. From (4.8) we deduce the continued fraction

$$\int_0^\infty \frac{d\Psi(s)}{x-s} = -\frac{\mathfrak{H}(c+1,0;x)}{\mathfrak{H}(c,\mu;x)}, \quad c \geq 0, \quad \mu + 4c^2(4c^2-1) \geq 0. \tag{4.9}$$

We are now in a position to study the first orthogonality measure corresponding to $(c = 0, \mu = 0)$. The denominator is merely

$$\mathfrak{H}(0,0;x) = \delta_0\left(\frac{x^{1/4}}{\sqrt{2}}K_0\right) = \cos\left(\frac{x^{1/4}}{2}K_0\right)\cosh\left(\frac{x^{1/4}}{2}K_0\right),$$

whose zeros give for support

$$x_n = \left((2n+1)\frac{\pi}{K_0}\right)^4, \qquad n = 0, 1, \ldots.$$

The numerator is simplified using the identity

$$2t(u)^5 = t(u) - \frac{d^2}{du^2}\left(\frac{t(u)^3}{3!}\right)$$

followed by several integrations by parts and the change of variable $K_0 - \sqrt{2}u \to u$. We are left with

$$\mathfrak{H}(1,0;x) = \frac{1}{(jx^{1/4})^2}\int_0^{K_0}\frac{du}{\sqrt{2}}\,\delta_2\left(\frac{x^{1/4}u}{\sqrt{2}}\right)\operatorname{cn} u.$$

This integral is reduced, in the appendix, to the more convenient form:

$$\mathfrak{H}(1,0;x) = \frac{-i}{4\sqrt{2x}}\left[\int_{-K_0}^{+K_0} du\,\operatorname{cd} u\cos\left(\frac{x^{1/4}u}{2}\right)\right.$$
$$\left. -2\cos\left(\frac{x^{1/4}K_0}{2}\right)\int_0^{K_0} du\,\cosh\left(\frac{x^{1/4}u}{2}\right)\left(\sqrt{2}\,\operatorname{dc} u - \operatorname{sc} u\right)\right].$$

When x is in the support of Ψ the second term vanishes while the first can be computed using the Fourier series of $\operatorname{cd} u$ [36, p. 511]. The residues give the measure masses

$$\psi_n = \frac{4\pi^2}{K_0^2} \frac{(2n+1)}{\sinh((2n+1)\pi)}, \qquad n = 0, 1, \dots$$

in agreement with Theorem 5 for $a = -1$. We have therefore Theorem 6.

THEOREM 6. *For* $\lambda_n = (4n+1)(4n+2)^2(4n+3)$ *and* $\mu_n = (4n-1)(4n)^2(4n+1)$ *asymptotic analysis gives for the* $\mathfrak{F}_n(0,0;x)$ *the orthogonality relation*

$$\sum_{l \geq 0} \psi_l \mathfrak{F}_m(0,0;x_l)\mathfrak{F}_n(0,0;x_l) = \frac{1}{4m+1} \left(\frac{(1/2)_m}{m!} \right)^2 \delta_{mn}$$

with

$$x_n = \left((2n+1)\frac{\pi}{K_0} \right)^4, \quad \psi_n = \frac{4\pi^2}{K_0^2} \frac{(2n+1)}{\sinh((2n+1)\pi)}, \quad n = 0, 1, \dots.$$

This measure is Nevanlinna extremal.

Let us observe that the derivation given in [32] and [33] does not settle whether the measures in Theorem 5 are Nevanlinna extremal or not. From Theorem 6 we know that this is the case for $a = -1$. The description of *all* the Nevanlinna extremal measures for the rates (4.1) with $c = \mu = 0$ is another work which demands the computation of the matrix A, B, C, D and will be explained elsewhere [7]. For the interested reader we mention a single result of this forthcoming analysis: among the measures of Theorem 5 only those with $a = \pm 1$ are extremal.

A second closed form orthogonality measure is obtained for $(c = 1/2, \mu = 0)$. The denominator is

$$\mathfrak{H}(1/2, 0; x) = \frac{1}{(jx^{1/4})^2} \delta_2 \left(\frac{x^{1/4}K_0}{\sqrt{2}} \right) = \frac{1}{\sqrt{x}} \sin \left(\frac{x^{1/4}K_0}{2} \right) \sinh \left(\frac{x^{1/4}K_0}{2} \right),$$

which gives for support $x_n = (2n\pi/K_0)^4$, $n = 1, 2, \dots$. To simplify the numerator we must use the relations

$$30t(u)^7 = -\frac{d^2}{du^2}\left(t(u)^5\right) + 20t(u)^3, \qquad t(u)^3 = -\frac{d^2}{du^2}t(u)$$

and repeated integrations by parts to eventually get

$$\mathfrak{H}(3/2, 0; x) = \frac{1}{2\,3!}\, \mathfrak{H}(1/2, 0; x) - \int_0^{K_0} \frac{du}{\sqrt{2}}\, \delta_0 \left(\frac{x^{1/4}K_0}{\sqrt{2}} \right) \operatorname{cn} u.$$

The right-hand side integral, using the appendix, reduces to

$$\frac{1}{4\sqrt{2}} \left[\int_{-K_0}^{+K_0} du\, \operatorname{nd} u \cos \left(\frac{x^{1/4}u}{2} \right) + 2 \sin \left(\frac{x^{1/4}K_0}{2} \right) \int_0^{K_0} du\, \sinh \left(\frac{x^{1/4}u}{2} \right) (\sqrt{2}\, \operatorname{dc} u - \operatorname{sc} u) \right].$$

Here too the second term vanishes on the support and the first one is deduced from the Fourier series for $\operatorname{nd} u$ given in [36, p. 511]. We conclude with the following theorem.

THEOREM 7. *For* $\lambda_n = (4n+3)(4n+4)^2(4n+5)$, $\mu_n = (4n+1)(4n+2)^2(4n+3)$, *asymptotic analysis gives for the* $\mathfrak{F}_n(1/2,0;x)$ *the orthogonality relation*

$$\sum_{l\geq 1} \psi_l \mathfrak{F}_m(1/2,0;x_l)\mathfrak{F}_n(1/2,0;x_l) = \frac{3}{4n+3}\left(\frac{n!}{(3/2)_n}\right)^2 \delta_{mn}$$

with

$$x_n = \left(\frac{2n\pi}{K_0}\right)^4, \quad \psi_n = \frac{4\pi^2}{K_0^2}\frac{2n\ x_n}{\sinh(2n\pi)}, \quad n = 1,2,\ldots$$

This measure is Nevanlinna extremal.

Let us notice that the process involved in Theorem 7 is dual to the process of Theorem 6. As is obvious, the general relation between their corresponding orthogonality measures (3.5) does not hold since the processes are not determined.

5. Conclusion. We think we have pushed one step further the applications of Carlitz's factorization technique; let us observe that its successes are also linked to the large body of knowledge gathered in elliptic function theory.

Further progress in the understanding of the quadratic birth and death processes will require new ideas about solutions of Heun's equation, an outstanding problem!

Despite the complexity of the spectra and the orthogonality measures for generic polynomial transition rates, the more striking fact remains the existence of very particular rates for which closed form results can be obtained. We are only at the beginning of the exploration of such an exciting field.

Appendix. Let us first recall the useful relations

$$\begin{aligned}
\delta_0(\rho\theta) &= \tfrac{1}{4}\left[\exp(\jmath\rho\theta) + \exp(-\jmath\rho\theta) + \exp(\bar{\jmath}\rho\theta) + \exp(-\bar{\jmath}\rho\theta)\right],\\
\delta_1(\rho\theta) &= \tfrac{1}{4}\left[\exp(\jmath\rho\theta) - \exp(-\jmath\rho\theta) + \imath\exp(\bar{\jmath}\rho\theta) - \imath\exp(-\bar{\jmath}\rho\theta)\right],\\
\delta_2(\rho\theta) &= \tfrac{1}{4}\left[\exp(\jmath\rho\theta) + \exp(-\jmath\rho\theta) - \exp(\bar{\jmath}\rho\theta) - \exp(-\bar{\jmath}\rho\theta)\right],\\
\delta_3(\rho\theta) &= \tfrac{1}{4}\left[\exp(\jmath\rho\theta) - \exp(-\jmath\rho\theta) - \imath\exp(\bar{\jmath}\rho\theta) + \imath\exp(-\bar{\jmath}\rho\theta)\right].
\end{aligned}$$

We are interested in

$$I_l = \jmath^{-l}\int_0^{K_0} \frac{d\theta}{\sqrt{2}}\mathrm{cn}\,\theta\ \delta_l(\rho\theta), \qquad l = 0,2, \qquad \rho = \frac{x^{1/4}}{\sqrt{2}}.$$

Using the preceding relations for the δ_l this problem reduces to the computation of

$$\mathcal{I} = \int_{-K_0}^{+K_0} d\theta \mathrm{cn}\,\theta \exp(\bar{\jmath}\rho\theta), \qquad \tilde{\mathcal{I}} = \mathcal{I}(\imath \to -\imath),$$

which gives in turn

$$I_0 = \frac{1}{4\sqrt{2}}(\mathcal{I} + \tilde{\mathcal{I}}), \qquad I_2 = \frac{\imath}{4\sqrt{2}}(\mathcal{I} - \tilde{\mathcal{I}}).$$

We first go to the complex plane by the change of variable $z = \jmath\theta$, which gives

$$\mathcal{I} = \int_{-\jmath K_0}^{\jmath K_0} dz\bar{\jmath}\mathrm{cn}\,(\bar{\jmath}z)\exp(-\imath\rho z).$$

The integration path is the first diagonal from $-\jmath K_0$ to $+\jmath K_0$.

We then use the addition theorem for lemniscate functions [36, pp. 496–497]

$$\operatorname{cn}(\bar{\jmath}z) = \frac{\operatorname{cn}^2(z/\sqrt{2}) + \imath \operatorname{sn}^2(z/\sqrt{2})\operatorname{dn}^2(z/\sqrt{2})}{\operatorname{cn}^2(z/\sqrt{2}) + \frac{1}{2}\operatorname{sn}^4(z/\sqrt{2})},$$

the duplication relations [36, p. 498]

$$\frac{1+\operatorname{cn}(\sqrt{2}z)}{2\operatorname{dn}(\sqrt{2}z)} = \frac{\operatorname{cn}^2(z/\sqrt{2})}{\operatorname{cn}^2(z/\sqrt{2}) + \frac{1}{2}\operatorname{sn}^4(z/\sqrt{2})},$$
$$\frac{1-\operatorname{cn}(\sqrt{2}z)}{2\operatorname{dn}(\sqrt{2}z)} = \frac{\operatorname{sn}^2(z/\sqrt{2})\operatorname{dn}^2(z/\sqrt{2})}{\operatorname{cn}^2(z/\sqrt{2}) + \frac{1}{2}\operatorname{sn}^4(z/\sqrt{2})},$$

and the change of variable $u = \sqrt{2}z$ to get

$$\mathcal{I} = \frac{1}{2}\int_{-(1+\imath)K_0}^{+(1+\imath)K_0} du \frac{1-\imath\operatorname{cn} u}{\operatorname{dn} u}\exp(-\imath\sigma u), \qquad \sigma = \frac{\rho}{\sqrt{2}}.$$

Let us observe that the integrand is continuous at the boundaries $\pm(1+\imath)K_0$ because

$$\operatorname{dn}(1+\imath)K_0 = 0, \qquad \operatorname{cn}(1+\imath)K_0 = -\imath,$$

but it cannot be split up into two pieces since each integral would diverge separately.

Using the Cauchy theorem we change the contour into three segments:

$$[-(1+\imath)K_0, -K_0], \quad [-K_0, +K_0], \quad [+K_0, (1+\imath)K_0].$$

Obvious transformations, including the use of Jacobi's imaginary transformation, lead eventually to the results

$$I_0 = \frac{1}{4\sqrt{2}}\left[\int_{-K_0}^{+K_0} du\ \operatorname{nd} u \cos\left(\frac{x^{1/4}u}{2}\right) + 2\sin\left(\frac{x^{1/4}K_0}{2}\right)\int_0^{K_0} du\ \sinh\left(\frac{x^{1/4}u}{2}\right)(\sqrt{2}\operatorname{dc} u - \operatorname{sc} u)\right]$$

and

$$I_2 = \frac{-i}{4\sqrt{2x}}\left[\int_{-K_0}^{+K_0} du\ \operatorname{cd} u \cos\left(\frac{x^{1/4}u}{2}\right) - 2\cos\left(\frac{x^{1/4}K_0}{2}\right)\int_0^{K_0} du\ \cosh\left(\frac{x^{1/4}u}{2}\right)(\sqrt{2}\operatorname{dc} u - \operatorname{sc} u)\right].$$

REFERENCES

[1] N. I. Akhiezer, *The classical moment problem*, Oliver and Boyd, Edinburgh, 1965.

[2] ———, *Elements of the theory of elliptic functions*, Trans. Amer. Math. Soc., Providence, RI, 1990.

[3] G. E. Andrews and R. Askey, *Classical orthogonal polynomials*, in Polynômes Orthogonaux et Applications, C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, and A. Ronveaux, eds., Lecture Notes in Mathematics, 1171, Springer-Verlag, Berlin, 1985, pp. 36–62.

[4] R. ASKEY AND M. E. ISMAIL, *Recurrence relations, continued fractions and orthogonal polynomials*, Memoirs Amer. Math. Soc., 300 (1984), pp. 1–108.
[5] R. ASKEY AND J. WIMP, *Associated Laguerre and Hermite polynomials*, Proc. Roy. Soc. Edinburgh, 96A (1984), pp. 15–37.
[6] C. BERG, *Markov's theorem revisited*, J. Approx. Theory, to appear.
[7] C. BERG AND G. VALENT, *Nevanlinna extremal measures for polynomials related to a quartic birth and death process*, Methods Appl. Anal., to appear.
[8] L. CARLITZ, *Some orthogonal polynomials related to elliptic functions*, Duke Math. J., 27 (1960), pp. 443–459.
[9] T. S. CHIHARA, *A characterization and a class of distribution functions for the Stieltjes–Wigert polynomials*, Canad. Math. Bull., 13 (1970), pp. 529–532.
[10] ———, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.
[11] ———, *On generalized Stieltjes–Wigert and related orthogonal polynomials*, J. Comput. Appl. Math., 5 (1979), pp. 291-297.
[12] T. S. CHIHARA AND M. E. ISMAIL, *Extremal measures for a system of orthogonal polynomials*, Construct. Approx., 9 (1993), pp. 111–119.
[13] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.
[14] G. GASPER AND M. RAHMAN, *Basic hypergeometric series*, Encyclopedia of Mathematics and Its Applications, Vol. 35, Cambridge University Press, Cambridge, UK, 1990.
[15] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series, and Products*, Academic Press, New York, 1980.
[16] M. E. ISMAIL, J. LETESSIER, AND G. VALENT, *Associated dual Hahn polynomials*, in Orthogonal Polynomials and Their Applications, M. Alfaro, J. S. Dehesa, F. J. Marcellan, J. L. Rubio de Francia, and J. Vinuesa, eds., Lecture Notes in Mathematics 1329, Springer-Verlag, Berlin, 1988, pp. 251–254.
[17] ———, *Quadratic birth and death processes and associated continuous dual Hahn polynomials*, SIAM J. Math. Anal., 20 (1989), pp. 727–737.
[18] M. E. ISMAIL, J. LETESSIER, D. MASSON, AND G. VALENT, *Birth and death processes and orthogonal polynomials*, in Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., NATO ASI series C, Vol. 294, Kluwer Academic Publishers, Boston, 1990, pp. 229–255.
[19] M. E. ISMAIL, J. LETESSIER, G. VALENT, AND J. WIMP, *Two families of associated Wilson polynomials*, Canad. J. Math., 42 (1990), pp. 659–695.
[20] M. E. ISMAIL, J. LETESSIER, AND G. VALENT, *Linear birth and death models and associated Laguerre polynomials*, J. Approx. Theory, 56 (1988), pp. 337–348.
[21] S. KARLIN AND J. MCGREGOR, *The classification of birth and death processes*, Trans. Amer. Math. Soc., 86 (1957), pp. 366–401.
[22] ———, *The differential equations of birth and death processes and the Stieltjes moment problem*, Trans. Amer. Math. Soc., 85 (1957), pp. 489–546.
[23] J. LETESSIER AND G. VALENT, *Exact eigenfunctions and spectrum for several cubic and quartic birth and death processes*, Phys. Lett., 108 A (1985), pp. 245–247.
[24] ———, *Some exact solutions of the Kolmogorov boundary value problem*, J. Approx. Theory Appl., 4 (1988), pp. 97–117.
[25] Y. L. LUKE, *The Special Functions and Their Approximations*, Vol. 2, Academic Press, New York, 1969.
[26] W. MAGNUS, F. OBERHETTINGER, AND R. P. SONI, *Formulas and theorems of the special functions of mathematical physics*, Springer-Verlag, New York, 1966.
[27] F. W. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[28] J. A. SHOHAT AND J. D. TAMARKIN, *The Problem of Moments*, American Mathematical Society, Providence, RI, 1950.
[29] T. STIELTJES, *Recherches sur les fractions continues*, Annales de la Faculté des Sciences de Toulouse, 8 (1894), pp. 1–122.
[30] G. SZEGÖ, *Orthogonal polynomials*, Amer. Math. Soc. Colloqium Publications, Vol. 23, Providence, RI, 1978.
[31] G. VALENT, *An integral transform involving Heun functions and a related eigenvalue problem*, SIAM J. Math. Anal., 17 (1986), pp. 688–703.
[32] ———, *Orthogonal polynomials for a quartic birth and death process*, Proceedings of the Granada Conference, 1991, to appear.
[33] ———, *Exact solutions of a quartic birth and death process and related orthogonal polynomials*, J. Comput. Appl. Math., submitted.
[34] W. VAN ASSCHE, *Orthogonal polynomials, associated polynomials and functions of the second kind*, J. Comput. Appl. Math., 37 (1991), pp. 237–249.

[35] E. VAN DOORN, *Stochastic monotonicity and queuing applications of birth-death processes*, Lecture Notes in Statistics 4, Springer-Verlag, Berlin, 1980.

[36] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, UK, 1965.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 776–811, March 1994

033

ON THE RELATIVE EXTREMA OF THE JACOBI POLYNOMIALS $P_n^{(0,-1)}(x)$*

R. WONG† AND J.-M. ZHANG‡

Abstract. Asymptotic approximations, complete with error bounds, are constructed for the Jacobi polynomials $P_{n-1}^{(0,1)}(\cos\theta)$ and $P_{n-1}^{(1,0)}(\cos\theta)$, as $n \to \infty$, which hold uniformly with respect to $\theta\epsilon[0, 0.78\pi]$. A corresponding approximation is also obtained for the zeros $\theta_{k,n}$ of $P_{n-1}^{(1,0)}(\cos\theta)$. These results are then used to prove the following conjecture of Askey: If $\nu_{k,n}$ denotes the relative extrema of the Jacobi polynomial $P_n^{(0,-1)}(x)$, ordered so that $\nu_{k+1,n}$ lies to the left of $\nu_{k,n}$, then $|\nu_{k,n-1}| < |\nu_{k,n}|$ for $k = 1, \ldots, n-1$ and $n = 1, 2, \ldots$.

Key words. Jacobi polynomials, zeros, relative extrema, uniform asymptotic approximation

AMS subject classifications. 33C45, 41A60

1. Introduction.

Let $-1 < y_{n-1,n} < \cdots < y_{1,n} < 1$ denote the critical points of the Legendre polynomial $P_n(x)$, i.e., $P_n'(y_{k,n}) = 0$, and put $y_{0,n} = 1$ and $y_{n,n} = -1$. If $\mu_{k,n} = P_n(y_{k,n})$, then it was observed by Todd [13] and proved by Szegö [12] that

$$|\mu_{k,n}| < |\mu_{k,n-1}|, \qquad k = 1, \ldots, n-1. \tag{1.1}$$

Now let $P_n^{(\alpha,\beta)}(x)$ denote the Jacobi polynomial. Szász [10] showed that these inequalities also hold for the relative extrema of $P_n^{(\alpha,\beta)}(x)/P_n^{(\alpha,\beta)}(1)$ when $\alpha = \beta > -\frac{1}{2}$. In [11, p. 190] it is stated that the same result for $\alpha > \beta > -\frac{1}{2}$ is probable, but still open, and indicated that graphical evidence suggests that the inequalities in (1.1) are reversed for the function

$$P_n^{(0,-1)}(x) = \frac{P_n(x) + P_{n-1}(x)}{2};$$

that is, if $\nu_{k,n}$ are the successive relative extrema of $P_n^{(0,-1)}(x)/P_n^{(0,-1)}(1)$ when x decreases from $+1$ to -1, then we have

$$|\nu_{k,n}| < |\nu_{k,n+1}|, \qquad k = 1, \ldots, n; \quad n = 1, 2, \ldots \tag{1.2}$$

The problem of proving this conjecture is reiterated in a more recent paper by Askey [2, p. 24, eq. (3.9)]. Establishing the inequalities in (1.2) solves a problem in Askey and Gasper [3, pp. 722–723]; see also [2, p. 24].

The purpose of this paper is to prove this conjecture. Our approach is based on asymptotic methods. In [9] and [16], we have succeeded in using these methods to prove a conjecture of Szegö concerning the monotonicity of the Lebesgue constants for

*Received by the editors December 24, 1991; accepted for publication (in revised form) March 13, 1993.

†Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2. The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A7359. Current address, Department of Mathematics, City Polytechnic of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong.

‡Department of Applied Mathematics, Tsinghua University, Beijing, China.

the Legendre series and a problem of Lorch and Szegö concerning the monotonicity of the inflection points of Bessel functions. Here, we first derive asymptotic expansions of $\nu_{k,n}$ for large values of n, and then show that (1.2) holds for $n \geq 25$. The validity of (1.2) for $1 \leq n \leq 25$ can be established by direct comparison of the numerical values of $\nu_{k,n}$.

The present paper is arranged as follows. In §2, we present some asymptotic formulas for $P_{n-1}^{(0,1)}(\cos\theta)$ and $P_{n-1}^{(1,0)}(\cos\theta)$, which are uniformly valid for $\theta\epsilon[0, \pi - \varepsilon]$, $\varepsilon > 0$. Then in §3 we derive asymptotic expansions of the zeros of $P_{n-1}^{(1,0)}(\cos\theta)$ for large values of n. Asymptotic expansions of $\nu_{k,n}$ are obtained in §4. The monotonicity of the relative extrema $\nu_{k,n}$ is proved in §§5, 6, and 7. Throughout this paper, it will be *assumed* that $n \geq 25$.

2. Uniform asymptotic approximations of Jacobi polynomials. Using the Liouville–Stekloff method, Baratella and Gatteschi [4] recently obtained an asymptotic formula for $P_n^{(\alpha,\beta)}(\cos\theta)$ when $-\frac{1}{2} \leq \alpha, \beta \leq \frac{1}{2}$, which is uniformly valid for $\theta\epsilon[0, \frac{\pi}{2}]$. Their method can be briefly described as follows. Let

$$u_n^{(\alpha,\beta)}(\theta) = \left(\sin\frac{\theta}{2}\right)^{\alpha+\frac{1}{2}}\left(\cos\frac{\theta}{2}\right)^{\beta+\frac{1}{2}} P_n^{(\alpha,\beta)}(\cos\theta). \tag{2.1}$$

Then it is well known that $u_n^{(\alpha,\beta)}(\theta)$ satisfies the differential equation

$$\frac{d^2u}{d\theta^2} + \left[N^2 + \frac{\frac{1}{4} - \alpha^2}{4\sin^2\frac{\theta}{2}} + \frac{\frac{1}{4} - \beta^2}{4\cos^2\frac{\theta}{2}}\right]u = 0, \tag{2.2}$$

where

$$N = n + \tfrac{1}{2}(\alpha + \beta + 1); \tag{2.3}$$

see [11, p. 67]. As in [4], we set

$$A = 1 - 4\alpha^2, \qquad B = 1 - 4\beta^2, \tag{2.4}$$

$$a(\theta) = \frac{2}{\theta} - \cot\frac{\theta}{2}, \qquad b(\theta) = \tan\frac{\theta}{2}, \tag{2.5}$$

$$f(\theta) = N\theta + \frac{1}{16N}[Aa(\theta) + Bb(\theta)], \tag{2.6}$$

and

$$F(\theta) = F_1(\theta) + F_2(\theta), \tag{2.7}$$

where

$$F_1(\theta) = \frac{1}{2}\frac{Aa'''(\theta) + Bb'''(\theta)}{16N^2 + Aa'(\theta) + Bb'(\theta)} - \frac{3}{4}\left[\frac{Aa''(\theta) + Bb''(\theta)}{16N^2 + Aa'(\theta) + Bb'(\theta)}\right]^2 \tag{2.8}$$

and

$$F_2(\theta) = \frac{A}{2\theta^3}\,\frac{Aa'(\theta)\theta + Bb'(\theta)\theta - Aa(\theta) - Bb(\theta)}{16N^2 + Aa(\theta)/\theta + Bb(\theta)/\theta}$$

(2.9)

$$\cdot\left[1 + \frac{1}{2}\,\frac{Aa'(\theta)\theta + Bb'(\theta)\theta - Aa(\theta) - Bb(\theta)}{16N^2\theta + Aa(\theta) + Bb(\theta)}\right] + \frac{[Aa'(\theta) + Bb'(\theta)]^2}{256N^2}.$$

Baratella and Gatteschi showed that the differential equation (2.2) can be converted into a Volterra integral equation, and that $u_n^{(\alpha,\beta)}(\theta)$ satisfies

$$\left[\frac{f(\theta)}{f'(\theta)}\right]^{-\frac{1}{2}} u_n^{(\alpha,\beta)}(\theta) = C_1 J_\alpha[f(\theta)] - \frac{\pi}{2}\int_0^\theta \left[\frac{f(t)}{f'(t)}\right]^{\frac{1}{2}} \Delta(t,\theta)F(t)u_n^{(\alpha,\beta)}(t)dt, \tag{2.10}$$

where

$$C_1 = 2^{-\frac{1}{2}}N^{-\alpha}\,\frac{\Gamma(n+\alpha+1)}{n!}\left[1 + \frac{1}{16N^2}\left(\frac{A}{6} + \frac{B}{2}\right)\right]^{-\alpha} \tag{2.11}$$

and

$$\Delta(t,\theta) = J_\alpha[f(\theta)]Y_\alpha[f(t)] - J_\alpha[f(t)]Y_\alpha[f(\theta)]. \tag{2.12}$$

Furthermore, they showed that for $-\frac{1}{2} \le \alpha, \beta \le \frac{1}{2}$, the integral

$$I = \frac{\pi}{2}\int_0^\theta \left[\frac{f(t)}{f'(t)}\right]^{\frac{1}{2}} \Delta(t,\theta)F(t)u_n^{(\alpha,\beta)}(t)dt \tag{2.13}$$

has the estimate

$$|I| \le \begin{cases} \theta^\alpha N^{-4}\binom{n+\alpha}{n}(0.00812A + 0.0828B), & 0 < \theta \le \theta^*, \\ \theta^{\frac{1}{2}}N^{-\alpha-\frac{7}{2}}\binom{n+\alpha}{n}(0.00526A + 0.535B), & \theta^* \le \theta \le \frac{\pi}{2}, \end{cases} \tag{2.14}$$

where θ^* is the root of the equation $f(\theta) = \frac{\pi}{2}$ and A, B are as given in (2.4). Regarding I as the error term, (2.10) provides a uniform asymptotic approximation for $u_n^{(\alpha,\beta)}(\theta)$, when $-\frac{1}{2} \le \alpha, \beta \le \frac{1}{2}$.

It is easily verified that (2.10) actually holds for $\alpha > -1$ and β arbitrary. Furthermore, the argument of Baratella and Gatteschi [4] can also be used to derive the following uniform asymptotic approximations for $u_{n-1}^{(0,1)}(\theta)$ and $u_{n-1}^{(1,0)}(\theta)$.

THEOREM 1. *Let $n \ge 25$ and*

$$f_1(\theta) \equiv n\theta + \frac{1}{16n}[a(\theta) - 3b(\theta)] \equiv n\theta - \frac{1}{16n}\hat{f}_1(\theta). \tag{2.15}$$

Then

$$u_{n-1}^{(0,1)}(\theta) = \left[\frac{f_1(\theta)}{f_1'(\theta)}\right]^{\frac{1}{2}}\{2^{-\frac{1}{2}}J_0[f_1(\theta)] - I\}, \tag{2.16}$$

where

$$|I| \leq \begin{cases} 0.0577n^{-\frac{7}{2}}, & 0 \leq \theta \leq 0.27\pi, \\ 0.3080n^{-\frac{7}{2}}, & 0 \leq \theta \leq \pi/2, \\ 22.739n^{-\frac{7}{2}}, & 0 \leq \theta \leq 0.78\pi. \end{cases} \tag{2.17}$$

THEOREM 2. *Let* $n \geq 25$ *and*

$$f_2(\theta) \equiv n\theta + \frac{1}{16n}[-3a(\theta) + b(\theta)] \equiv n\theta + \frac{1}{16n}\hat{f}_2(\theta). \tag{2.18}$$

Then

$$u_{n-1}^{(1,0)}(\theta) = \left[\frac{f_2(\theta)}{f_2'(\theta)}\right]^{\frac{1}{2}} \{2^{-\frac{1}{2}} J_1[f_2(\theta)] - I\}, \tag{2.19}$$

where

$$|I| \leq \begin{cases} 0.0294n^{-\frac{7}{2}}, & 0 \leq \theta \leq 0.27\pi, \\ 0.1259n^{-\frac{7}{2}}, & 0 \leq \theta \leq \pi/2, \\ 4.1103n^{-\frac{7}{2}}, & 0 \leq \theta \leq 0.78\pi. \end{cases} \tag{2.20}$$

From the Maclaurin expansions [1, p. 75]

$$a(\theta) = \frac{\theta}{6} + \frac{\theta^3}{8 \cdot 45} + \cdots + \frac{(-1)^{n-1}2^{2n}B_{2n}}{(2n)!}\left(\frac{\theta}{2}\right)^{2n-1} + \cdots, \qquad |\theta| < 2\pi, \tag{2.21}$$

$$b(\theta) = \frac{\theta}{2} + \frac{\theta^3}{24} + \cdots + \frac{(-1)^{n-1}2^{2n}(2^{2n}-1)B_{2n}}{(2n)!}\left(\frac{\theta}{2}\right)^{2n-1} + \cdots, \qquad |\theta| < \pi, \tag{2.22}$$

it is easily seen that the functions $\hat{f}_1(\theta)$ and $\hat{f}_2(\theta)$, defined in (2.15) and (2.18), respectively, and all their derivatives are positive and increasing in $0 < \theta < \pi$. This fact, and some similar ones, will be frequently used later in our argument.

3. Zeros of $P_{n-1}^{(1,0)}(\cos\theta)$. To derive asymptotic formulas for the zeros of $P_{n-1}^{(1,0)}(\cos\theta)$, we shall make use of the following result due to Gatteschi [5] as stated in Hethcote [6].

LEMMA 1. *In the interval* $[b-\rho, b+\rho]$, *suppose* $f(t) = g(t) + \varepsilon(t)$, *where* $f(t)$ *is continuous,* $g(t)$ *is differentiable,* $g(b) = 0, m = \min|g'(t)| > 0$, *and*

$$E = \max|\varepsilon(t)| < \min\{|g(b-\rho)|, |g(b+\rho)|\};$$

then there exists a zero c *of* $f(t)$ *in the interval such that* $|c-b| \leq E/m$.

Our discussion of the zeros of $P_{n-1}^{(1,0)}(\cos\theta)$ will be divided into three cases: (i) $0 < \theta < \pi/2$, (ii) $0 < \theta < 0.78\pi$, and (iii) $0.73\pi < \theta < \pi$.

First we consider the case $0 < \theta < \pi/2$. By Theorem 2, we have

$$\left(\sin\frac{\theta}{2}\right)^{\frac{3}{2}}\left(\cos\frac{\theta}{2}\right)^{\frac{1}{2}} P_{n-1}^{(1,0)}(\cos\theta) = \left[\frac{f_2(\theta)}{f_2'(\theta)}\right]^{\frac{1}{2}}\left\{\frac{1}{\sqrt{2}}J_1[f_2(\theta)] - I\right\} \tag{3.1}$$

for $0 \le \theta \le \pi/2$, where

$$|I| \le 0.1259 n^{-\frac{7}{2}}. \tag{3.2}$$

To apply Lemma 1 to (3.1), we let $t = f_2(\theta)$, and take

$$f(t) = \sqrt{2}\left[\frac{f_2(\theta)}{f_2'(\theta)}\right]^{-\frac{1}{2}} \left(\sin\frac{\theta}{2}\right)^{\frac{3}{2}} \left(\cos\frac{\theta}{2}\right)^{\frac{1}{2}} P_{n-1}^{(1,0)}(\cos\theta). \tag{3.3}$$

Equation (3.1) then becomes

$$f(t) = J_1(t) + \varepsilon(t) \tag{3.4}$$

with $\varepsilon(t) = -\sqrt{2}\, I$. From (3.2), it follows that

$$|\varepsilon(t)| \le 0.1781 n^{-\frac{7}{2}}, \qquad 0 \le t \le f_2(\pi/2). \tag{3.5}$$

Set

$$\rho_1 = 0.4006 n^{-3} \tag{3.6}$$

and let K_1 be the largest postive integer k satisfying

$$j_{1,k} + \rho_1 \le f_2(\pi/2). \tag{3.7}$$

From (2.18), it is clear that

$$\frac{j_{1,k}}{n} \le \frac{\pi}{2} + 0.0001, \qquad k = 1, \dots, K_1. \tag{3.8}$$

For each $k \le K_1$, by Taylor's theorem, there exists $\zeta \epsilon (j_{1,k} - \rho_1, j_{1,k})$ such that

$$|J_1(j_{1,k} - \rho_1)| \ge |J_1'(j_{1,k})|\rho_1 - \tfrac{1}{2}|J_1''(\zeta)|\rho_1^2. \tag{3.9}$$

Since $|J_\alpha(\zeta)| \le 1/\sqrt{2}$ if $\alpha \ge 1$, using a recurrence relation and the Bessel differential equation, we have

$$|J_1''(\zeta)| \le \tfrac{1}{4}[|J_3(\zeta)| + 3|J_1(\zeta)|] \le 2^{-\frac{1}{2}}. \tag{3.10}$$

In [9, p. 163], it has been shown that

$$(-1)^k J_0(j_{1,k}) = \sqrt{\frac{2}{\pi}}\left(j_{1,k}^{-\frac{1}{2}} - \frac{3}{16} j_{1,k}^{-\frac{5}{2}}\right) + \delta_1,$$

where

$$|\delta_1| \le \begin{cases} 0.0582 k^{-\frac{7}{2}}, & (k \ge 2), \\ 0.0360 k^{-\frac{7}{2}} & (k \ge 25). \end{cases}$$

Since $J_1'(j_{1,k}) = J_0(j_{1,k})$, it follows that

$$|J_1'(j_{1,k})| \ge 0.7883 j_{1,k}^{-\frac{1}{2}}, \qquad k \ge 20. \tag{3.11}$$

From the numerical values of $J_1'(j_{1,k})$ given in [1, p. 409], it is easily verified that (3.11) actually holds for $k \geq 1$. A combination of (3.6), (3.9)–(3.11) gives

$$|J_1(j_{1,k} - \rho_1)| \geq \frac{0.7883}{\sqrt{j_{1,k}}} \times 0.4006n^{-3} - 2^{-\frac{3}{2}} \times (0.4006n^{-3})^2 > 0.1781n^{-\frac{7}{2}} \geq E, \tag{3.12}$$

where

$$E = \max\{|\varepsilon(t)| : j_{1,k} - \rho_1 \leq t \leq j_{1,k} + \rho_1\}.$$

Using the same argument, it can be shown that

$$E < |J_1(j_{1,k} + \rho_1)|. \tag{3.13}$$

Let

$$m = \min\{|J_1'(t)| : j_{1,k} - \rho_1 < t < j_{1,k} + \rho_1\}.$$

By the mean-value theorem,

$$m \geq |J_1'(j_{1,k})| - |J_1''(\zeta)|\rho_1,$$

where $\zeta \epsilon (j_{1,k} - \rho_1, j_{1,k} + \rho_1)$. From (3.10) and (3.11) it follows that

$$m \geq |J_1'(j_{1,k})|\left\{1 - \frac{\rho_1}{\sqrt{2}|j_1'(j_{1,k})|}\right\} \geq \frac{0.7881}{\sqrt{j_{1,k}}}.$$

With $g(t) = J_1(t)$, the conditions of Lemma 1 are now all met, and hence we have the following result.

LEMMA 2. *For each $k \leq K_1$ and $n \geq 25$, the function $f(t)$ given in* (3.4) *has a zero t_k satisfying*

$$|t_k - j_{1,k}| \leq \frac{E}{m} \leq \frac{0.2260\sqrt{j_{1,k}}}{n^{\frac{7}{2}}}. \tag{3.14}$$

THEOREM 3. *Let $\theta_{k,n}$ denote the root of the equation $f_2(\theta) = t_k$. For each $k \leq K_1$ and $n \geq 25$, $\theta_{k,n}$ is the kth zero of $P_{n-1}^{(1,0)}(\cos\theta)$ and satisfies $\theta_{k,n} < \pi/2$. Furthermore,*

$$\theta_{k,n} = \frac{j_{1,k}}{n} - \frac{1}{16n^2}\left[-3a\left(\frac{j_{1,k}}{n}\right) + b\left(\frac{j_{1,k}}{n}\right)\right] + \varepsilon_1, \tag{3.15}$$

where

$$|\varepsilon_1| \leq 0.2263\sqrt{j_{1,k}}n^{-\frac{9}{2}}. \tag{3.16}$$

Proof. For each $k \leq K_1$, we have, by Lemma 2 and (3.7),

$$t_k < j_{1,k} + \rho_1 \leq f_2\left(\frac{\pi}{2}\right).$$

Since $f_2(\theta_{k,n}) = t_k$ and $f_2(\theta)$ is increasing, it follows that $\theta_{k,n} < \frac{\pi}{2}$.

We now rewrite Lemma 2 in the form

$$t_k = j_{1,k} + \varepsilon_2 \tag{3.17}$$

with

$$|\varepsilon_2| \le \frac{0.2260\sqrt{j_{1,k}}}{n^{\frac{7}{2}}}, \tag{3.18}$$

and let

$$\varepsilon_3 = \theta_{k,n} - \frac{j_{1,k}}{n}. \tag{3.19}$$

By the mean-value theorem,

$$f_2(\theta_{k,n}) = f_2\left(\frac{j_{1,k}}{n}\right) + f_2'(\eta)\varepsilon_3,$$

where η is between $\theta_{k,n}$ and $j_{1,k}/n$. In view of (2.18) and (3.17), $f_2(\theta_{k,n}) = t_k$ can be written as

$$\frac{1}{16n}\left[-3a\left(\frac{j_{1,k}}{n}\right) + b\left(\frac{j_{1,k}}{n}\right)\right] + \left[n + \frac{1}{16n}\hat{f}_2'(\eta)\right]\varepsilon_3 = \varepsilon_2.$$

As a consequence, we have

$$\varepsilon_3 = \left\{-\frac{1}{16n^2}\left[-3a\left(\frac{j_{1,k}}{n}\right) + b\left(\frac{j_{1,k}}{n}\right)\right] + \frac{\varepsilon_2}{n}\right\} \Big/ \left[1 + \frac{1}{16n^2}\hat{f}_2'(\eta)\right]. \tag{3.20}$$

In view of the formula

$$\frac{1}{1+x} = 1 - \frac{x}{1+x},$$

(3.20) gives

$$\varepsilon_3 = -\frac{1}{16n^2}\left[-3a\left(\frac{j_{1,k}}{n}\right) + b\left(\frac{j_{1,k}}{n}\right)\right] + \varepsilon_1, \tag{3.21}$$

where

$$\varepsilon_1 = \frac{\varepsilon_2}{n} + \left[\frac{\varepsilon_2}{n} - \frac{1}{16n^2}\hat{f}_2\left(\frac{j_{1,k}}{n}\right)\right] \cdot \frac{1}{16n^2}\hat{f}_2'(\eta) \Big/ \left[1 + \frac{1}{16n^2}\hat{f}_2'(\eta)\right].$$

Since $\eta < (\pi/2) + 0.0001$, $\hat{f}_2'(\eta) < 0.4318$. From (2.21) and (2.22), it is also easily shown that $\hat{f}_2(\theta) < \frac{1}{30}\theta^3$. Therefore,

$$|\varepsilon_1| \le \frac{|\varepsilon_2|}{n} + \frac{0.0270}{n^2}\left[\frac{|\varepsilon_2|}{n} + \frac{1}{480n^2}\left(\frac{j_{1,k}}{n}\right)^3\right].$$

The estimate (3.16) now follows from (3.8) and (3.18), and the asymptotic formula (3.15) is obtained by inserting (3.21) in (3.19).

From (3.3) and (3.4), it is evident that for each k, the root of $f_2(\theta) = t_k$ is a zero of $P_{n-1}^{(1,0)}(\cos\theta)$. Since the kth zero $\theta_{n-1,k}^{(1,0)}$ of $P_{n-1}^{(1,0)}(\cos\theta)$ satisfies

$$\lim_{n\to\infty} n\theta_{n-1,k}^{(1,0)} = j_{1,k}$$

(see [1, p. 787]), by comparing this with (3.15) we conclude that $\theta_{n-1,k}^{(1,0)} = \theta_{k,n}$, thus proving the theorem. □

To extend the result of Theorem 3 to zeros of $P_{n-1}^{(1,0)}(\cos\theta)$ in a larger interval, we set

$$\rho_2 = 13.2n^{-3} \tag{3.22}$$

and let K_2 be the largest positive integer k satisfying

$$j_{1,k} + \rho_2 \le f_2(0.78\pi). \tag{3.23}$$

Since the proof of our next result is exactly the same as that of Theorem 3, it will be omitted.

THEOREM 4. *For each $k \le K_2$ and $n \ge 25$, the root $\theta_{k,n}$ of $f_2(\theta) = t_k$ is the kth zero of $P_{n-1}^{(1,0)}(\cos\theta)$ and satisfies $\theta_{k,n} < 0.78\pi$. Furthermore,*

$$\theta_{k,n} = \frac{j_{1,k}}{n} - \frac{1}{16n^2}\left[-3a\left(\frac{j_{1,k}}{n}\right) + b\left(\frac{j_{1,k}}{n}\right)\right] + \varepsilon_4, \tag{3.24}$$

where

$$|\varepsilon_4| \le 7.4349\sqrt{j_{1,k}}\,n^{-\frac{9}{2}}. \tag{3.25}$$

To deal with the zeros of $P_{n-1}^{(1,0)}(\cos\theta)$ in the interval $(0.73\pi, \pi)$, we make use of the reflection formula [11, p. 59]

$$P_{n-1}^{(1,0)}(x) = (-1)^{n-1}P_{n-1}^{(0,1)}(-x). \tag{3.26}$$

Let $\tilde{\theta} = \pi - \theta$. Then from (2.1) it is easily seen that

$$(-1)^{n-1}u_{n-1}^{(1,0)}(\theta) = u_{n-1}^{(0,1)}(\tilde{\theta}). \tag{3.27}$$

By Theorem 1, we have

$$u_{n-1}^{(0,1)}(\tilde{\theta}) = \left[\frac{f_1(\tilde{\theta})}{f_1'(\tilde{\theta})}\right]^{\frac{1}{2}}\left\{2^{-\frac{1}{2}}J_0[f_1(\tilde{\theta})] - I\right\}, \tag{3.28}$$

where

$$|I| \le 0.0577n^{-\frac{7}{2}}, \qquad 0.73\pi \le \theta < \pi. \tag{3.29}$$

Make the change of variable $t = f_1(\tilde{\theta})$, and let $g(t) = J_0(t)$ and

$$f(t) = J_0(t) + \varepsilon(t) \tag{3.30}$$

with

$$|\varepsilon(t)| = \sqrt{2}\, I \le \sqrt{2} \times 0.0577 n^{-\frac{7}{2}}. \tag{3.31}$$

Furthermore, let

$$\rho_3 = 0.1814 n^{-3} \tag{3.32}$$

and K_3 be the largest positive integer satisfying

$$j_{0,k} + \rho_3 \le f_1(0.27\pi). \tag{3.33}$$

Clearly,

$$\frac{j_{0,k}}{n} < 0.27\pi, \qquad k = 1, \dots, K_3. \tag{3.34}$$

Using arguments similar to that given for Lemma 2, it can be verified that for every positive integer $k \le K_3$,

$$\begin{aligned} E &= \max\{|\varepsilon(t)| : j_{0,k} - \rho_3 \le t \le j_{0,k} + \rho_3\} < 0.0817 n^{-\frac{7}{2}} \\ &< \min\{|J_0(j_{0,k} - \rho_3)|, |J_0(j_{0,k} + \rho_3)|\} \end{aligned}$$

and

$$\begin{aligned} m &= \min\{|J_0'(t)| : j_{0,k} - \rho_3 \le t \le j_{0,k} + \rho_3\} \\ &\ge |J_0'(j_{0,k})| - |J_0''(\zeta)|\rho_3 \ge 0.7976 j_{0,k}^{-\frac{1}{2}}, \end{aligned}$$

where $\zeta \epsilon (j_{0,k} - \rho_3, j_{0,k} + \rho_3)$. Hence by Lemma 1 and (3.34), we have the following.

LEMMA 3. *For each $k \le K_3$ and $n \ge 25$, the function $f(t)$ given in* (3.30) *has a zero t_k satisfying*

$$|t_k - j_{0,k}| \le \frac{E}{m} \le 0.0943 n^{-3}. \tag{3.35}$$

The proof of the following theorem is similar to that of Theorem 3.

THEOREM 5. *Let $\tilde{\theta}_{k,n}$ denote the root of $f_2(\tilde{\theta}) = t_k$, and let $\theta_{n-k,n} = \pi - \tilde{\theta}_{k,n}$. For each $k \le K_3$ and $n \ge 25$, $\theta_{n-k,n}$ is the $(n-k)$th zero of $P_{n-1}^{(1,0)}(\cos\theta)$ and satisfies $\theta_{n-k,n} \ge 0.73\pi$. Furthermore,*

$$\theta_{n-k,n} = \pi - \frac{j_{0,k}}{n} + \frac{1}{16n^2}\left[a\left(\frac{j_{0,k}}{n}\right) - 3b\left(\frac{j_{0,k}}{n}\right)\right] + \varepsilon_5, \tag{3.36}$$

where

$$|\varepsilon_5| \le 0.1025 n^{-4}. \tag{3.37}$$

4. Asympototic approximations of $\nu_{k,n}$. In view of the identity [11, §4.21.7]

$$\frac{d}{dx} P_n^{(0,-1)}(x) = \frac{n}{2} P_{n-1}^{(1,0)}(x), \tag{4.1}$$

we know that the critical points of $P_n^{(0,-1)}(x)$ are exactly the zeros of $P_{n-1}^{(1,0)}(x)$. Thus the relative extrema $\nu_{k,n}$ of $P_n^{(0,-1)}(x)$ are given by

$$\nu_{k,n} = P_n^{(0,-1)}(\cos\theta_{k,n}), \qquad k = 1, \dots, n-1, \tag{4.2}$$

where $\theta_{k,n}$, as we have shown in §3, is the kth zero of $P_{n-1}^{(1,0)}(\cos\theta)$. To obtain the asymptotic behaviour of $\nu_{k,n}$, we first note that

$$u_n^{(0,-1)}(\theta) = u_{n-1}^{(0,1)}(\theta), \tag{4.3}$$

which can be obtained from the reflection formula (3.26) and the identity [11, p. 64]

$$P_n^{(-1,0)}(x) = \frac{x-1}{2} P_{n-1}^{(1,0)}(x). \tag{4.4}$$

By Theorem 1,

$$u_n^{(0,-1)}(\theta) = \left[\frac{f_1(\theta)}{f_1'(\theta)}\right]^{\frac{1}{2}} \{2^{-\frac{1}{2}} J_0[f_1(\theta)] - I\}, \tag{4.5}$$

where

$$|I| \le 0.3080 n^{-\frac{7}{2}}, \qquad 0 \le \theta \le \frac{\pi}{2}. \tag{4.6}$$

Thus, in view of (2.1), we have for $k \le K_1$,

$$\nu_{k,n} = \left[\frac{f_1(\theta_{k,n})}{f_1'(\theta_{k,n})}\right]^{\frac{1}{2}} \left(\cot\frac{\theta_{k,n}}{2}\right)^{\frac{1}{2}} \{2^{-\frac{1}{2}} J_0[f_1(\theta_{k,n})] - I\}. \tag{4.7}$$

In the following, we shall derive the asymptotic expansion of each of the quantities $f_1(\theta_{k,n})$, $[f_1(\theta_{k,n})/f_1'(\theta_{k,n})]^{\frac{1}{2}}$, and $J_0[f_1(\theta_k, n)]$.

By the mean value theorem,

$$\hat{f}_1(\theta_{k,n}) = \hat{f}_1\left(\frac{j_{1,k}}{n}\right) + \hat{f}_1'(\zeta)\left(\theta_{k,n} - \frac{j_{1,k}}{n}\right),$$

where ζ lies between $\theta_{k,n}$ and $j_{1,k}/n$. From (2.15) and Theorem 3, it follows that

$$f_1(\theta_{k,n}) = j_{1,k} + \frac{1}{4n}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right] + \varepsilon_6, \tag{4.8}$$

where, in view of (2.18),

$$\varepsilon_6 = n\varepsilon_1 - \frac{1}{16n}\hat{f}_1'(\zeta)\left[\varepsilon_1 - \frac{1}{16n^2}\hat{f}_2\left(\frac{j_{1,k}}{n}\right)\right].$$

Since ζ is between $\theta_{k,n}$ and $j_{1,k}/n$, by Theorem 3 and the inequality in (3.8), $0 < \zeta < (\pi/2) + 0.0001$ and $|\hat{f}_1'(\zeta)| \le 2.8109$. Since $\hat{f}_2(\theta) < \frac{1}{30}\theta^3$ as remarked in the proof of Theorem 3, we conclude from (3.16) that

$$|\varepsilon_6| \le 0.2275\sqrt{j_{1,k}}\, n^{-\frac{7}{2}}. \tag{4.9}$$

In a similar manner, we have from (2.15),

$$\frac{f_1(\theta_{k,n})}{f_1'(\theta_{k,n})} = \frac{1 - \frac{1}{16n^2}[\hat{f}_1(\theta_{k,n})/\theta_{k,n}]}{1 - \frac{1}{16n^2}\hat{f}_1'(\theta_{k,n})}\,\theta_{k,n}.$$

The identity

$$\frac{1}{1-x} = 1 + x + \frac{x^2}{1-x}$$

then gives

$$\frac{f_1(\theta_{k,n})}{f_1'(\theta_{k,n})} = \theta_{k,n}\left[1 + \frac{g_2(\theta_{k,n})}{16n^2} + \varepsilon_7'\right], \tag{4.10}$$

where

$$g_2(\theta) = -\frac{\hat{f}_1(\theta)}{\theta} + \hat{f}_1'(\theta) = \left[\frac{a(\theta)}{\theta} - 3\frac{b(\theta)}{\theta}\right] - [a'(\theta) - 3b'(\theta)] \tag{4.11}$$

and

$$\begin{aligned}\varepsilon_7' = &-\frac{1}{256n^4}\frac{\hat{f}_1(\theta_{k,n})}{\theta_{k,n}}\hat{f}_1'(\theta_{k,n}) \\ &+ \frac{1}{256n^4}\left[\hat{f}_1'(\theta_{k,n})\right]^2\left[1 - \frac{1}{16n^2}\frac{\hat{f}_1(\theta_{k,n})}{\theta_{k,n}}\right] \Big/ \left[1 - \frac{1}{16n^2}\hat{f}_1'(\theta_{k,n})\right].\end{aligned}$$

From the power series representations (2.21) and (2.22), it is easily shown that $\hat{f}_1'(\theta) > \hat{f}_1(\theta)/\theta$ and hence $\varepsilon_7' > 0$. Since $\theta_{k,n} < \frac{\pi}{2}$ and $\frac{2}{\pi}\hat{f}_1\left(\frac{\pi}{2}\right) \le 1.7360$ and $\hat{f}_1'\left(\frac{\pi}{2}\right) \le 2.8106$, it follows that

$$0 < \varepsilon_7' < 0.05n^{-4}. \tag{4.12}$$

Applying Taylor's theorem to (4.10), we obtain

$$\left[\frac{f_1(\theta_{k,n})}{f_1'(\theta_{k,n})}\right]^{\frac{1}{2}} = \sqrt{\theta_{k,n}}\left[1 + \frac{g_2(\theta_{k,n})}{32n^2} + \varepsilon_7\right], \tag{4.13}$$

where

$$\varepsilon_7 = \frac{1}{2}\varepsilon_7' - \frac{1}{8(1+\zeta)^{\frac{3}{2}}}\left\{\frac{g_2(\theta_{k,n})}{16n^2} + \varepsilon_7'\right\}^2 \tag{4.14}$$

and

$$0 < \zeta < \frac{g_2(\theta_{k,n})}{16n^2} + \varepsilon_7' < 0.0002.$$

The last estimate follows from the fact that $0 < g_2(\theta) < g_2(\frac{\pi}{2}) \le 1.0747$ for $0 < \theta < \frac{\pi}{2}$. Since the second term on the right-hand side of (4.14) is positive and smaller than the first term, we have from (4.12)

$$0 < \varepsilon_7 < 0.025n^{-4}. \tag{4.15}$$

Inserting (4.13) in (4.7) yields the following result.

LEMMA 4. *For $k \leq K_1$ and $n \geq 25$, $\nu_{k,n}$ has the asymptotic representation*

$$\nu_{k,n} = \frac{g_1(\theta_{k,n})}{\sqrt{2}} \left[1 + \frac{g_2(\theta_{k,n})}{32n^2} + \varepsilon_7\right] \{J_0[f_1(\theta_{k,n})] + \varepsilon_8\}, \tag{4.16}$$

where

$$g_1(\theta) = \left(\theta \cot \frac{\theta}{2}\right)^{\frac{1}{2}}, \tag{4.17}$$

$g_2(\theta)$ is given by (4.11), *ε_7 satisfies* (4.15), *and*

$$|\varepsilon_8| = |-\sqrt{2}I| \leq 0.4356n^{-\frac{7}{2}}. \tag{4.18}$$

Our next result gives the asymptotic expansion of $J_0[f_1(\theta_{k,n})]$.

LEMMA 5. *For $k \leq K_1$ and $n \geq 25$,*

$$\begin{aligned} J_0[f_1(\theta_{k,n})] = J_0(j_{1,k}) &+ \frac{J_0''(j_{1,k})}{32n^2}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^2 \\ &+ \frac{J_0'''(j_{1,k})}{384n^3}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^3 + \varepsilon_9, \end{aligned} \tag{4.19}$$

where

$$|\varepsilon_9| \leq 0.0415\sqrt{j_{1,k}}|J_0(j_{1,k})|n^{-\frac{9}{2}}. \tag{4.20}$$

Proof. Put

$$h = \frac{1}{4n}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right] + \varepsilon_6$$

so that (4.8) becomes $f_1(\theta_{k,n}) = j_{1,k} + h$. By Taylor's theorem

$$J_0[f_1(\theta_{k,n})] = J_0(j_{1,k}) + \frac{h^2}{2}J_0''(j_{1,k}) + \cdots + \frac{h^5}{120}J_0^{(5)}(\zeta),$$

where ζ lies between $f_1(\theta_{k,n})$ and $j_{1,k}$. The expansion (4.19) now follows by letting

$$\begin{aligned} \varepsilon_9 = &\frac{1}{4n}J_0''(j_{1,k})\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]\varepsilon_6 + \frac{1}{2}J_0''(j_{1,k})\varepsilon_6^2 \\ &+ \frac{1}{32n^2}J_0'''(j_{1,k})\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^2\varepsilon_6 \\ &+ \frac{1}{8n}J_0'''(j_{1,k})\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]\varepsilon_6^2 + \frac{1}{6}J_0'''(j_{1,k})\varepsilon_6^3 \\ &+ \frac{h^4}{24}J_0^{(4)}(j_{1,k}) + \frac{h^5}{120}J_0^{(5)}(\zeta). \end{aligned}$$

Using the differential equation and the recurrence relations for Bessel functions, it is easily verified that

$$\begin{aligned} |J_0''(j_{1,k})| &= |-J_0(j_{1,k})| = |J_0(j_{1,k})|, \\ |J_0'''(j_{1,k})| &= \frac{1}{j_{1,k}}|J_0''(j_{1,k})| = \frac{1}{j_{1,k}}|J_0(j_{1,k})|, \\ |J_0^{(4)}(j_{1,k})| &= |(3j_{1,k}^{-2} - 1)J_0''(j_{1,k})| \leq |J_0(j_{1,k})|, \end{aligned}$$

and

$$|J_0^{(5)}(\zeta)| \le \frac{5}{8}|J_1(\zeta)| + \frac{5}{16}|J_3(\zeta)| + \frac{1}{16}|J_5(\zeta)| \le \frac{1}{\sqrt{2}}.$$

Since $J_0(j_{1,k}) = J_1'(j_{1,k})$, by (3.11) we also have

$$|J_0^{(5)}(\zeta)| \le \frac{1}{\sqrt{2}|J_0(j_{1,k})|}|J_0(j_{1,k})| \le 0.8971\sqrt{j_{1,k}}|J_0(j_{1,k})|.$$

Note that both $b(\theta) - a(\theta)$ and $[b(\theta) - a(\theta)]/\theta^{1/8}$ are increasing in $0 < \theta < \pi$. Hence, on account of (3.8), we have $|a(\theta) - b(\theta)| \le 0.7269$ and $|a(\theta) - b(\theta)|/\theta^{1/8} \le 0.6870$ for $\theta \le \frac{\pi}{2} + 0.0001$. The desired estimate (4.20) is now obtained by combining the above results together. □

In exactly the same manner, one can derive the corresponding asymptotic expansion for $K_1 < k < K_2$. Thus we have the following.

LEMMA 6. *For $K_1 < k < K_2$ and $n \ge 25$,*

$$\nu_{k,n} = \frac{g_1(\theta_{k,n})}{\sqrt{2}}\left[1 + \frac{g_2(\theta_{k,n})}{32n^2} + \varepsilon_{10}\right]\{J_0[f_1(\theta_{k,n})] + \varepsilon_{11}\}, \tag{4.21}$$

where $g_1(\theta)$ and $g_2(\theta)$ are as defined in (4.17) *and* (4.11), *respectively,*

$$|\varepsilon_{10}| \le 1.2978n^{-4} \tag{4.22}$$

and

$$|\varepsilon_{11}| \le 32.1585n^{-\frac{7}{4}}. \tag{4.23}$$

Furthermore,

$$\begin{aligned} J_0[f_1(\theta_{k,n})] = J_0(j_{1,k}) &+ \frac{J_0''(j_{1,k})}{32n^2}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^2 \\ &+ \frac{J_0'''(j_{1,k})}{384n^3}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^3 + \varepsilon_{12}, \end{aligned} \tag{4.24}$$

where

$$|\varepsilon_{12}| \le 12.4871\sqrt{j_{1,k}}|J_0(j_{1,k})|n^{-\frac{9}{2}}. \tag{4.25}$$

To deal with the case $n - K_3 \le k \le n - 1$, we let, as in Theorem 5,

$$\tilde{\theta}_{n-k,n} = \pi - \theta_{k,n}.$$

By (3.35),

$$\tilde{\theta}_{n-k,n} = \frac{j_{0,n-k}}{n} - \frac{1}{16n^2}\left[a\left(\frac{j_{0,n-k}}{n}\right) - 3b\left(\frac{j_{0,n-k}}{n}\right)\right] + \varepsilon_5, \tag{4.26}$$

where

$$|\varepsilon_5| \le 0.1025n^{-4}. \tag{4.27}$$

Applying the reflection formula [11, p. 59] and identity (4.4) to (4.2), it is easily verified that

$$\nu_{k,n} = (-1)^{n+1}\left(\tan\frac{\tilde{\theta}_{n-k,n}}{2}\right)^{\frac{1}{2}} u_{n-1}^{(1,0)}(\tilde{\theta}_{n-k,n}).$$

Since $\tilde{\theta}_{n-k,n} < 0.27\pi$ by Theorem 5, it follows from (2.19) that

$$\nu_{k,n} = (-1)^{n+1}\left(\tan\frac{\tilde{\theta}_{n-k,n}}{2}\right)^{\frac{1}{2}}\left[\frac{f_2(\tilde{\theta}_{n-k,n})}{f_2'(\tilde{\theta}_{n-k,n})}\right]^{\frac{1}{2}}\left\{\frac{1}{\sqrt{2}}J_1\left[f_2(\tilde{\theta}_{n-k,n})\right] - I\right\}, \tag{4.28}$$

where

$$|I| \le 0.0294n^{-\frac{7}{2}}. \tag{4.29}$$

Using the same argument as that for (4.8), it can be shown that

$$f_2(\tilde{\theta}_{n-k,n}) = j_{0,n-k} - \frac{1}{4n}\left[a\left(\frac{j_{0,n-k}}{n}\right) - b\left(\frac{j_{0,n-k}}{n}\right)\right] + \varepsilon_{13}, \tag{4.30}$$

where

$$|\varepsilon_{13}| \le 0.1029n^{-3}. \tag{4.31}$$

Moreover, in a manner similar to that given in Lemma 4, one can obtain the following result.

LEMMA 7. *For $k = n - K_3, \ldots, n-1$ and $n \ge 25$,*

$$\begin{aligned}\nu_{k,n} = {} & (-1)^{n+1}\frac{g_3(\tilde{\theta}_{n-k,n})}{\sqrt{2}}\left\{1 + \frac{1}{32n^2}g_4(\tilde{\theta}_{n-k,n}) + \varepsilon_{14}\right\} \\ & \cdot\left\{J_1\left[f_2(\tilde{\theta}_{n-k,n})\right] - \sqrt{2}I\right\},\end{aligned} \tag{4.32}$$

where

$$g_3(\theta) = \left(\theta\tan\frac{\theta}{2}\right)^{\frac{1}{2}}, \tag{4.33}$$

$$g_4(\theta) = -3\frac{a(\theta)}{\theta} + \frac{b(\theta)}{\theta} + 3a'(\theta) - b'(\theta), \tag{4.34}$$

and

$$|\varepsilon_{14}| \le 0.0001n^{-4}. \tag{4.35}$$

5. Monotonicity of $\nu_{k,n}$: $1 \le k \le K_1$. Lemma 4 gives

$$\begin{aligned}\sqrt{2}(\nu_{k,n} - \nu_{k,n+1}) = {} & \left[1 + \frac{g_2(\theta_{k,n})}{32n^2} + \varepsilon_7\right]\{J_0[f_1(\theta_{k,n})] + \varepsilon_8\}D_1 \\ & + g_1(\theta_{k,n+1})\{J_0[f_1(\theta_{k,n})] + \varepsilon_8\}D_2 \\ & + g_1(\theta_{k,n+1})\left\{1 + \frac{g_2(\theta_{k,n+1})}{32(n+1)^2} + \varepsilon_7(n+1)\right\}D_3,\end{aligned} \tag{5.1}$$

where

$$D_1 = g_1(\theta_{k,n}) - g_1(\theta_{k,n+1}), \tag{5.2}$$

$$D_2 = \frac{g_2(\theta_{k,n})}{32n^2} + \varepsilon_7(n) - \frac{g_2(\theta_{k,n+1})}{32(n+1)^2} - \varepsilon_7(n+1), \tag{5.3}$$

and

$$D_3 = J_0\left[f_1(\theta_{k,n})\right] + \varepsilon_8(n) - J_0\left[f_1(\theta_{k,n+1})\right] - \varepsilon_8(n+1). \tag{5.4}$$

In (5.3) and (5.4), we have indicated the dependence of ε_7 and ε_8 on n. In what follows, we shall estimate each of the quantities D_1, D_2, and D_3.

By the mean value theorem,

$$D_1 = g_1'(\xi_1)(\theta_{k,n} - \theta_{k,n+1}), \tag{5.5}$$

where ξ_1 lies between $\theta_{k,n}$ and $\theta_{k,n+1}$. Thus, to estimate D_1, it suffices to consider $g_1'(\xi_1)$ and $\theta_{k,n} - \theta_{k,n+1}$ separately.

LEMMA 8. *For $n \geq 25$, we have*

$$\theta_{k,n} - \theta_{k,n+1} = \frac{j_{1,k}}{n(n+1)} + \varepsilon_{15}, \tag{5.6}$$

where

$$|\varepsilon_{15}| \leq 0.0414\frac{j_{1,k}}{n^3(n+1)} + 0.4526\sqrt{j_{1,k}}n^{-\frac{9}{2}}. \tag{5.7}$$

Proof. Let

$$g_5(\theta) = \theta^2\left[-3a(\theta) + b(\theta)\right] = \theta^2\hat{f}_2(\theta); \tag{5.8}$$

cf. (2.18). Straightforward computation gives $\frac{2}{\pi}\hat{f}_2\left(\frac{\pi}{2}\right) = 0.1148$ and $\hat{f}_2'\left(\frac{\pi}{2}\right) = 0.4318$. Therefore

$$0 < g_5'(\theta) \leq 0.6614\theta^2, \qquad 0 < \theta \leq \frac{\pi}{2}. \tag{5.9}$$

By Theorem 3,

$$\theta_{k,n} - \theta_{k,n+1} = \frac{j_{1,k}}{n(n+1)} + \varepsilon_{15},$$

where

$$\begin{aligned}\varepsilon_{15} = &-\frac{1}{16n^2}\left[-3a\left(\frac{j_{1,k}}{n}\right) + b\left(\frac{j_{1,k}}{n}\right)\right] + \varepsilon_1(n)\\ &+\frac{1}{16(n+1)^2}\left[-3a\left(\frac{j_{1,k}}{n+1}\right) + b\left(\frac{j_{1,k}}{n+1}\right)\right] - \varepsilon_1(n+1).\end{aligned}$$

In terms of $g_5(\theta)$,

$$\varepsilon_{15} = -\frac{1}{16j_{1,k}^2}\left[g_5\left(\frac{j_{1,k}}{n}\right) - g_5\left(\frac{j_{1,k}}{n+1}\right)\right] + \varepsilon_1(n) - \varepsilon_1(n+1).$$

By the mean value theorem again,

$$|\varepsilon_{15}| \le \frac{1}{16 j_{1,k}^2} g_5'(\xi) \frac{j_{1,k}}{n(n+1)} + 2|\varepsilon_1(n)|,$$

where $\xi \in (j_{1,k}/(n+1), j_{1,k}/n)$. The final result (5.7) now follows from (5.9) and (3.16). □

LEMMA 9. *For* $0 < \theta \le \frac{\pi}{2}$,

$$g_1'(\theta) \le -\frac{1}{12}\theta g_1(\theta) \le 0; \tag{5.10}$$

and for $k \le K_1$ *and* $n \ge 5$,

$$g_1(\theta_{k,n+1}) \ge g_1(\theta_{k,n}) \ge 0.9890 g_1(\theta_{k,n+1}). \tag{5.11}$$

Proof. From (4.17), we have

$$g_1'(\theta) = \frac{1}{2}\left(\frac{1}{\theta}\cot\frac{\theta}{2}\right)^{\frac{1}{2}}\left(1 - \frac{\theta}{\sin\theta}\right).$$

The inequalities in (5.10) follow from the expansion [1, p. 75]

$$\frac{\theta}{\sin\theta} = 1 + \frac{1}{6}\theta^2 + \cdots + (-1)^{n-1}\frac{2(2^{2n}-1)B_{2n}}{(2n)!}\theta^{2n} + \cdots, \qquad 0 < \theta < \pi. \tag{5.12}$$

From (5.12), we also have

$$\begin{aligned} |g_1'(\theta)| &\le \frac{1}{2\theta}\left(\frac{\theta}{\sin\theta} - 1\right) g_1(\theta) \\ &\le \frac{1}{\pi}\left(\frac{\pi}{2} - 1\right) g_1(\theta) \le 0.1817 g_1(\theta). \end{aligned} \tag{5.13}$$

By the mean value theorem, there exists ζ between $\theta_{k,n}$ and $\theta_{k,n+1}$ such that

$$g_1(\theta_{k,n}) = g_1(\theta_{k,n+1}) + g_1'(\zeta)(\theta_{k,n} - \theta_{k,n+1}). \tag{5.14}$$

Since the zeros of $P_n^{(1,0)}(x)$ interlace between the zeros of $P_{n-1}^{(1,0)}(x)$, we have $\theta_{k,n+1} < \zeta < \theta_{k,n}$. Coupling (5.13) and (5.14) gives

$$g_1(\theta_{k,n}) \ge g_1(\theta_{k,n+1}) - 0.1817 g_1(\zeta)(\theta_{k,n} - \theta_{k,n+1}).$$

Since $g_1(\theta)$ is decreasing and $\theta_{k,n} - \theta_{k,n+1} < 0.0605$ by Lemma 8 and (3.8), (5.11) follows. □

A combination of (5.5), (5.10), (3.15), and (5.6) yields

$$\begin{aligned} D_1 &\le -\frac{1}{12}\xi_1 g_1(\xi_1)(\theta_{k,n} - \theta_{k,n+1}) \\ &\le -\frac{1}{12}\theta_{k,n+1} g_1(\theta_{k,n})(\theta_{k,n} - \theta_{k,n+1}) \\ &\le -\frac{0.9890}{12}\left\{\frac{j_{1,k}}{n+1} - \frac{1}{16(n+1)^2}\left[-3a\left(\frac{j_{1,k}}{n+1}\right) + b\left(\frac{j_{1,k}}{n+1}\right)\right] + \varepsilon_1(n+1)\right\} \\ &\quad \times \left\{\frac{j_{1,k}}{n(n+1)} + \varepsilon_{15}\right\} g_1(\theta_{k,n+1}) \\ &\le -0.0824(1 + \varepsilon_{15}')\frac{j_{1,k}^2}{n(n+1)^2} g_1(\theta_{k,n+1}) < 0, \end{aligned} \tag{5.15}$$

where

$$|\varepsilon_{15}'| = \left|\varepsilon_{15}\frac{n(n+1)}{j_{1,k}}\right| \le 0.0895n^{-2}.$$

The second to the last inequality in (5.15) is obtained by using (3.16) and the facts that $j_{1,k} \ge j_{1,1} \ge 3.8317$ and $(1/\theta)\hat{f}_2(\theta) < 0.1148$ for $0 < \theta < (\pi/2) + 0.0001$.

We now proceed with the estimation of D_2 in (5.3). By (3.15) and the mean value theorem, we have

$$g_2(\theta_{k,n}) = g_2\left(\frac{j_{1,k}}{n}\right) + \varepsilon_{16}, \tag{5.16}$$

where

$$\varepsilon_{16} = g_2'(\eta)\left\{-\frac{1}{16n^2}\left[-3a\left(\frac{j_{1,k}}{n}\right) + b\left(\frac{j_{1,k}}{n}\right)\right] + \varepsilon_1\right\}$$

and η lies in between $j_{1,k}/n$ and $\theta_{k,n}$. Using the series representations (2.21) and (2.22), it can be readily shown that $g_2'(\theta)$ is increasing in $0 < \theta < \pi$. Therefore, $0 < g_2'(\eta) < 2.2843$ for $0 < \eta < \frac{\pi}{2} + 0.0001$, which together with the fact that $\frac{1}{\theta}\hat{f}_2(\theta) < 0.1148$ for $\theta < \frac{\pi}{2} + 0.0001$, yields

$$|\varepsilon_{16}| \le 0.0186\frac{j_{1,k}}{n^3}. \tag{5.17}$$

Put $\tilde{g}_2(\theta) = \theta^2 g_2(\theta)$. By the same argument, it can be proved that $\frac{1}{\theta^3}\tilde{g}_2'(\theta)$ is increasing in $0 < \theta < \pi$. Since $g_2(\theta_0) = 1.0749$ and $g_2'(\theta_0) = 2.2843$, where $\theta_0 = \frac{\pi}{2} + 0.0001$, it follows that $\tilde{g}_2'(\theta_0) = 9.0142$ and

$$\tilde{g}_2'(\theta) = \theta^3\frac{\tilde{g}_2'(\theta)}{\theta^3} \le 2.3254\theta^3, \qquad 0 < \theta < \theta_0.$$

Inserting (5.16) in (5.3) gives

$$\begin{aligned} D_2 &= \frac{1}{32j_{1,k}^2}\left[\tilde{g}_2\left(\frac{j_{1,k}}{n}\right) - \tilde{g}_2\left(\frac{j_{1,k}}{n+1}\right)\right] + \varepsilon_7(n) \\ &\quad + \frac{1}{32n^2}\varepsilon_{16}(n) - \varepsilon_7(n+1) - \frac{1}{32(n+1)^2}\varepsilon_{16}(n+1) \\ &= \frac{1}{32j_{1,k}^2}\tilde{g}_2'(\zeta)\frac{j_{1,k}}{n(n+1)} + \varepsilon_7(n) + \frac{1}{32n^2}\varepsilon_{16}(n) \\ &\quad - \varepsilon_7(n+1) - \frac{1}{32(n+1)^2}\varepsilon_{16}(n+1), \end{aligned}$$

where $j_{1,k}/n + 1 < \zeta < j_{1,k}/n$. Consequently,

$$\begin{aligned} |D_2| &\le \frac{2.3254}{32}\frac{j_{1,k}^2}{n^4(n+1)} + 2|\varepsilon_7| + \frac{|\varepsilon_{16}|}{16n^2} \\ &\le 0.0727\frac{j_{1,k}^2}{n^4(n+1)} + 0.05n^{-4} + 0.0012\frac{j_{1,k}}{n^5}. \end{aligned} \tag{5.18}$$

Finally, we come to the estimation of D_3. In view of (5.4) and (4.19), we can write

$$D_3 = \frac{D_{31}}{32}J_0''(j_{1,k}) + \frac{D_{32}}{384}J_0'''(j_{1,k}) + D_{33} + D_{34}, \tag{5.19}$$

where

$$D_{31} = n^{-2}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^2 - (n+1)^{-2}\left[a\left(\frac{j_{1,k}}{n+1}\right) - b\left(\frac{j_{1,k}}{n+1}\right)\right]^2, \tag{5.20}$$

$$D_{32} = n^{-3}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^3 - (n+1)^{-3}\left[a\left(\frac{j_{1,k}}{n+1}\right) - b\left(\frac{j_{1,k}}{n+1}\right)\right]^3, \tag{5.21}$$

and

$$D_{33} = \varepsilon_8(n) - \varepsilon_8(n+1), \qquad D_{34} = \varepsilon_9(n) - \varepsilon_9(n+1). \tag{5.22}$$

Put

$$g_6(\theta) = \theta^2\left[a(\theta) - b(\theta)\right]^2 \tag{5.23}$$

and

$$g_7(\theta) = \theta^3\left[a(\theta) - b(\theta)\right]^3. \tag{5.24}$$

Direct computation from (2.5) gives $a(\theta_0) - b(\theta_0) = -0.7268\ldots$ and $a'(\theta_0) - b'(\theta_0) = -0.8106\ldots$, where $\theta_0 = \frac{\pi}{2} + 0.0001$. From the series representations (2.21) and (2.22), it is also easily shown that both $\frac{1}{\theta}\left[a(\theta) - b(\theta)\right]$ and $a'(\theta) - b'(\theta)$ are negative and decreasing. As a result, we have

$$0 \le g_6'(\zeta) \le 1.17835\zeta^3, \qquad 0 < \zeta < \theta_0. \tag{5.25}$$

By a similar argument, we get

$$0 \ge g_7'(\eta) \ge -1.28471\eta^4, \qquad 0 < \eta < \theta_0. \tag{5.26}$$

From (5.20), it follows that there exists $\zeta \in (j_{1,k}/n+1, j_{1,k}/n)$ such that

$$\begin{aligned} |D_{31}| &= \frac{1}{j_{1,k}^2}\left|g_6\left(\frac{j_{1,k}}{n}\right) - g_6\left(\frac{j_{1,k}}{n+1}\right)\right| \\ &= \frac{1}{j_{1,k}^2}\left|g_6'(\zeta)\right|\frac{j_{1,k}}{n(n+1)} \le 1.17835\frac{j_{1,k}^2}{n^4(n+1)}. \end{aligned}$$

The last inequality is obtained by using (5.25). In the same manner, we conclude that

$$\begin{aligned} |D_{32}| &= \frac{1}{j_{1,k}^3}\left|g_7\left(\frac{j_{1,k}}{n}\right) - g_7\left(\frac{j_{1,k}}{n+1}\right)\right| \\ &= \frac{1}{j_{1,k}^3}\left|g_7'(\zeta)\right|\frac{j_{1,k}}{n(n+1)} \le 1.28471\frac{j_{1,k}^2}{n^5(n+1)}. \end{aligned}$$

From (4.18) and (4.20), it is also evident that

$$|D_{33}| \le 2|\varepsilon_8(n)| \le 0.8712n^{-\frac{7}{2}}$$

and

$$|D_{34}| \le 2|\varepsilon_9(n)| \le 0.083\sqrt{j_{1,k}}\,|J_0(j_{1,k})|\,n^{-\frac{9}{2}}.$$

A combination of these estimates yields

$$|D_3| \le \Big\{\frac{1.17835}{32\sqrt{n}(n+1)} + \frac{1.28471}{384n^{\frac{3}{2}}(n+1)j_{1,k}} + \frac{0.8712}{j_{1,k}^2\,|J_0(j_{1,k})|} + \frac{0.083}{j_{1,k}^{\frac{3}{2}}n}\Big\}\frac{j_{1,k}^2}{n^{\frac{7}{2}}}\,|J_0(j_{1,k})|\,;$$

cf. the estimates of $J_0''(j_{1,k})$ and $J_0'''(j_{1,k})$ in §4. Since $|J_0''(j_{1,k})| \ge 0.7883 j_{1,k}^{-1/2}$ by (3.11) and $j_{1,1} \ge 3.831706$, the last estimate reduces to

$$|D_3| \le 0.1481\frac{j_{1,k}^2}{n^{\frac{7}{2}}}\,|J_0''(j_{1,k})|\,. \tag{5.27}$$

LEMMA 10. *For $k \le K_1$ and $n \ge 25$, we have*

$$\frac{\sqrt{2}}{J_0(j_{1,k})}(\nu_{k,n} - \nu_{k,n+1}) \le -0.0525\frac{j_{1,k}^2}{n^3}g_1(\theta_{k,n+1}) < 0. \tag{5.28}$$

Proof. By (4.19), equation (5.1) gives

$$\begin{aligned}\sqrt{2}(\nu_{k,n} - \nu_{k,n+1}) = {}& [1+\varepsilon_{17}(n)](1+\varepsilon_{18})J_0(j_{1,k})D_1\\ &+ g_1(\theta_{k,n+1})(1+\varepsilon_{18})J_0(j_{1,k})D_2\\ &+ g_1(\theta_{k,n+1})[1+\varepsilon_{17}(n+)]D_3,\end{aligned}$$

where

$$\varepsilon_{17} = \frac{g_2(\theta_{k,n})}{32n^2} + \varepsilon_7$$

and

$$\begin{aligned}\varepsilon_{18} = {}& \frac{1}{32n^2}\frac{J_0''(j_{1,k})}{J_0(j_{1,k})}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^2\\ &+ \frac{1}{384n^3}\frac{J_0'''(j_{1,k})}{J_0(j_{1,k})}\left[a\left(\frac{j_{1,k}}{n}\right) - b\left(\frac{j_{1,k}}{n}\right)\right]^3 + \frac{\varepsilon_8+\varepsilon_9}{J_0(j_{1,k})}.\end{aligned}$$

Since $g_2\left(\frac{\pi}{2}\right) < 1.0747$ (see the argument following (4.14)), it is clear from (4.15) that

$$|\varepsilon_{17}| \le 0.0337n^{-2}.$$

Using the fact that $-0.7269 \le a(\theta_0) - b(\theta_0) < 0$, where $\theta_0 = \frac{\pi}{2} + 0.0001$, it can be easily shown that

$$|\varepsilon_{18}| \le 0.0444n^{-2};$$

see the arguments for (5.27). Therefore, from (5.15), (5.18), and (5.27), it follows that

$$\begin{aligned}\frac{\sqrt{2}}{J_0(j_{1,k})}(\nu_{k,n} - \nu_{k,n+1}) \le {}& -0.0824(1-|\varepsilon_{15}'|)(1-|\varepsilon_{17}|)(1-|\varepsilon_{18}|)\\ &\times \frac{j_{1,k}^2}{n(n+1)^2}g_1(\theta_{k,n+1})\\ &+ g_1(\theta_{k,n+1})|D_2|(1+|\varepsilon_{18}|)\\ &+ g_1(\theta_{k,n+1})[1+|\varepsilon_{17}(n+1)|]\frac{|D_3|}{|J_0(j_{1,k})|}\\ \le {}& -0.0525\frac{j_{1,k}^2}{n^3}g_1(\theta_{k,n+1}) < 0,\end{aligned}$$

thus proving the lemma. □

THEOREM 6a. *For $k \le K_1$ and $n \ge 25$, we have*

$$|\nu_{k,n}| < |\nu_{k,n+1}|.$$

Proof. Let $j'_{0,k}$ denote the kth positive critical point of $J_0(x)$. Since $J_0'(x) = -J_1(x)$, we conclude that $j'_{0,k} = j_{1,k}$. Furthermore, since $J_0(j'_{0,1}) < 0$, we also have

$$\operatorname{sgn}\{J_0(j_{1,k})\} = (-1)^k.$$

Next we prove that

$$\operatorname{sgn}(\nu_{k,n}) = (-1)^k.$$

Since the function of $y = P_n^{(0,-1)}(x)$ satisfies the differential equation [12, Thm. 4.2.1]

$$(1-x^2)y'' - (1+x)y' + n^2 y = 0, \qquad -1 < x < 1,$$

and $y(1) = 1$, at an extremum $\xi = \cos\theta_{k,n}$ we have

$$y''(\xi)y(\xi) = y''(\xi)\nu_{k,n} < 0.$$

Thus, each extremum yielding a positive value is a maximum, and each extremum yielding a negative value is a minimum. Furthermore, putting $x = 1$ in the above differential equation shows that $y'(1) > 0$. If there were an extremum after the largest zero, there would have to be a positive minimum so that the graph could rise to $y(1) = 1$. But there cannot be a positive minimum. Therefore, the first extremum (from the right) must occur between the two largest zeros. Since there can be only one extremum between each pair of zeros, successive extrema (in decreasing order) must alternate in sign. Since the first extremum is negative, this proves that $\operatorname{sgn}(\nu_{k,n}) = (-1)^k$, as desired.

Coupling these results together with Lemma 10, we obtain

$$\begin{aligned}(-1)^{k+1} &= \operatorname{sgn}\{\nu_{k,n} - \nu_{k,n+1}\} = \operatorname{sgn}\{(\operatorname{sgn}\nu_{k,n})|\nu_{k,n}| - (\operatorname{sgn}\nu_{k,n+1})|\nu_{k,n+1}|\} \\ &= (-1)^k \operatorname{sgn}\{|\nu_{k,n}| - |\nu_{k,n+1}|\},\end{aligned}$$

for $n \ge 25$. The desired inequality now follows. □

6. Monotonicity of $\nu_{k,n}$: $K_1 < k \le K_2$. In this case, we proceed exactly in the same manner as in §5, except that we will make use of Lemma 6, instead of Lemmas 4 and 5, and Theorem 4, instead of Theorem 3. This will result in replacing ε_1 by ε_4 in the second equation following (5.9) used to define ε_{15}. The notation D_i, D_{ij}, and ε'_{15} in this section will have the same meaning as those given in the previous one, except that their estimates are now considerably bigger.

Using an argument similar to that given in Lemma 9, it can be shown that

$$g_1(\theta_{k,n}) \ge 0.9257 g_1(\theta_{k,n+1}), \qquad K_1 < k \le K_2.$$

From (3.8), it also follows that

$$j_{1,k} \ge \frac{\pi}{2} n \ge 39.2699, \qquad K_1 < k.$$

Consequently, we have

$$D_1 \le -0.0771(1+\varepsilon'_{15})\frac{j_{1,k}^2}{n(n+1)^2}g_1(\theta_{k,n+1}) < 0,$$

$$|\varepsilon'_{15}| \le 4.1369n^{-2},$$

$$|D_2| \le 0.8754\frac{j_{1,k}^2}{n^4(n+1)} + 0.4596\frac{j_{1,k}}{n^5} + 2.5956n^{-4},$$

$$|D_{31}| \le 16.8243\frac{j_{1,k}^2}{n^4(n+1)},$$

$$|D_{32}| \le 74.6039\frac{j_{1,k}^2}{n^5(n+1)},$$

$$|D_{33}| \le 64.3107n^{-\frac{7}{2}},$$

$$|D_{34}| \le 24.9742\sqrt{j_{1,k}}|J_0(j_{1,k})|n^{-\frac{9}{2}},$$

$$|D_3| \le 0.3398\frac{j_{1,k}^2}{n^{\frac{7}{2}}}|J_0(j_{1,k})|.$$

The proofs of the following two results are exactly the same as those of Lemma 10 and Theorem 6a, and hence will be omitted.

LEMMA 11. *For $K_1 < k \le K_2$ and $n \ge 25$, we have*

$$\frac{\sqrt{2}}{J_0(j_{1,k})}(\nu_{k,n} - \nu_{k,n+1}) \le -0.0013\frac{j_{1,k}^2}{n^3}g_1(\theta_{k,n+1}) < 0.$$

THEOREM 6b. *For $K_1 < k \le K_2$ and $n \ge 25$, we have*

$$|\nu_{k,n}| < |\nu_{k,n+1}|.$$

7. Monotonicity of $\nu_{k,n}$: $n - K'_3 \le k \le n-1$. Similar to (3.32) and (3.33), we let

$$\rho'_3 = 0.1814(n+1)^{-3}$$

and K'_3 be the largest k satisfying

$$j_{0,k+1} + \rho'_3 \le (0.27\pi)(n+1) - \frac{1}{16(n+1)}\hat{f}_1(0.27\pi). \tag{7.1}$$

Since the difference $\triangle j_{0,k} \equiv j_{0,k+1} - j_{0,k}$ increases with k, $j_{0,k+1} - j_{0,k} \ge j_{0,2} - j_{0,1} \ge 3.115252$. By comparing (3.33) with (7.1), it can then be shown that $K'_3 \le K_3$. Since $\hat{f}_1(0.27\pi) > 0$, we have from (3.34) and (7.1)

$$\frac{j_{0,n-k}}{n} < 0.27\pi \quad \text{and} \quad \frac{j_{0,n-k+1}}{n+1} < 0.27\pi. \tag{7.2}$$

Using the inequalitites [7]

$$\left(m-\frac{1}{4}\right)\pi \le j_{0,m} \le \left(m-\frac{1}{4}\right)\pi + \frac{1}{8\pi\left(m-\frac{1}{4}\right)}, \qquad m=1,2,\ldots, \tag{7.3}$$

it can be proved that $K_3' \le 0.29n$ and consequently

$$\frac{j_{0,n-k}}{n} < \frac{j_{0,n+1-k}}{n+1} \quad \text{for } n-K_3' \le k \le n-1. \tag{7.4}$$

LEMMA 12. *For $n-K_3' \le k \le n-1$ and $n \ge 25$, we have*

$$\tilde{\theta}_{n-k,n} - \tilde{\theta}_{n+1-k,n+1} = \frac{-k\pi - \frac{1}{4}\pi + (n+1)\delta_{n-k} - n\delta_{n+1-k}}{n(n+1)} + \varepsilon_{19}, \tag{7.5}$$

where

$$|\varepsilon_{19}| \le 0.4803n^{-3} \tag{7.6}$$

and

$$0 \le \delta_m \le \frac{1}{8\left(m-\frac{1}{4}\right)\pi}, \qquad m=1,2,\ldots. \tag{7.7}$$

Proof. By Theorem 5,

$$\tilde{\theta}_{n-k,n} - \tilde{\theta}_{n+1-k,n+1} = \frac{j_{0,n-k}}{n} - \frac{j_{0,n+1-k}}{n+1} + \varepsilon_{19}, \tag{7.8}$$

where

$$\varepsilon_{19} = -\frac{1}{16n^2}\left[a\left(\frac{j_{0,n-k}}{n}\right) - 3b\left(\frac{j_{0,n-k}}{n}\right)\right] - \varepsilon_5(n)$$

$$+ \frac{1}{16(n+1)^2}\left[a\left(\frac{j_{0,n+1-k}}{n+1}\right) - 3b\left(\frac{j_{0,n+1-k}}{n+1}\right)\right] + \varepsilon_5(n+1).$$

In view of (2.15), we can write

$$\begin{aligned} |\varepsilon_{19}| &\le 2|\varepsilon_5| + \frac{1}{16}\left[\frac{1}{n^2} - \frac{1}{(n+1)^2}\right]\left|\hat{f}_1\left(\frac{j_{0,n-k}}{2}\right)\right| \\ &\quad + \frac{1}{16(n+1)^2}\left|\hat{f}_1\left(\frac{j_{0,n-k}}{n}\right) - \hat{f}_1\left(\frac{j_{0,n-k+1}}{n+1}\right)\right| \\ &\le 2|\varepsilon_5| + \frac{1}{16}\,\frac{2n+1}{n^2(n+1)^2}\left|\hat{f}_1\left(\frac{j_{0,n-k}}{n}\right)\right| \\ &\quad + \frac{1}{16(n+1)^2}|\hat{f}_1'(\zeta_1)| \cdot \left|\frac{j_{0,n-k}}{n} - \frac{j_{0,n+1-k}}{n+1}\right|, \end{aligned} \tag{7.9}$$

where $\zeta_1 \epsilon (j_{0,n-k}/n, j_{0,n+1-k}/n+1)$. The result in (7.3) can be written as

$$j_{0,m} = \left(m - \frac{1}{4}\right)\pi + \delta_m,$$

where δ_m satisfies (7.7). This gives

$$\frac{j_{0,n-k}}{n} - \frac{j_{0,n+1-k}}{n+1} = \frac{-k\pi - \frac{1}{4}\pi + (n+1)\delta_{n-k} - n\delta_{n+1-k}}{n(n+1)}. \tag{7.10}$$

Using (7.4) and the fact that the difference $\Delta j_{0,m} \equiv j_{0,m+1} - j_{0,m}$ forms an increasing sequence (see [11, p. 20, Thm. 1.82.2]), it can be shown that

$$\left|\frac{j_{0,n-k}}{n} - \frac{j_{0,n+1-k}}{n+1}\right| \le \frac{3.1416}{n}. \tag{7.11}$$

Inserting (7.10) in (7.8) yields the desired approximation (7.5). Since $|\hat{f}_1(0.27\pi)| < 1.2115$ and $|\hat{f}_1'(0.27\pi)| < 1.6330$. Substituting these estimates and (7.11) in (7.9), we obtain the required inequality (7.6). □

Now put

$$D = (-1)^{n+1}\sqrt{2}(\nu_{k,n} - \nu_{k,n+1}). \tag{7.12}$$

By Lemma 7,

(7.13)

$$\begin{aligned} D = D_1 \cdot \left\{1 + \frac{1}{32n^2} g_4(\tilde{\theta}_{n-k,n}) + \varepsilon_{14}(n)\right\} \cdot \left\{J_1[f_2(\tilde{\theta}_{n-k,n})] - \sqrt{2}\, I(n)\right\} \\ + g_3(\tilde{\theta}_{n+1-k,n+1}) \cdot D_2 \cdot \left\{J_1[f_2(\tilde{\theta}_{n-k,n})] - \sqrt{2}\, I(n)\right\} \\ + g_3(\tilde{\theta}_{n+1-k,n+1}) \cdot \left\{1 + \frac{1}{32(n+1)^2} g_4(\tilde{\theta}_{n+1-k,n+1}) + \varepsilon_{14}(n+1)\right\} \cdot D_3, \end{aligned}$$

where $\tilde{\theta}_{n-k,n} = \pi - \theta_{k,n}$ (see Theorem 5),

$$D_1 = g_3(\tilde{\theta}_{n-k,n}) - g_3(\tilde{\theta}_{n+1-k,n+1}), \tag{7.14}$$

$$D_2 = \frac{1}{32n^2} g_4(\tilde{\theta}_{n-k,n}) + \varepsilon_{14}(n) - \frac{1}{32(n+1)^2} g_4(\tilde{\theta}_{n+1-k,n+1}) - \varepsilon_{14}(n+1), \tag{7.15}$$

and

$$D_3 = J_1[f_2(\tilde{\theta}_{n+1-k,n+1})] - \sqrt{2}\, I(n+1) + J_1[f_2(\tilde{\theta}_{n-k,n})] - \sqrt{2}\, I(n). \tag{7.16}$$

As in §5, we shall first estimate the quantities D_1, D_2, and D_3.

From (7.14), we have

$$D_1 = g_3'(\tilde{\theta}_{n+1-k,n+1})(\tilde{\theta}_{n-k,n} - \tilde{\theta}_{n+1-k,n+1}) + \tfrac{1}{2}g_3''(\zeta_2)(\tilde{\theta}_{n-k,n} - \tilde{\theta}_{n+1-k,n+1})^2, \tag{7.17}$$

where it can be shown by using (7.8) and (7.3) that

$$\tilde{\theta}_{n-k,n} < \zeta_2 < \tilde{\theta}_{n+1-k,n+1} \quad \text{for } n - K_3' \le k \le n-1. \tag{7.18}$$

LEMMA 13. *For $0 < \theta \le 0.27\pi$, the function $g_3(\theta)$ in (4.33) is increasing and satisfies*

$$0 < g_3''(\theta) \le 0.2951 g_3(\theta). \tag{7.19}$$

Proof. Since $g_3(\theta)$ is positive and

$$g_3'(\theta) = \frac{1}{2}g_3(\theta)\left(\frac{1}{\theta} + \frac{1}{\sin\theta}\right), \tag{7.20}$$

it is clear that $g_3'(\theta)$ is positive and hence $g_3(\theta)$ is increasing. From (7.20), we also have

$$g_3''(\theta) = \frac{1}{4}g_3(\theta)\left[-\frac{1}{\theta^2} + \frac{1}{\sin^2\theta} - \frac{2\cos\theta}{\sin^2\theta} + \frac{2}{\theta\sin\theta}\right].$$

Using the series representations of (2.21), (2.22), and (5.12), it can be shown that

$$\left(\frac{1}{\sin^2\theta} - \frac{1}{\theta^2}\right) + \frac{2}{\sin\theta}\left(\frac{1}{\theta} - \cot\theta\right) = 1 + \sum_{s=1}^{\infty} c_s\theta^s,$$

where $c_s \ge 0$ for all $s \ge 1$. Hence for $0 < \theta \le 0.27\pi$,

$$0 < -\frac{1}{\theta^2} + \frac{1}{\sin^2\theta} - \frac{2\cos\theta}{\sin^2\theta} + \frac{2}{\theta\sin\theta} \le 1.1801.$$

The result (7.19) now follows. □

By Lemma 13

$$0 \le g_3''(\zeta_2) < 0.2951 g_3(\tilde{\theta}_{n+1-k,n+1})$$

for $\tilde{\theta}_{n-k,n} < \zeta_2 < \tilde{\theta}_{n+1-k,n+1}$. Inserting (7.20) and (7.5) in (7.17) gives

$$\begin{aligned} D_1 \le{}& \frac{g_3(\tilde{\theta}_{n+1-k,n+1})}{\tilde{\theta}_{n+1-k,n+1}}\left\{\frac{-k\pi - \frac{1}{4}\pi + (n+1)\delta_{n-k} - n\delta_{n+1-k}}{n(n+1)} + \varepsilon_{19}\right\} \\ &+ 0.1476 g_3(\tilde{\theta}_{n+1-k,n+1})\left\{\frac{-k\pi - \frac{1}{4}\pi + (n+1)\delta_{n-k} - n\delta_{n+1-k}}{n(n+1)} + \varepsilon_{19}\right\}^2. \end{aligned} \tag{7.21}$$

In view of (2.15), (4.26) can be written as

$$\tilde{\theta}_{n+1-k,n+1} = \frac{j_{0,n+1-k}}{n+1}\left\{1 + \frac{1}{16(n+1)^2}\hat{f}_1\left(\frac{j_{0,n+1-k}}{n+1}\right)\frac{n+1}{j_{0,n+1-k}} + \varepsilon_5 \cdot \frac{n+1}{j_{0,n+1-k}}\right\}.$$

Since $\hat{f}_1(\theta_0)/\theta_0 \le 1.4283$ where $\theta_0 = 0.27\pi$, this yields

$$\tilde{\theta}_{n+1-k,n+1} \le 1.0002\frac{j_{0,n+1-k}}{n+1}, \qquad n - K_3' \le k \le n-1. \tag{7.22}$$

Coupling (7.21) and (7.22), we obtain

$$D_1 \le g_3(\tilde{\theta}_{n+1-k,n+1})\frac{n+1}{1.0002 j_{0,n+1-k}}\left\{-\frac{(k+\frac{1}{4})\pi-(n+1)\delta_{n-k}+n\delta_{n+1-k}}{n(n+1)}+\varepsilon_{19}\right\}$$

$$+0.1476 g_3(\tilde{\theta}_{n+1-k,n+1})\left\{-\frac{(k+\frac{1}{4})\pi-(n+1)\delta_{n-k}+n\delta_{n+1-k}}{n(n+1)}+\varepsilon_{19}\right\}^2 \tag{7.23}$$

$$= g_3(\tilde{\theta}_{n+1-k,n+1})\left\{-\frac{(k+\frac{1}{4})\pi-(n+1)\delta_{n-k}+n\delta_{n+1-k}}{n j_{0,n+1-k}}\right\}\left\{\frac{1}{1.0002}+\varepsilon_{20}\right\} < 0,$$

where

$$\varepsilon_{20} = -\frac{n(n+1)\varepsilon_{19}}{(1.0002)[(k+\frac{1}{4})\pi-(n+1)\delta_{n-k}+n\delta_{n+1-k}]}$$

$$+0.1476\cdot\left\{-\frac{j_{0,n+1-k}\cdot n}{(k+\frac{1}{4})\pi-(n+1)\delta_{n-k}+n\delta_{n+1-k}}\right\}\cdot$$

$$\cdot\left\{-\frac{(k+\frac{1}{4})\pi-(n+1)\delta_{n-k}+n\delta_{n+1-k}}{n(n+1)}+\varepsilon_{19}\right\}^2.$$

From (7.2) and (7.3), we have $k+\frac{1}{4} \ge 0.73(n+1)$. Using this fact and (7.2) and (7.7), it can be shown that

$$|\varepsilon_{20}| \le \frac{0.4053}{n}. \tag{7.24}$$

To estimate D_2 in (7.15), we write

$$D_2 = D_{21} + D_{22} + D_{23},$$

where

$$D_{21} = \frac{1}{32}\left[\frac{1}{n^2} - \frac{1}{(n+1)^2}\right] g_4(\tilde{\theta}_{n-k,n}),$$

$$D_{22} = \frac{1}{32(n+1)^2}[g_4(\tilde{\theta}_{n-k,n}) - g_4(\tilde{\theta}_{n+1-k,n+1})],$$

and

$$D_{23} = \varepsilon_{14}(n) - \varepsilon_{14}(n+1).$$

Using the series representations (2.21) and (2.22), it can be shown that both $|g_4(\theta)|$ and $|g_4'(\theta)|$ are increasing in $0 < \theta < \pi$. Hence for $0 < \theta \le 0.27\pi$,

$$|g_4(\theta)| \le |g_4(0.27\pi)| \le 0.0573$$

and

$$|g_4'(\theta)| \le |g_4'(0.27\pi)| \le 0.1595.$$

Consequently,

$$|D_{21}| \le \frac{|g_4(\tilde{\theta}_{n-k,n})|}{16n^2(n+1)} \le \frac{0.0036}{n^3}$$

and, in view of (7.8) and (7.11),

$$|D_{22}| \le \frac{|g_4'(\xi)|}{32(n+1)^2}|\tilde{\theta}_{n-k,n} - \tilde{\theta}_{n+1-k,n+1}| \le \frac{0.0157}{n^3},$$

where $\xi\epsilon(\tilde{\theta}_{n-k,n}, \tilde{\theta}_{n+1-k,n+1})$. From (4.35), we also have

$$|D_{23}| \le 2|\varepsilon_{14}| \le 0.0002n^{-4}.$$

Thus for $n - K_3' \le k \le n-1$,

$$|D_2| \le \frac{0.0195}{n^3}. \tag{7.25}$$

Finally, we come to the estimation of D_3 given in (7.16), which we shall write as

$$D_3 = D_{31} + D_{32} \tag{7.26}$$

with

$$D_{31} = J_1[f_2(\tilde{\theta}_{n-k,n})] + J_1[f_2(\tilde{\theta}_{n+1-k,n+1})] \tag{7.27}$$

and

$$D_{32} = -\sqrt{2}\, I(n) - \sqrt{2}\, I(n+1), \tag{7.28}$$

where $I = I(n)$ satisfies (4.29). To estimate D_{31}, we first recall the well-known asymptotic approximation [1, p. 364]

$$J_1(x) = \sqrt{\frac{2}{\pi x}}\left[\cos\left(x - \frac{3}{4}\pi\right) P(x) - \sin\left(x - \frac{3}{4}\pi\right) Q(x)\right], \tag{7.29}$$

where

$$P(x) = 1 + \eta_1(x), \qquad 0 < \eta_1(x) < \frac{15}{128}x^{-2}, \tag{7.30}$$

and

$$Q(x) = \frac{3}{8x} + \eta_2(x), \qquad -\frac{105}{1024}x^{-3} < \eta_2(x) < 0. \tag{7.31}$$

From (4.30) and the equation preceding (7.10), we obtain

$$\begin{aligned} J_1[f_2(\tilde{\theta}_{n-k,n})] = (-1)^{n-k+1}&\left[\frac{2}{\pi f_2(\tilde{\theta}_{n-k,n})}\right]^{\frac{1}{2}} \\ &\cdot \{(\cos\varphi_n)\cdot P[f_2(\tilde{\theta}_{n-k,n})] + (\sin\varphi_n)\cdot Q[f_2(\tilde{\theta}_{n-k,n})]\}, \end{aligned} \tag{7.32}$$

where

$$\varphi_n = \frac{1}{4n}\left[a\left(\frac{j_{0,n-k}}{n}\right) - b\left(\frac{j_{0,n-k}}{n}\right)\right] - \delta_{n-k} - \varepsilon_{13}, \tag{7.33}$$

δ satisfies (7.7), and ε satisfies (4.31). We next present some preliminary results concerning $f_2(\tilde{\theta}_{n-k,n})$ and φ_n.

LEMMA 14. *For $n - K_3' \le k \le n-1$ and $n \ge 25$, we have*

$$0 < f_2(\tilde{\theta}_{n+1-k,n+1}) - f_2(\tilde{\theta}_{n-k,n}) \le 3.1447. \tag{7.34}$$

Proof. The first inequality follows from (7.18) and the fact that $f_2(\theta)$ is increasing in $0 < \theta < \pi$. The asymptotic approximation (4.30) gives

$$\begin{aligned} &f_2(\tilde{\theta}_{n+1-k,n+1}) - f_2(\tilde{\theta}_{n-k,n}) \\ &\quad = j_{0,n+1-k} - j_{0,n-k} \\ &\qquad - \frac{1}{4(n+1)}\left[a\left(\frac{j_{0,n+1-k}}{n+1}\right) - b\left(\frac{j_{0,n+1-k}}{n+1}\right)\right] \\ &\qquad + \frac{1}{4n}\left[a\left(\frac{j_{0,n-k}}{n}\right) - b\left(\frac{j_{0,n-k}}{n}\right)\right] + \varepsilon_{13}(n+1) - \varepsilon_{13}(n). \end{aligned}$$

Since $j_{0,m+1} - j_{0,m} < \pi$ for all $m \ge 1$ (see the argument for (7.11)) and $0 < b(\theta) - a(\theta) < b(0.27\pi) - a(0.27\pi) < 0.3085$, we obtain

$$f_2(\tilde{\theta}_{n+1-k,n+1}) - f_2(\tilde{\theta}_{n-k,n}) < \pi + 0.0030 + 2|\varepsilon_{13}|.$$

In view of (4.31), this proves the second inequality in (7.34). □

LEMMA 15. *For $n - K_3' \le k \le n-1$ and $n \ge 25$,*

$$f_2(\tilde{\theta}_{n-k,n}) > j_{0,n-k}.$$

Proof. From (2.15) and (2.16), we have

$$\frac{1}{\theta}[b(\theta) - a(\theta)] > \frac{1}{3}, \qquad 0 < \theta < \pi.$$

Therefore, (4.30) gives

$$f_1(\tilde{\theta}_{n-k,n}) > j_{0,n-k} + \frac{j_{0,n-k}}{12n^2} + \varepsilon_{13} > j_{0,n-k}. \qquad \square$$

LEMMA 16. *For $n - K_3' \le k \le n-4$ and $n \ge 25$, we have*

$$\text{(i)} \quad |\cos\varphi_n - \cos\varphi_{n+1}| \le \frac{0.0002}{n+1} + \frac{0.0006}{n-k-\frac{1}{4}},$$

$$\text{(ii)} \quad |\sin\varphi_n - \sin\varphi_{n+1}| \le \frac{0.0139}{n+1} + \frac{0.0398}{n-k-\frac{1}{4}},$$

$$\text{(iii)} \quad |P[f_2(\tilde{\theta}_{n-k,n})] - P[f_2(\tilde{\theta}_{n+1-k,n+1})]| \le \frac{0.0191}{(n-k+\frac{3}{4})^2},$$

$$\text{(iv)} \quad |Q[f_2(\tilde{\theta}_{n-k,n})] - Q[f_2(\tilde{\theta}_{n+1-k,n+1})]| \le 0.0068.$$

Proof. (i) From (2.21) and (2.22), it is easily seen that both $b(\theta) - a(\theta)$ and $b'(\theta) - a'(\theta)$ are positive and increasing in $0 < \theta < \pi$. Straightforward computation gives

$$(7.35)\quad b(0.27\pi) - a(0.27\pi) = 0.30842\ldots, \qquad b'(0.27\pi) - a'(0.27\pi) = 0.42909\ldots.$$

By the mean value theorem,

$$\cos\varphi_n - \cos\varphi_{n+1} = (-\sin\xi)(\varphi_n - \varphi_{n+1}),$$

where ξ lies between φ_n and φ_{n+1}. From (7.33), we obtain

$$|\sin\xi| < |\xi| < \frac{1}{4n}|a(0.27\pi) - b(0.27\pi)|$$

$$+ \frac{1}{8\pi(n-k-\frac{1}{4})} + |\varepsilon_{13}| \le 0.01371.$$

From (7.33), it also follows that

$$|\varphi_n - \varphi_{n+1}| \le \left|\frac{1}{4(n+1)}\left[a\left(\frac{j_{0,n+1-k}}{n+1}\right) - b\left(\frac{j_{0,n+1-k}}{n+1}\right)\right]\right.$$

$$(7.36)\qquad \left. - \frac{1}{4n}\left[a\left(\frac{j_{0,n-k}}{n}\right) - b\left(\frac{j_{0,n-k}}{n}\right)\right]\right|$$

$$+ |\delta_{n-k} - \delta_{n+1-k}| + 2|\varepsilon_{13}|.$$

Note that

$$-\frac{0.0772}{n(n+1)} \le \frac{1}{4}\left(\frac{1}{n} - \frac{1}{n+1}\right)\left[a\left(\frac{j_{0,n-k}}{n}\right) - b\left(\frac{j_{0,n-k}}{n}\right)\right] < 0,$$

and that

$$0 < \frac{1}{4(n+1)}\left\{\left[a\left(\frac{j_{0,n-k}}{n}\right) - b\left(\frac{j_{0,n-k}}{n}\right)\right] - \left[a\left(\frac{j_{0,n+1-k}}{n+1}\right) - b\left(\frac{j_{0,n+1-k}}{n+1}\right)\right]\right\}$$

$$= \frac{1}{4(n+1)}[a'(\eta) - b'(\eta)]\left[\frac{j_{0,n+1-k}}{n+1} - \frac{j_{0,n-k}}{n}\right] \le \frac{0.3371}{n(n+1)}$$

in view of (7.11) and (7.35). Therefore the first term on the right-hand side of (7.36) is dominated by $0.3371/n(n+1)$. Furthermore, since $j_{0,m+1} - j_{0,m} < \pi$ for all $m \ge 1$, we have from the equation preceding (7.10)

$$0 < \delta_{n-k} - \delta_{n+1-k} < \frac{1}{8\pi(n-k-\frac{1}{4})} < \frac{0.0398}{n-k-\frac{1}{4}}.$$

Consequently,

$$|\varphi_n - \varphi_{n+1}| \le \frac{0.0139}{n+1} + \frac{0.0398}{n-k-\frac{1}{4}}$$

and

$$|\cos\varphi_n - \cos\varphi_{n+1}| \le \frac{0.0002}{n+1} + \frac{0.0006}{n-k-\frac{1}{4}}.$$

(ii) Similar to (i), we have

$$|\sin\varphi_n - \sin\varphi_{n+1}| \le |\varphi_n - \varphi_{n+1}| \le \frac{0.0139}{n+1} + \frac{0.0398}{n-k-\frac{1}{4}}.$$

(iii) Equation (7.30) gives

$$|P[f_2(\tilde{\theta}_{n-k,n})] - P[f_2(\tilde{\theta}_{n+1-k,n+1})]| \le |\eta_1[f_2(\tilde{\theta}_{n-k,n})] - \eta_1[f_2(\tilde{\theta}_{n+1-k,n+1})]|.$$

Since $\eta_1(x)$ is positive and $f_2(\theta)$ is increasing, we also have from (7.30) and Lemma 15

$$|P[f_2(\tilde{\theta}_{n-k,n})] - P[f_2(\tilde{\theta}_{n+1-k,n+1})]| \le \frac{15}{128}[f_2(\theta_{n-k,n})]^{-2} \le \frac{15}{128(j_{0,n-k})^2},$$

which, in view of (7.3), yields

$$|P[f_2(\tilde{\theta}_{n-k,n})] - P[f_2(\tilde{\theta}_{n+1-k,n+1})]| \le \frac{0.0119}{(n-k-\frac{1}{4})^2} \le \frac{0.0191}{(n-k+\frac{3}{4})^2}.$$

(iv) Similarly, since $\eta_2(x)$ is negative and $f_2(\theta)$ is increasing, (7.31) gives

$$|Q[f_2(\tilde{\theta}_{n-k,n})] - Q[f_2(\tilde{\theta}_{n+1-k,n+1})]| \le \frac{3}{8}\frac{f_2(\tilde{\theta}_{n+1-k,n+1}) - f_2(\tilde{\theta}_{n-k,n})}{f_2(\tilde{\theta}_{n+1-k,n+1})f_2(\tilde{\theta}_{n-k,n})} + \frac{105}{1024}[f_2(\tilde{\theta}_{n-k,n})]^{-3}.$$

By Lemmas 14 and 15,

$$|Q[f_2(\tilde{\theta}_{n-k,n})] - Q[f_2(\tilde{\theta}_{n+1-k,n+1})]| \le 0.0068,$$

thus proving the lemma. □

Returning to (7.27), we write

$$D_{31} = D_{31}^{(1)} + D_{32}^{(2)}, \tag{7.37}$$

where

$$D_{31}^{(1)} = J_1[f_2(\tilde{\theta}_{n-k,n})] \cdot \left\{1 - \left[\frac{f_2(\tilde{\theta}_{n-k,n})}{f_2(\tilde{\theta}_{n+1-k,n+1})}\right]^{\frac{1}{2}}\right\} \tag{7.38}$$

and

$$D_{32}^{(2)} = \left[\frac{f_2(\tilde{\theta}_{n-k,n})}{f_2(\tilde{\theta}_{n+1-k,n+1})}\right]^{\frac{1}{2}} \cdot J_1[f_2(\tilde{\theta}_{n-k,n})] + J_1[f_2(\tilde{\theta}_{n+1-k,n+1})]. \tag{7.39}$$

By Lemmas 14 and 15,

$$0 \le 1 - \left[\frac{f_2(\tilde{\theta}_{n-k,n})}{f_2(\tilde{\theta}_{n+1-k,n+1})}\right]^{\frac{1}{2}} = \frac{f_2(\tilde{\theta}_{n+1-k,n+1}) - f_2(\tilde{\theta}_{n-k,n})}{\sqrt{f_2(\tilde{\theta}_{n+1-k,n+1})} \cdot \left[\sqrt{f_2(\tilde{\theta}_{n+1-k,n+1})} + \sqrt{f_2(\tilde{\theta}_{n-k,n})}\right]}$$

$$\le \frac{3.1447}{j_{0,n+1-k}[1 + \sqrt{j_{0,n-k}}/\sqrt{j_{0,n+1-k}}]}$$

for $n - K_3' \le k \le n-1$. We now restrict k to the smaller range $n - K_3' \le k \le n-4$. For k in this range, we have from (7.3)

$$\sqrt{\frac{j_{0,n-k}}{j_{0,n+1-k}}} \ge \left\{\frac{n-k-0.25}{n-k+0.7527}\right\}^{\frac{1}{2}} \ge 0.8882.$$

Hence

$$0 \le 1 - \left[\frac{f_0(\tilde{\theta}_{n-k,n})}{f_0(\tilde{\theta}_{n+1-k,n+1})}\right]^{\frac{1}{2}} \le \frac{0.5310}{n-k+\frac{3}{4}}$$

and

$$|D_{31}^{(1)}| \le \frac{0.5310}{n-k+\frac{3}{4}} |J_1[f_2(\tilde{\theta}_{n-k,n})]|, \qquad n - K_3' \le k \le n-4. \tag{7.40}$$

Inserting (7.32) in (7.39) gives

$$D_{31}^{(2)} = (-1)^{n-k+1} \cdot \left[\frac{2}{\pi f_2(\tilde{\theta}_{n+1-k,n+1})}\right]^{\frac{1}{2}}$$

$$\cdot \{\cos\varphi_n \cdot P[f_2(\tilde{\theta}_{n-k,n})] - \cos\varphi_{n+1} \cdot P[f_2(\tilde{\theta}_{n+1-k,n+1})]$$

$$+ \sin\varphi_n \cdot Q[f_2(\tilde{\theta}_{n-k,n})] - \sin\varphi_{n+1} \cdot Q[f_2(\tilde{\theta}_{n+1-k,n+1})]\},$$

which in turn yields

$$|D_{31}^{(2)}| \le \left[\frac{2}{\pi f_2(\tilde{\theta}_{n+1-k,n+1})}\right]^{\frac{1}{2}}$$
$$\cdot \{|\cos\varphi_n - \cos\varphi_{n+1}| \cdot P[f_2(\tilde{\theta}_{n-k,n})]$$

$$+ |P[f_2(\tilde{\theta}_{n-k,n})] - P[f_2(\tilde{\theta}_{n+1-k,n+1})]| + |\sin\varphi_n - \sin\varphi_{n+1}| \cdot |Q[f_2(\tilde{\theta}_{n-k,n})]|$$

$$+ |\sin\varphi_{n+1}| \cdot |Q[f_2(\tilde{\theta}_{n-k,n})] - Q[f_2(\tilde{\theta}_{n+1-k,n+1})]|\}.$$

Now we apply Lemma 16 and approximations (7.30) and (7.31). Since $|\sin\varphi_{n+1}| < |\varphi_{n+1}|$, it follows from (7.33) and (7.35) that

$$
\begin{aligned}
|D_{31}^{(2)}| \le{}& \left[\frac{2}{\pi f_2(\tilde\theta_{n+1-k,n+1})}\right]^{\frac12} \\
&\cdot\left\{\left(\frac{0.0002}{n+1}+\frac{0.0006}{n-k-\frac14}\right)\left[1+\frac{0.0191}{(n-k+\frac34)^2}\right]\right. \\
&+\frac{0.0191}{(n-k+\frac34)^2}+\left(\frac{0.0139}{n+1}+\frac{0.0398}{n-k-\frac14}\right)\cdot\frac{0.1194}{n-k-\frac14} \\
&\left.+\left(\frac{0.0773}{n+1}+\frac{0.0398}{n-k+\frac34}\right)\times 0.0068\right\} \\
\le{}& \left(\frac{0.0008}{n+1}+\frac{0.0068}{n-k+\frac34}\right)\left[\frac{2}{\pi f_2(\tilde\theta_{n+1-k,n+1})}\right]^{\frac12}
\end{aligned}
$$

for $n-K_3' \le k \le n-4$. From (4.30), we obtain

$$
\begin{aligned}
\frac{J_1[f_2(\tilde\theta_{n-k,n})]}{J_1(j_{0,n-k})} ={}& 1-\frac{J_1'(j_{0,n-k})}{J_1(j_{0,n-k})}\left\{\frac{1}{4n}\left[a\left(\frac{j_{0,n-k}}{n}\right)-b\left(\frac{j_{0,n-k}}{n}\right)\right]-\varepsilon_{13}\right\} \\
&+\frac12\frac{J_1''(\xi)}{J_1(j_{0,n-k})}\left\{\frac{1}{4n}\left[a\left(\frac{j_{0,n-k}}{n}\right)-b\left(\frac{j_{0,n-k}}{n}\right)\right]-\varepsilon_{13}\right\}^2,
\end{aligned}
\tag{7.41}
$$

where ξ lies between $j_{0,n-k}$ and $f_2(\tilde\theta_{n-k,n})$. As in Lemma 5, it can be shown that

$$
|J_1'(j_{0,n-k})| = \frac{1}{j_{0,n-k}}|J_1(j_{0,n-k})| \tag{7.42}
$$

and

$$
|J_1''(\xi)| \le \frac34|J_1'(\xi)|+\frac14|J_3(\xi)| \le \frac{1}{\sqrt2}. \tag{7.43}
$$

From [9, p. 166], we also have

$$
(-1)^{k+1}j_{0,k}J_1(j_{0,k}) = \sqrt{\frac{2}{\pi}}\left(j_{0,k}^{\frac12}+\frac{1}{10}j_{0,k}^{-\frac32}\right)+\delta_2, \tag{7.44}
$$

where

$$
|\delta_2| \le \begin{cases} 0.1251k^{-\frac52}, & k \ge 2, \\ 0.0819k^{-\frac52}, & k \ge 25, \end{cases}
$$

which in turn gives

$$
|J_1(j_{0,k})| \ge 0.7977j_{0,k}^{-\frac12}, \qquad k \ge 20. \tag{7.45}
$$

The numerical values of $J_1(j_{0,k})$ given in [1, p. 409] shows that (7.45) in fact holds for $k \geq 1$. Since $\frac{1}{\theta}[b(\theta) - a(\theta)]$ is positive and increasing in $0 < \theta < \pi$, a combination of (7.41), (7.42), (7.43), and (7.45) yields

$$|J_1[f_2(\tilde{\theta}_{n-k,n})]| \geq 0.9998|J_1(j_{0,n-k})| \geq 0.7975(j_{0,n-k})^{-\frac{1}{2}}. \tag{7.46}$$

Therefore by Lemma 15,

$$|D_{31}^{(2)}| \leq \left(\frac{0.0009}{n+1} + \frac{0.0069}{n-k+\frac{3}{4}}\right)|J_1[f_2(\tilde{\theta}_{n-k,n})]|. \tag{7.47}$$

From (7.37) and (7.40), it follows that

$$|D_{31}| \leq \left(\frac{0.5379}{n-k+\frac{3}{4}} + \frac{0.0009}{n+1}\right)|J_1[f_2(\tilde{\theta}_{n-k,n})]|. \tag{7.48}$$

The estimation of D_{32} in (7.28) is considerably simpler. From (4.29), (7.46), and (7.45), we have

$$|D_{32}| \leq 2\sqrt{2}|I(n)| \leq \frac{0.0832}{0.9998}\frac{n^{-\frac{7}{2}}}{|J_1(j_{0,n-k})|}|J_1[f_2(\tilde{\theta}_{n-k,n})]|$$

$$\leq 0.1044\sqrt{\frac{j_{0,n-k}}{n}}n^{-3}|J_1[f_2(\tilde{\theta}_{n-k,n})]| \tag{7.49}$$

$$\leq \frac{0.0002}{n+1}|J_1[f_2(\tilde{\theta}_{n-k,n})]|.$$

In the last inequality, we have also made use of (3.34). Coupling (7.48) and (7.49) gives

$$|D_3| \leq |D_{31}| + |D_{32}| \leq \left(\frac{0.5379}{n-k+\frac{3}{4}} + \frac{0.0011}{n+1}\right)|J_1[f_2(\tilde{\theta}_{n-k,n})]|. \tag{7.50}$$

LEMMA 17. *For $n - K_3' \leq k \leq n-4$ and $n \geq 25$, the quantity D in* (7.12) *satisfies*

$$\frac{D}{J_1(j_{0,n-k})} \leq -\frac{0.1747}{n-k+\frac{3}{4}}g_3(\tilde{\theta}_{n+1-k,n+1}) < 0.$$

Proof. Similar to (7.46), one can show from (7.41) that

$$\frac{|J_1[f_2(\tilde{\theta}_{n-k,n})]|}{|J_1(j_{0,n-k})|} \leq 1.0002 \tag{7.51}$$

and, in view of (7.44),

$$\operatorname{sgn}\{J_1[f_2(\tilde{\theta}_{n-k,n})]\} = \operatorname{sgn}\{J_1(j_{0,n-k})\} = (-1)^{n-k+1}. \tag{7.52}$$

Since $g_4(0.27\pi) < g_4(\theta) < 0$ and $|g_4(0.27\pi)| = 0.0573$, it can be shown from (4.35) that

$$\left\{1 + \frac{1}{32n^2}g_4(\tilde{\theta}_{n-k,n}) + \varepsilon_{14}(n)\right\} \geq 0.9999$$

and

$$\left\{1+\frac{1}{32(n+1)^2}g_4(\tilde{\theta}_{n+1-k,n+1})+\varepsilon_{14}(n+1)\right\}\le 1.0001.$$

By (4.29) and (7.45), we also have

$$\frac{\sqrt{2}|I(n)|}{|J_1(j_{0,n-k})|}\le\frac{\sqrt{2}}{0.7977}\sqrt{\frac{j_{0,n-k}}{n}}\cdot\frac{0.0294}{n^3}\le 0.0001,$$

from which it follows that

$$\left\{\frac{J_1[f_2(\tilde{\theta}_{n-k,n})]}{J_1(j_{0,n-k})}-\frac{\sqrt{2}\,I(n)}{J_1(j_{0,n-k})}\right\}\ge 0.9997$$

and

$$\left\{\frac{J_1[f_2(\tilde{\theta}_{n-k,n})]}{J_1(j_{0,n-k})}-\frac{\sqrt{2}\,I(n)}{J_1(j_{0,n-k})}\right\}\le 1.0003$$

on account of (7.46) and (7.51). Now insert these estimates in (7.13) and make use of (7.23), (7.25), and (7.50). The result is

$$\frac{D}{J_1(j_{0,n-k})}\le\left\{-\frac{(k+\frac{1}{4})\pi-(n+1)\delta_{n-k}+n\delta_{n+1-k}}{nj_{0,n+1-k}}\times 0.9831\right.$$

$$\left.+\frac{0.0195}{n^3}\times 1.0003+\frac{0.5379}{n-k+\frac{3}{4}}+\frac{0.0011}{n+1}\right\}\cdot g_3(\tilde{\theta}_{n+1-k,n+1}).$$

Since $(k+\frac{1}{4})\ge(0.73n)$ by (7.02), using (7.3) and (7.7), we obtain

$$\frac{D}{J_1(j_{0,n-k})}\le\left\{-\frac{0.7138}{n-k+\frac{3}{4}}+\frac{0.5378}{n-k+\frac{3}{4}}+\frac{0.0014}{n-k+\frac{3}{4}}\right\}g_3(\tilde{\theta}_{n+1-k,n+1}),$$

$$\le-\frac{0.1746}{n-k+\frac{3}{4}}g_3(\tilde{\theta}_{n+1-k,n+1})<0,$$

thus proving the lemma. □

LEMMA 18. *For $k=n-3,n-2$ and $n-1$, and for $n\ge 25$, the quantity D in* (7.12) *satisfies*

$$\frac{D}{J_1(j_{0,n-k})}\le-\frac{0.1038}{n-k+\frac{3}{4}}\cdot g_3(\tilde{\theta}_{n+1-k,n+1})<0.$$

Proof. The numerical table in [1, p. 409] shows that the remainder δ_m in the equation preceding (7.10) satisfies

$$0\le\delta_1\le 0.0487,\qquad 0\le\delta_2\le 0.0223,\qquad 0\le\delta_3\le 0.0144.$$

Hence we have from (7.23)

$$D_1\le-\frac{(n-3)\pi+\frac{1}{4}\pi-(n+1)\times 0.0487}{[(n-k+\frac{3}{4})\pi+0.0223]\cdot n}\times 0.9835\times g_3(\tilde{\theta}_{n+1-k,n+1})$$

$$\le-\frac{0.8461}{n-k+\frac{3}{4}}g_3(\tilde{\theta}_{n+1-k,n+1}).$$

By (7.25),

$$|D_2| \leq \frac{0.0195}{n^3}.$$

Similar to (7.40), (7.47), and (7.49), one can also verify that

$$|D_{31}^{(1)}| \leq \frac{0.6069}{n-k+\frac{3}{4}}|J_1[f_2(\tilde{\theta}_{n-k,n})]|,$$

$$|D_{31}^{(2)}| \leq \left(\frac{0.0085}{n+1} + \frac{0.0884}{n-k+\frac{3}{4}}\right) \cdot |J_1[f_2(\tilde{\theta}_{n-k,n})]|,$$

and

$$|D_{32}| \leq \frac{0.0002}{n+1}|J_1[f_2(\tilde{\theta}_{n-k,n})]|.$$

As in Lemma 17, the final result now follows from (7.13). □

THEOREM 6c. *For $n - K_3' \leq k \leq n-1$ and $n \geq 25$, we have*

$$|\nu_{k,n}| < |\nu_{k,n+1}|.$$

Proof. As in the proof of Theorem 6a, it can be shown that

$$\operatorname{sgn}\{J_1(j_{0,n-k})\} = (-1)^{n-k+1}$$

and

$$\operatorname{sgn}\{\nu_{k,n}\} = (-1)^k.$$

Hence by (7.12) and Lemmas 17 and 18,

$$\operatorname{sgn}\{\nu_{k,n} - \nu_{k,n+1}\} = (-1)^{n+1}\operatorname{sgn}\{D\}$$

$$= (-1)^n \operatorname{sgn}\{J_1(j_{0,n-k})\} = (-1)^{k+1}$$

for $n \geq 25$. Therefore

$$\operatorname{sgn}\{|\nu_{k,n}| - |\nu_{k,n+1}|\} = -1,$$

and the theorem is proved. □

8. Conclusion. We are now ready to state and prove the following main result of this paper.

THEOREM 7. *For $n = 1, 2, \ldots$ and $k = 1, \ldots, n$, we have*

$$|\nu_{k,n}| < |\nu_{k,n+1}|. \tag{8.1}$$

Proof. From the reflection formula [11, p. 59] and the identity in (4.4), it is easily seen that

$$P_n^{(0,-1)}(x) = \frac{x+1}{2}P_{n-1}^{(0,1)}(x).$$

Since $y_{n,n} = -1$, we have $\nu_{n,n} = 0$, i.e., (8.1) obviously holds for $k = n$. Therefore we need to consider only the case $k = 1, \ldots, n-1$ and $n = 2, 3, \ldots$. For $2 \leq n \leq 25$, the validity of (8.1) is evident from the values of $\nu_{k,n}$, which can be obtained by direct computation. (A numerical table of these values is available upon request.)

For $n \geq 25$, the results in Theorems 6a, 6b, and 6c show that (8.1) holds for $1 \leq k \leq K_2$ and $n - K_3' \leq k \leq n - 1$. Thus, to complete the proof, it suffices to show that

$$n - K_3' \leq K_2 + 1. \tag{8.2}$$

We shall prove this by contradiction. Assume that (8.2) is not true. Then there exists a positive integer k_0 such that either

$$k_0 > K_2 \quad \text{and} \quad k_0 < n - K_3', \tag{8.3}$$

or

$$k_0 > K_2 + 1 \quad \text{and} \quad k_0 - 1 < n - K_3'. \tag{8.4}$$

If (8.3) holds, then by (7.1) and (3.23)

$$\frac{j_{0,n+1-k_0}}{n+1} > 0.27\pi - \frac{1}{16(n+1)^2}\hat{f}_1(0.27\pi) - \frac{\rho_3'}{n+1} \tag{8.5}$$

and

$$\frac{j_{1,k_0}}{n} > 0.78\pi + \frac{1}{16n^2}\hat{f}_2(0.78\pi) - \frac{\rho_2}{n}. \tag{8.6}$$

Adding up (8.5) and (8.6) gives

$$\frac{j_{1,k_0}}{n} + \frac{j_{0,n+1-k_0}}{n+1} > 1.0499\pi. \tag{8.7}$$

On the other hand, it follows from (7.3) and the inequality [7]

$$j_{1,k} \leq (k + \tfrac{1}{4})\pi, \qquad k = 1, 2, \ldots,$$

that

$$\frac{j_{1,k_0}}{n} + \frac{j_{0,n+1-k_0}}{n+1} < 1.0388\pi, \tag{8.8}$$

which contradicts (8.7). A similar contradiction results if (8.4) holds. This completes the proof of the theorem. □

REFERENCES

[1] M. ABRAMOWITZ AND I. STEGUN, ED., *Handbook of Mathematical Functions,* National Bureau of Standards Appl. Math. Ser. 55, Washington DC, 1964.

[2] R. ASKEY, *Graphs as an aid to understanding special functions,* in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, 1990, pp. 3–33.

[3] R. ASKEY AND G. GASPER, *Positive Jacobi polynomial sums,* II, Amer. J. Math., 98 (1976), pp. 709–737.

[4] P. BARATELLA AND L. GATTESCHI, *The bounds for the error terms of an asymptotic approximation of Jacobi polynomials,* in Orthogonal Polynomials and Their Applications, M. Alfaro et al., eds., Lecture Notes in Math., Vol. 1329, Springer-Verlag, Berlin, New York, 1988, pp. 203–221.

[5] L. GATTESCHI, *Limitazione degli errori nelle formule asintotiche per le funzioni speciali,* Rend. Sem. Mat. Univ. Politec. Torino, 16 (1956–1957), pp. 83–97.

[6] H. W. HETHCOTE, *Error bounds for asymptotic approximations of zeros of transcendental functions,* SIAM J. Math. Anal., 1 (1970), pp. 147–152.

[7] ———, *Bounds for zeros of some special functions,* Proc. Amer. Math. Soc., 25 (1970), pp. 72–74.

[8] F. W. J. OLVER, *Asymptotics and Special Functions,* Academic Press, New York, 1974.

[9] C. K. QU AND R. WONG, *Szegö's conjecture on Lebesgue constants for Legendre series,* Pacific J. Math., 135 (1988), pp. 157–188.

[10] O. SZÁSZ, *On the relative extrema of ultraspherical polynomials,* Bol. Un. Mat. Ital., 5 (1950), pp. 125–127.

[11] G. SZEGÖ, *Orthogonal Polynomials,* Amer. Math. Soc. Colloq. Publ. 23, American Mathematical Society, Providence, RI, 1939, fourth ed., 1975.

[12] ———, *On the relative extrema of Legendre polynomials,* Bol. Un. Mat. Ital., 5 (1950), pp. 120–121.

[13] J. TODD, *On the relative extrema of the Laguerre orthogonal functions,* Bol. Un. Mat. Ital., 5 (1950), pp. 122–125.

[14] G. N. WATSON, *Theory of Bessel Functions,* Cambridge University Press, Cambridge, 1944.

[15] R. WONG, *Asymptotic Approximations of Integrals,* Academic Press, Boston, MA, 1989.

[16] R. WONG AND T. LANG, *On the points of inflection of Bessel functions of positive order,* II, Canad. J. Math., 43 (1991), pp. 628–651.

SIAM J. MATH. ANAL.
Vol. 25, No. 2, pp. 812–814, March 1994

034

TOWARDS A WZ EVOLUTION OF THE MEHTA INTEGRAL*

DORON ZEILBERGER†

Abstract. The celebrated Mehta integral is shown to be equivalent to a simple algebraic-differential identity, which is completely routine for any fixed number of variables.

Key words. WZ form, computer-assisted proof of identities

AMS subject classifications. 33A30, 33A99

Askey [As] proposed the problem of proving the Mehta (see [M]) integral identity

(Mehta)
$$\frac{1}{(2\pi)^{n/2}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-x_1^2/2-\cdots-x_n^2/2)\prod_{1\le i<j\le n}(x_i-x_j)^{2c}dx_1\cdots dx_n=\prod_{j=1}^{n}\frac{(cj)!}{j!}$$

without using Selberg's integral (see [M]). This problem was solved by Anderson [An]. Here we use the method of [WZ1] and [WZ2] to initiate another Selberg-free proof that we believe is of independent interest. We show that (Mehta) for any given n is equivalent to the following elegant identity ($d:=n(n-1)/2$):

(Mehta′)
$$\left\{\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2}+2c\sum_{1\le i<j\le n}\frac{1}{x_i-x_j}\left(\frac{\partial}{\partial x_i}-\frac{\partial}{\partial x_j}\right)\right\}^{d}\prod_{1\le i<j\le n}(x_i-x_j)^2=2^d d!n!\cdot\prod_{1\le s<r\le n}(rc+s).$$

(Mehta′) is purely routine for any specific n, but at the time of writing we are unable to prove it directly for general n. Of course, we do have *a* proof, since we are going to show that (Mehta′) and (Mehta) are equivalent, but what we are after is a *direct* proof. The author is offering a prize of 25 US dollars for such a proof.

The present method also obviously extends to the Macdonald–Mehta integral [M], which was proved by Beckner and Regev [BR] for the classical root systems (see [M]), by Garvan [G] for the exceptional root system F_4, and by Opdam [O] for E_6, E_7, and E_8. It follows that the present approach should also yield new proofs for all the exceptional root systems, at least in principle, but most likely in practice also. More important, it seems to have a high chance of producing a uniform, intrinsic, classification-independent proof. We leave to the reader, as an instructive exercise, the task of finding the root-system analog of (Mehta′) that is equivalent to the now-proved Mehta–Macdonald conjecture, and we offer an additional 25 dollars for an intrinsic proof.

Our proposed proof of (Mehta) will be a *derivation* rather than a *verification*, and will follow the method of [WZ2]. Let us call the left of (Mehta) $L(c)$, and the integrand $F(c;x_1,\ldots,x_n)$. We know from the general theory of [WZ2] that for some r

* Received by the editors April 2, 1992; accepted for publication (in revised form) May 11, 1993.
† Department of Mathematics, Temple University, Philadelphia, Pennsylvania 19122. This work was supported in part by the National Science Foundation.

and some rational functions $P_1, \dots, P_n$, in $(c, x_1, \dots, x_n)$, and some rational functions in $c, a_0(c), \dots, a_r(c)$,

$$\text{(WZ)} \qquad \sum_{s=0}^{r} a_s(c) F(c+s; x_1, \dots, x_n) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}(P_i F).$$

Let us be optimistic and try out $r = 1$. Without loss of generality, set $a_1 := 1$. Substituting F in (WZ), performing all differentiations, and dividing throughout by F leads to the following equation for P_i and a_0:

$$(1) \qquad a_0 + \prod_{1 \le i < j \le n} (x_i - x_j)^2 = \sum_{i=1}^{n} \frac{\partial P_i}{\partial x_i} + 2c \sum_{i=1}^{n} \left(\sum_{j \ne i} \frac{1}{x_i - x_j} \right) P_i - \sum_{i=1}^{n} x_i P_i.$$

To be even more optimistic, assume that P_i are polynomials, rather than mere rational functions, in their dependence on $(x_1, \dots, x_n)$, and are furthermore the components of the gradient of another polynomial P, i.e., $P_i = \partial P / \partial x_i$, $i = 1, \dots, n$, for some polynomial P. Equation (1) then becomes

(2)

$$a_0 + \prod_{1 \le i < j \le n} (x_i - x_j)^2 = \left\{ \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} + 2c \sum_{i=1}^{n} \left(\sum_{j \ne i} \frac{1}{x_i - x_j} \right) \frac{\partial}{\partial x_i} \right\} P - \sum_{i=1}^{n} \left(x_i \frac{\partial}{\partial x_i} \right) P.$$

If we can find such a polynomial P, and compute the corresponding a_0, then it would follow from (WZ), upon integrating with respect to $x_1, \dots, x_n$, that $L(c)$ satisfies the recurrence $L(c+1) = -a_0(c) L(c)$, which combined with $L(0) = 1$ would enable one to find $L(c)$. Note that the mere existence of a_0, which we will shortly prove, is given by the left side of (Mehta′) and implies that $L(c)$ is of *closed form*, which from a theoretical point of view is almost as good as knowing what it is exactly.

Let us write P as a sum of its homogeneous parts

$$P = \sum_{j=2}^{2d} P^{(j)}, \quad P^{(j)} \text{ homog. of deg. } j,$$

where, as above, d equals $n(n-1)/2$. Denote the operator inside the braces of (2) or (Mehta′) by $\mathbf{Z}$. Using Euler's formula, we get

(3)

$$a_0 + \prod_{1 \le i < j \le n} (x_i - x_j)^2 = \sum_{j=2}^{2d} \mathbf{Z} P^{(j)} - \sum_{j=2}^{2d} j P^{(j)} = (-2d) P^{(2d)} + \sum_{j=2}^{2d} (\mathbf{Z} P^{(j)} - (j-2) P^{(j-2)}).$$

By equating corresponding homogeneous parts, we get

(4)

$$P^{(2d)} = -(2d)^{-1} \prod_{1 \le i < j \le n} (x_i - x_j)^2, \quad P^{(j-2)} = \frac{1}{j-2} \mathbf{Z} P^{(j)}, \quad j = 2d, 2d-2, \dots, 4.$$

It is easy to see that $\mathbf{Z}$ maps homogeneous symmetric polynomials to homogeneous symmetric polynomials, and that it reduces the degree by 2. Iterating (4) and comparing the constant part of (3) finally yields that both P and a_0 indeed exist, and that

$$a_0 = -(2^d d!)^{-1} \mathbf{Z}^d \left[\prod_{1 \le i < j \le n} (x_i - x_j)^2 \right].$$

Hence Mehta's integral is indeed expressible in closed form, and proving that its value coincides with the value implied by (Mehta) amounts to proving (Mehta′).

The referee empirically noticed the following generalization. For a, any positive integer, we have

$$\text{(Mehta}'') \quad \left\{ \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} + 2c \sum_{1 \le i < j \le n} \frac{1}{x_i - x_j} \left(\frac{\partial}{\partial x_i} - \frac{\partial}{\partial x_j} \right) \right\}^{ad} \prod_{1 \le i < j \le n} (x_i - x_j)^{2a}$$
$$= 2^{ad} (da)! n! a \prod_{2 \le r \le n} \prod_{1 \le s \le ra, s \ne 0 \bmod r} (rc + s).$$

It turns out that (Mehta″) follows from (Mehta) exactly the same way as (Mehta′); just replace c by ca, in (Mehta), and repeat the argument! In particular, it follows that (Mehta′) implies (Mehta″), albeit indirectly, via (Mehta).

Acknowledgment. Many thanks are due to Herb Wilf for countless discussions and inspiration.

REFERENCES

[An] G. ANDERSON, *Letter to R. Askey.*

[As] R. A. ASKEY, *Lost and found mathematics*, joint MAA-AAAS invited talk, Columbus, OH, August 1990.

[BR] W. BECKNER AND A. REGEV, *manuscript*, 1980.

[G] F. G. GARVAN, *Some Macdonald–Mehta integrals by brute force*, in q-Series and Partitions, D. Stanton, ed., IMA volumes 18, Springer-Verlag, New York, pp. 77–98.

[M] I. G. MACDONALD, *Some conjectures for root systems*, SIAM J. Math. Anal., 13 (1982), pp. 988–1007.

[O] E. OPDAM, *Some applications of hypergeometric shift operators*, Invent. Math., 98 (1989), pp. 1–18.

[WZ1] H. S. WILF AND D. ZEILBERGER, *Rational function certification of hypergeometric multi-integral/sum/"q" identities*, Bull. Amer. Math. Soc., 24 (1992), pp. 143–148.

[WZ2] ———, *An algorithmic proof theory for hypergeometric (ordinary and "q") multi-sum/integral identities*, Invent. Math., 108 (1992), pp. 575–633.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 815–835, May 1994

ASYMPTOTIC STABILITY FOR INTERMITTENTLY CONTROLLED NONLINEAR OSCILLATORS*

PATRIZIA PUCCI† AND JAMES SERRIN‡

Abstract. The authors prove a number of asymptotic stability theorems for intermittently damped quasi-variational systems, extending and generalizing previous work on the subject.

Key words. global asymptotic stability, intermittent damping, control set

AMS subject classifications. 34DXX, 35A15

1. Introduction.

The problem of global asymptotic stability of solutions of second order equations with intermittent damping has been studied by Smith, Thurston and Wong, Artstein and Infante, and Hatvani and Totik. In this paper we give various generalizations of this work and extensions to quasi-variational systems.

As in our earlier work [5], [7] on asymptotic stability, we consider vector unknowns $u : J \to \mathbb{R}^N$ and systems having the general form

$$\big(\nabla \mathcal{L}(t,u,u')\big)' - \nabla_u \mathcal{L}(t,u,u') = Q(t,u,u'), \qquad t \in J, \tag{1.1}$$

where J is a half open interval of the form $[T,\infty)$ and $\mathcal{L}(t,u,p) = G(u,p) - F(t,u)$, and where G, F, Q are given continuously differentiable functions. The most important of the conditions which will be imposed on (1.1) are that

$$G(u,\cdot) \text{ is strictly convex in } \mathbb{R}^N; \qquad G(u,0) = 0, \quad \nabla G(u,0) = 0, \tag{1.2}$$

$$\big(\nabla_u F(t,u), u\big) > 0 \quad \text{for } u \neq 0; \qquad F(t,u) = 0, \tag{1.3}$$

$$\big(Q(t,u,p), p\big) \leq 0. \tag{1.4}$$

Here $(\cdot\,,\cdot)$ denotes the inner product in $\mathbb{R}^N$ and

$$\nabla = \nabla_p = \left(\frac{\partial}{\partial p_1}, \dots, \frac{\partial}{\partial p_N}\right), \qquad \nabla_u = \left(\frac{\partial}{\partial u_1}, \dots, \frac{\partial}{\partial u_N}\right).$$

The function F represents a restoring potential and Q a general nonlinear damping, expressed by (1.4). In §2 we shall give a complete set of hypotheses, while explicit examples are given in [5] and [7].

Since $\nabla G(u,0) = \nabla_u G(u,0) = \nabla_u F(t,0) = Q(t,u,0) = 0$ it is clear that the rest state $u = 0$ is a solution of (1.1). This state is said to be *a global attractor* for the system if any bounded solution u, defined on some interval J, has the property

$$u(t),\ u'(t) \to 0 \quad \text{as } t \to \infty.$$

By the concept of *intermittent damping* we mean that certain restrictions or controls are placed on the damping term on a sequence of nonoverlapping intervals $I_n = [a_n, b_n]$ of J, with $a_n \to \infty$; on the other hand, in the gaps *between* these intervals either no restrictions are imposed or, alternatively, the damping is assumed

* Received by the editors November 23, 1992; accepted for publication April 8, 1993. This research has been partially supported by the Italian Ministero della Universitá e della Ricerca Scientifica e Tecnologica.

† Dipartimento di Matematica, Università di Perugia, 06123 Perugia, Italy.

‡ Department of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

to be bounded from zero but to be otherwise uncontrolled. We emphasize that the intervals I_n may be arbitrarily widely spaced, leaving gaps between them that can be as long as one wishes.

Our purpose is to show that under appropriate conditions on the measures $|I_n|$ and on the damping term $Q(t,u,p)$ for $t \in \cup_1^\infty I_n$, the rest state $u = 0$ becomes a global attractor for (1.1).

From a mechanical point of view the system (1.1) can be considered as the governing law of a holonomic dynamical system, having N degrees of freedom and subject to nonlinear damping. The notion of intermittent damping then occurs if the system is positively damped in the time intervals I_n, but has its damping either *switched off* or *unrestricted* at other times. The system is oscillatory when no damping is present, because $(f(t,u),u) > 0$ for $u \neq 0$; that is, it is not possible to have any solution, other than the trivial one $u = 0$, approaching a limit as $t \to \infty$ (see [5, §5] for a more complete discussion). From this point of view the question we consider is whether the damping which occurs on the time intervals I_n is sufficient to drive the solution to its rest state as $t \to \infty$. The following example provides a specific illustration of this situation in perhaps its simplest form.

Consider the system

$$u'' + A(t,u,u')u' + f(u) = 0, \tag{1.5}$$

where A is a continuous $N \times N$ nonnegative definite matrix and $f(u) = \nabla_u F(u)$. This system arises from (1.1) by the specializations

$$G(p) = \tfrac{1}{2}|p|^2, \qquad Q(t,u,p) = -A(t,u,p)p.$$

We suppose that $(f(u),u) > 0$ for $u \neq 0$, and that A is bounded and uniformly positive definite for $t \in I = \cup_1^\infty I_n$ and (u,p) in any given compact set of $\mathbb{R}^N \times \mathbb{R}^N$; no restrictions, however, other than nonnegativity, are placed on A in the set $J \setminus I$. Then the following rather unexpected result holds:

If the measures of the intervals I_n satisfy

$$\sum_1^\infty |I_n|^3 = \infty, \tag{1.6}$$

then $u = 0$ is a global attractor for (1.5).

The exponent 3 *is the best possible one: that is, without further restrictions no smaller exponent can yield the general conclusion.*

A stronger result is valid if the damping matrix A has the decomposition

$$A(t,u,p) = \beta(t,u,p)I + B(t,u,p), \tag{1.7}$$

where $B(t,u,p)$ is bounded and nonnegative definite for $t \in J$ and (u,p) in any compact set of $\mathbb{R}^N \times \mathbb{R}^N$; the coefficient $\beta(t,u,p)$ is such that for every compact set K of $\mathbb{R}^N \times \mathbb{R}^N$ there exist positive constants β_1, β_2 such that

$$\beta(t,u,p) \geq \beta_1 \quad \text{in } J \times K, \tag{1.8}$$

$$\beta(t,u,p) \leq \beta_2 \quad \text{in } I \times K. \tag{1.9}$$

Then $u = 0$ is a global attractor for (1.5) *provided that*

$$\sum_1^\infty |I_n|^2 = \infty. \tag{1.10}$$

Again the exponent is best possible.

The above results are special cases, respectively, of Corollaries 3 and 4 in §3; see the comments at the end of §3. Indeed, in those results the damping need not even be bounded on I but only have a controlled L^1 norm. Moreover, the constant β_1 in (1.8) can be replaced by a nonnegative measurable function $\hat{\sigma}$ satisfying a *positive mean value criterion*; see condition (2.13) below.

References [1]–[3], [8], and [9] treat the case $N = 1$ of (1.5); moreover, in [8] the coefficient A is independent of u, u' and $f(u)$ is linear. Our results are improvements of the corresponding ones in these papers, even when restricted to the cases treated there.

In §2 we present the setting of the paper and state two important preliminary theorems upon which our further results are based. The main results for the system (1.1) are given in §3 and proved in §§4 and 5. In §6 we present specific examples, showing that the exponents 2 and 3 in the above results are best possible.

2. Preliminaries. We consider vector solutions $u = (u_1, \dots, u_N)$ of the quasi-variational ordinary differential system

$$(2.1) \qquad \big(\nabla G(u,u')\big)' - \nabla_u G(u,u') + f(t,u) = Q(t,u,u'), \qquad t \in J = [T,\infty),$$

where ∇ denotes the gradient operator with respect to the variable p and

$$f(t,u) = \nabla_u F(t,u).$$

It will be supposed throughout the paper that

$$G \in C^1(\mathbb{R}^N \times \mathbb{R}^N; \mathbb{R}), \quad F \in C^1(J \times \mathbb{R}^N; \mathbb{R}), \quad Q \in C(J \times \mathbb{R}^N \times \mathbb{R}^N; \mathbb{R}^N),$$

and also that the following natural conditions hold.

(H_1) $G(u,\cdot)$ is strictly convex in $\mathbb{R}^N$ for all $u \in \mathbb{R}^N$; with $G(u,0) = 0$ and $\nabla G(u,0) = 0$. For all $U > 0$ there exists a positive constant $\Theta = \Theta(U)$ and an exponent $m > 1$ independent of U such that

$$(2.2) \qquad |\nabla G(u,p)| \le \Theta |p|^{m-1} \quad \text{for all } |u| \le U \text{ and } |p| \le 1.$$

(H_2) $F(t,0) = 0$ for all $t \in J$. For all u_0, U with $0 < u_0 \le U$ there exists a constant $\kappa > 0$ and a nonnegative function $\psi \in L^1(J)$ such that

$$(2.3) \qquad \big(f(t,u),u\big) \ge \kappa \quad \text{when } t \in J \quad \text{and} \quad |u| \in [u_0, U],$$

$$(2.4) \qquad |F_t(t,u)| \le \psi(t) \quad \text{when } t \in J(\text{a.e.}) \quad \text{and} \quad |u| \le U.$$

(H_3) $\big(Q(t,u,p),p\big) \le 0$ for all $t \in J$, $u \in \mathbb{R}^N$ and $p \in \mathbb{R}^N$.

If F does not depend on t, then (2.3) follows from the condition $\big(f(u),u\big) > 0$ for $u \ne 0$, while (2.4) is irrelevant. Finally, when $N = 1$ any function f is of gradient type, with $F(t,u) = \int_0^u f(t,s)\,ds$.

Obviously (H_1) is satisfied by any strictly convex homogeneous function $G = G(p)$ of degree $m > 1$, and in particular by the model function $G(p) = |p|^m/m$, $m > 1$; another example is $G(p) = \sqrt{1+|p|^2} - 1$, with $m = 2$. The system (1.5) arises when $G(p) = \frac{1}{2}|p|^2$, with the corresponding exponent $m = 2$.

The next hypothesis places in evidence the concept of a control set $I \subset J$ where the damping term Q is subject to restrictions.

(H_4) For all $U > 0$ there exists a measurable control set $I \subset J$ and a number $\gamma \geq 1$ such that

$$|Q(t,u,p)| \cdot |p| \leq \gamma \left|\big(Q(t,u,p),p\big)\right| \quad \text{for all } t \in I, |u| \leq U \text{ and } p \in \mathbb{R}^N. \tag{2.5}$$

Moreover, there exists a positive measurable damping function $\delta : I \to \mathbb{R}$ and numbers μ, $q > 0$ such that

$$\big(Q(t,u,p),u\big) \leq \delta(t)\,|p|^{\mu} \quad \text{for } t \in I, |u| \leq U \text{ and } |p| \leq q. \tag{2.6}$$

Although I, δ and γ, μ, q may depend on U, for simplicity we do not specifically indicate this dependence. When $N = 1$ condition (2.5) holds automatically with $\gamma = 1$.

In [5] we considered the asymptotic stability of the system (2.1) when the damping magnitude $|Q|$ is controlled from below, but not bounded away from zero. Specifically the following condition was required:

For every $U > 0$ there exist a nonnegative measurable damping control $\sigma : I \to \mathbb{R}$ and an exponent $\nu > 0$ such that

$$|Q(t,u,p)| \geq \sigma(t) \min\{1,\ |p|^{\nu}\} \quad \text{for all } t \in I, |u| \leq U \text{ and } p \in \mathbb{R}^N. \tag{2.7}$$

A further technical hypothesis concerning the function G was also assumed:

For every $U > 0$ and $p_0 > 0$ there is a constant such that

$$\big(\nabla_u G(u,p),u\big) \leq \text{Const.}\ \big(\nabla G(u,p),p\big) \quad \text{whenever } |u| \leq U \text{ and } |p| \geq p_0. \tag{2.8}$$

Note that (2.8) holds whenever $G(u,p) = g(u)\overline{G}(p)$, with $g(u) > 0$ a smooth function in $\mathbb{R}^N$ and $\overline{G}$ satisfying (H_1).

Under the natural assumptions (H_1)–(H_4), together with (2.7) and (2.8), the following result is valid; see [5, Thm. 4.2] and the modified version of this result proved in §3.2 of [4]. This will be the basis for the first main theorem in §3. In its statement we agree that the function δk is extended to all of J by the definition $\delta(t)k(t) = 0$ for $t \in J \setminus I$.

THEOREM A. *Assume that for every $U > 0$ there exists a bounded absolutely continuous function k on J such that*

$$k \notin L^1(J), \qquad k = 0 \quad \text{on } J \setminus I, \tag{2.9}$$

$$0 \leq k \leq \text{Const. } \sigma \text{ on } I, \qquad |k'| \leq \text{Const. } \sigma^{\lambda} k^{1-\lambda} \quad \text{a.e. on } I, \tag{2.10}$$

where

$$\lambda = \begin{cases} \dfrac{m-1}{\nu+1} & \text{if } \nu > m-2, \\ 1 & \text{if } \nu \leq m-2 \quad (\text{and } m > 2). \end{cases} \tag{2.11}$$

Suppose furthermore that there exists a constant $M > 0$ for which

$$\int_T^t \delta(s) k^{\mu+1}(s)\,ds \leq M \int_T^t k(s)\,ds, \qquad t \in J. \tag{2.12}$$

Then the rest state $u = 0$ is a global attractor for the system (2.1).

In [7] we also studied asymptotic stability for the complementary situation in which the damping magnitude $|Q|$ is bounded from zero when $|u|$ and $|p|$ are bounded from zero. In particular, the following condition was assumed:

There is (i) a continuous function $\varphi : \mathbb{R}^N \times \mathbb{R}^N \to [0,\infty)$ with

$$\varphi(u,p) > 0 \quad \text{when } u \neq 0 \text{ and } p \neq 0;$$

and (ii) a measurable function $\hat{\sigma} : J \to [0,\infty)$ and a positive function a on $(0,1)$ satisfying

$$\int_L \hat{\sigma}(t)\,dt \geq a(|L|) > 0 \quad \text{for all intervals } L \subset J \text{ with } |L| \in (0,1);$$

such that for all $U > 0$ there holds

$$\big|\big(Q(t,u,p),p\big)\big| \geq \hat{\sigma}(t)\,\varphi(u,p) \quad \text{for all } t \in J, |u| \leq U \text{ and } |p| \leq q, \tag{2.13}$$

where $q = q(U) > 0$ is given in (H_4).

This is in fact condition (H_3) of [7], in the weaker version involving (ii) which was given in §7 of [7]. (In this context, condition (ii) was first introduced by Hatvani [2].) If $Q(t,u,p) = \hat{\sigma}(t)\,\hat{Q}(u,p)$ and $\big(\hat{Q}(u,p),p\big) < 0$ for $u,p \neq 0$, it is easy to see that (2.13) is satisfied.

Two further technical hypotheses were introduced in [7]; they are required only when $N > 1$, though in fact the second, (V_2), automatically holds when $N = 1$ with $\varepsilon(p) = 0$, $g(u) = \frac{1}{2}u^2$ and $C = 0$ in view of (H_1) and (H_3); see also [7, Lemma 2.1].

(V_1) For all $U > 0$ and $p_0 > 0$ there is a nonnegative measurable function $h \notin L^1(J)$ such that

$$\big|\big(Q(t,u,p),p\big)\big| \geq h(t) \quad \text{for all } t \in J, |u| \leq U \text{ and } |p| \geq p_0.$$

(V_2) For all $U > 0$ there exists a continuous function $\varepsilon(p)$ with $\varepsilon(0) = 0$, such that

$$\big(Q(t,u,p),u\big) \leq \varepsilon(p)$$

when $t \in J$, $|u| \leq U$, $|p| \leq q$ and $\big(\nabla G(u,p),u\big) \geq 0$. Moreover, there exists a C^1 function $g(u)$ and a constant $C \geq 0$ such that

$$\frac{\big(\nabla_u g(u),p\big)}{|p|} - \frac{\big(\nabla G(u,p),u\big)}{|\nabla G(u,p)|} \leq C\,\frac{\varphi(u,p)}{|p|},$$

when $|u| \leq U$, $|p| \leq q$ and $(\nabla G(u,p),u) < 0$. Again q and φ are given in (2.13).

It is worth noting that (V_2) is satisfied if the vectors p, $\nabla G(u,p)$ and $-Q(t,u,p)$ all have the same direction when $p \neq 0$.

Again under the natural hypotheses (H_1)–(H_4), and also assuming (V_1)–(V_2) and (2.13), we have the following result; see [7, Thm. 2], its extension in §7 of [7], and the modified version of this result proved in §3.1 of [4].

THEOREM B. *Suppose that for all $U > 0$ there is a bounded absolutely continuous function k on J such that* (2.9), (2.12), *and*

$$|k'| \leq \begin{cases} \text{Const. } k^{2-m}, & 1 < m < 2, \\ \text{Const.}, & m \geq 2, \end{cases} \quad \text{a.e. in } J, \tag{2.14}$$

are satisfied. Then the rest state is a global attractor for the system (2.1).

Theorem B will be the basis for the second main theorem in §3. Now let

$$H(u,p) = \big(\nabla G(u,p), p\big) - G(u,p)$$

be the Legendre transform in the variable p of the action function $G(u,p)$. The following observation shows that, when

$$H(u,p) \to \infty \quad \text{as } |p| \to \infty \tag{2.15}$$

uniformly for u in compact subsets of $\mathbb{R}^N$, then several of the earlier hypotheses can be weakened, while (2.8) can be omitted entirely.

We first recall that solutions of (2.1) have the property that

$$H(u(t), u'(t)) + F(t, u(t)) \to \text{limit} \quad \text{as } t \to \infty.$$

see [5, (3.7)] or [7, Lemma 5.1(i)]. Hence in turn, since $F(t,u) \geq 0$ by (H_2), the function

$$H(u(t), u'(t))$$

is bounded along any solution $u = u(t)$, $t \in J$. Thus by (2.15), *for any bounded solution of* (2.1) *the function* $u'(t)$ *is also bounded on* J.

It follows that in applying the hypotheses of Theorems A and B for any given bounded solution of (2.1), *one can restrict consideration to compact subsets of vectors* (u,p) *in* $\mathbb{R}^N \times \mathbb{R}^N$. In particular, when (2.15) holds, the condition (2.5) can be weakened to the following form:

For every compact set K in $\mathbb{R}^N \times \mathbb{R}^N$ there exists a measurable control set $I \subset J$ and a number $\gamma \geq 1$ such that

$$|Q(t,u,p)| \cdot |p| \leq \gamma \big|\big(Q(t,u,p), p\big)\big| \quad \text{for all } t \in I \text{ and } (u,p) \in K. \tag{2.5}'$$

Analogous restatements of (2.7) and (V_1) also hold when (2.15) is assumed. Finally, (2.8) is automatically satisfied when $|u \leq U$ and $p_o \leq |p| \leq P$, with the constant depending only on U, P and p_o. Thus (2.8) can be omitted if (2.15) holds.

We conclude the section with a useful estimate.

LEMMA. *Let* (2.5)–(2.7) *hold. Then for all* $U > 0$ *there is a positive constant* $c = c(U)$ *such that*

$$\delta(t) \geq c\,\sigma(t) \quad \textit{for } t \in I. \tag{2.16}$$

If (2.5), (2.6), *and* (2.13) *hold, then for all* $U > 0$ *and* $\vartheta \in (0,1)$ *there is a positive constant* $d = d(U, \vartheta)$ *such that*

$$\frac{1}{|L|} \int_L \delta(t)\,dt \geq d \quad \textit{for all intervals } L \subset J \textit{ with } |L| \geq \vartheta. \tag{2.17}$$

Proof. Fix $U > 0$. By (2.5) and (2.6), with $u = -p$ and $|p| = \min\{1U,\ q\} = r > 0$, we get

$$\delta(t)\,|p|^\mu \geq -\big(Q(t,-p,p),p\big) \geq |Q(t,-p,p)| \cdot |p|/\gamma.$$

On the other hand, by (2.7) and the fact that $|p| \leq \min\{1, U\}$,

$$|Q(t,-p,p)| \geq \sigma(t)\,|p|^\nu,$$

proving (2.16) with $c = r^{\nu+1-\mu}/\gamma$.

Next by (2.13)

$$|Q(t,-p,p)|\cdot|p|\geq\hat{\sigma}(t)\,\varphi(-p,p),$$

so that $\delta(t)\geq\hat{d}\,\hat{\sigma}(t)$ in J, with $\hat{d}=\max\{\varphi(-p,p)\ :\ |p|=r\}\cdot r^{-\mu}/\gamma$. In turn,

$$\frac{1}{|L|}\int_L\delta(t)\,dt\geq\frac{\hat{d}}{|L|}\int_L\hat{\sigma}(t)\,dt\geq\frac{\hat{d}}{2\vartheta}\,a(\vartheta)=d>0$$

by application of inequality (7.2) of [7] with $\lambda=\vartheta$. This completes the proof.

3. Main results. Here we state our main theorems and related consequences. *It is assumed throughout, without further comment, that the conditions* (H_1)–(H_4) *are satisfied.*

Let $(I_n)_n$ be a sequence of nonoverlapping intervals $I_n=[a_n,b_n]$ of J with $a_n<b_n$ and, for the results of this section, $a_n\to\infty$. Let the control set in (H_3)–(H_4) have the form $I=\cup_1^\infty I_n$, and introduce the notation

$$d_n=\frac{1}{|I_n|}\int_{I_n}\delta(t)\,dt\quad\text{(possibly }\infty),$$

which will be used throughout the paper.

Our first result is based on Theorem A of §2. Conditions (2.7) and (2.8) are of course required here, and also for the corresponding corollaries.

THEOREM 1. *Suppose that for every* $U>0$ *there are positive constants* A,B *such that*

$$(3.1)\qquad\sum_1^\infty\sigma_n\cdot\min\left\{|I_n|^q,\ \frac{A}{B+x_n}|I_n|\right\}=\infty,\qquad q=\begin{cases}\dfrac{m+\nu}{m-1}, & \nu>m-2,\\ 2, & \nu\leq m-2,\end{cases}$$

where

$$\sigma_n=\inf_{I_n}\sigma(t)\quad\textit{and}\quad x_n=\sigma_n\,d_n^\ell,\qquad\ell=1/\mu.$$

Then $u=0$ *is a global attractor for* (2.1).

The proof of Theorem 1 is given in §4. In the case $N=1$, $G(p)=p^2/2$, $Q=-a(t)\,p$ and $f(u)=u$, Smith [8] obtained the weaker result that $u=0$ is a global attractor when

$$\sum_1^\infty\sigma_n|I_n|\cdot\min\left\{|I_n|^2,\ \frac{1}{(1+\Delta_n)^2}\right\}=\infty,\qquad\Delta_n=\max_{I_n}a(t);$$

in particular, for this case

$$m=2,\quad\mu=\nu=1,\quad q=3,\quad\sigma(t)=a(t),\quad\delta(t)=Ua(t),$$

so that by taking $A=B=U$ in (3.1) we get

$$\frac{A}{B+x_n}=\frac{1}{1+\sigma_n d_n/U}\geq\frac{1}{1+\Delta_n^2}\geq\frac{1}{(1+\Delta_n)^2}.$$

Several special cases of Theorem 1 are of particular importance. In stating the results, we recall throughout that the hypotheses are understood to apply for each fixed $U>0$.

COROLLARY 1. *Suppose that*

$$\sup_n x_n < \infty. \tag{3.2}$$

Then $u = 0$ *is a global attractor for* (2.1) *if*

$$\sum_1^\infty \sigma_n \min\{|I_n|^q,\ |I_n|\} = \infty. \tag{3.3}$$

Proof. Condition (3.1) with $A = 1 + \sup_n x_n$ and $B = 1$ follows at once from (3.2), (3.3).

In view of (2.16) it is clear that (3.2) holds whenever δ is bounded on $I = \cup_1^\infty I_n$. From Corollary 1 we also get the following consequence:

Suppose that

$$\sup_n x_n < \infty, \qquad \inf_n |I_n| > 0.$$

Then $u = 0$ *is a global attractor if* $\sum_1^\infty \sigma_n = \infty$.

A related result, applying however only for the scalar case of (1.5), appears in [3, Cor. 4.2].

COROLLARY 2. *Suppose that*

$$\inf_n x_n > 0.$$

Then $u = 0$ *is a global attractor for* (2.1) *if*

$$\sum_1^\infty d_n^{-\ell} \min\{|I_n|^q, |I_n|\} = \infty, \quad \textit{where } \ell = 1/\mu. \tag{3.4}$$

Proof. Condition (3.1) with $A = 1 + \inf_n x_n$, $B = 1$ follows easily from (3.4) together with the relations

$$\sigma_n = x_n d_n^{-\ell} \geq x\, d_n^{-\ell}, \qquad \frac{\sigma_n A}{1 + x_n} = \frac{x_n(1+x)}{1+x_n}\, d_n^{-\ell} \geq x\, d_n^{-\ell},$$

where $x = \inf x_n > 0$.

COROLLARY 3. *Suppose that*

$$\inf_I \sigma(t) > 0, \qquad \sup_n d_n < \infty. \tag{3.5}$$

Then $u = 0$ *is a global attractor for* (2.1) *if*

$$\sum_1^\infty |I_n|^q = \infty. \tag{3.6}$$

Proof. This is an immediate consequence of Corollary 1 or 2. For example, (3.5) implies that

$$\inf_n x_n \geq c^\ell \left(\inf_n \sigma_n\right)^{1+\ell} > 0 \quad \text{and} \quad \inf_n d_n^{-\ell} \geq \left(\sup_n d_n\right)^{-\ell} > 0,$$

where c is the constant in (2.16). Hence by Corollary 2 the rest state is a global attractor provided that

$$\sum_1^\infty \min\{|I_n|^q, |I_n|\} = \infty.$$

But this series diverges if and only if (3.6) diverges (since $q > 1$).

Our second main result is based on Theorem B. In this case conditions (2.13) and (V_1)–(V_2) are required (instead of (2.7)–(2.8)). We recall again that the control set has the form $I = \cup_1^\infty I_n$.

THEOREM 2. *Suppose that for every $U > 0$ there exists a positive constant A such that*

$$\sum_1^\infty \min\left\{|I_n|^{\bar q}, \frac{A}{d_n^\ell}|I_n|\right\} = \infty, \qquad \bar q = \begin{cases} \dfrac{m}{m-1}, & 1 < m \le 2, \\ 2, & m \ge 2. \end{cases} \tag{3.7}$$

Then $u = 0$ is a global attractor for (2.1).

The proof of Theorem 2 is given in §5. The hypotheses of Theorem 2 hold for the system (1.5) when A has the decomposition (1.7), see the comments at the end of the section.

Theorem 2 has the following consequences.

COROLLARY 4. *Suppose that*

$$\sup_n d_n < \infty.$$

Then $u = 0$ is a global attractor for (2.1) *if*

$$\sum_1^\infty |I_n|^{\bar q} = \infty. \tag{3.8}$$

Proof. Taking $A = (\sup_n d_n)^\ell$, we see that the series in (3.7) is greater than

$$\sum_1^\infty \min\left\{|I_n|^{\bar q}, |I_n|\right\}.$$

This diverges if and only if $\sum_1^\infty |I_n|^{\bar q}$ diverges (since $\bar q > 1$).

For the canonical case $m = 2$, $\mu = 1$, $\nu = 1$, the exponents q and $\bar q$ in Corollaries 3 and 4 have the respective values 3 and 2, these values being the best possible as shown in §6.

COROLLARY 5. *Suppose there is a positive constant such that*

$$d_n \ge \text{Const.} \begin{cases} |I_n|^{-\mu} & \text{if } m \ge 2, \\ |I_n|^{-\mu/(m-1)} & \text{if } 1 < m \le 2. \end{cases} \tag{3.9}$$

Then $u = 0$ is a global attractor for (2.1) *if*

$$\sum_1^\infty |I_n|\, d_n^{-\ell} = \infty. \tag{3.10}$$

Proof. Let the constant in (3.9) be denoted by D, and choose $A = D^\ell$ in (3.7). Then the second term in braces in (3.7) is less than the first, so that (3.10) implies (3.7).

COROLLARY 6 (Criterion of Thurston–Wong type). *Let* $\inf_n |I_n| > 0$. *Then* $u = 0$ *is a global attractor for* (2.1) *if* (3.10) *is satisfied.*

Proof. Let $U > 0$ be fixed as usual. From (2.17) with $d = d(U, \vartheta) > 0$, where $\vartheta = \min\{\frac{1}{2}, \inf_n |I_n|\}$, we have $d_n \geq d > 0$ for all n, and in turn (3.9) obviously holds because $\inf_n |I_n| > 0$.

Thurston and Wong discovered the special case of Corollary 6 when $N = 1$, $|I_n| = 1$, $G(p) = p^2/2$, $Q(t, u, p) = -a(t, u, p)\, p$, and f is independent of t. Their assumptions imply $\mu = \ell = 1$, in which case (3.10) takes exactly their form $\sum_1^\infty (\int_{I_n} \delta(t)\, dt)^{-1} = \infty$.

Artstein and Infante [1, cond. (2.7)] showed for the same case that $u = 0$ is a global attractor provided

$$\frac{1}{K^2} \sum_1^K d_n \leq B$$

for some constant B independent of K. In fact, more generally, without any restrictions on the measures of I_n, and whatever the value of μ, condition (3.10) is implied by

$$\frac{1}{K^{\mu+1}} \sum_1^K c_n \leq B, \quad \text{where } c_n = \frac{1}{|I_n|^{\mu+1}} \int_{I_n} \delta(t)\, dt = \frac{d_n}{|I_n|^\mu}. \tag{3.11}$$

To see this, note that for any positive integers $0 < L < K$ we have, by Hölder's inequality,

$$K - L = \sum_L^K 1 \leq \left(\sum_L^K c_n \right)^{1/(\mu+1)} \left(\sum_L^K c_n^{-1/\mu} \right)^{\mu/(\mu+1)},$$

so that in turn

$$\sum_L^K c_n^{-\ell} \geq \left[\frac{1}{(K-L)^{\mu+1}} \sum_L^K c_n \right]^{-1/\mu}. \tag{3.12}$$

But, by (3.11), if $K = 2L$ then

$$\frac{1}{(K-L)^{\mu+1}} \sum_L^K c_n \leq 2^{\mu+1} B.$$

Hence from (3.12) it follows that

$$\sum_1^{2^\nu} c_n^{-\ell} \geq \frac{\nu}{2(2B)^\ell}, \qquad \nu = 1, 2, \ldots.$$

Thus the series (3.10) diverges. We have proved the following result.

COROLLARY 7 (Criterion of Artstein–Infante type). *Suppose that* $\inf_n |I_n| > 0$, *or more generally that* (3.9) *holds. Then* $u = 0$ *is a global attractor for* (2.1) *if* (3.11) *is satisfied.*

In essentially the same way, condition (3.4) in Corollary 2 can also be deduced from the Artstein–Infante type condition (3.11), provided $\inf_n |I_n| > 0$.

Remark. When $\inf_n |I_n| > 0$ the criteria (3.4) of Corollary 2 and (3.10) of Corollary 6 are equivalent. Since by (2.16) the condition $\inf_n x_n > 0$ holds whenever $\inf_I \sigma(t) > 0$, one can see a connection between the hypotheses of these corollaries. On the other hand, the assumptions of Theorem 2 are different enough from those of Theorem 1 that the corollaries are not directly comparable.

The system (1.5). We show that Corollaries 3 and 4 apply to the system (1.5). For Corollary 3 the hypotheses of Theorem A must be verified, on the basis of the assumptions immediately following (1.5).

Fix a compact subset K of $\mathbb{R}^N \times \mathbb{R}^N$, and let $\alpha > 0$ be such that

$$\big(A(t,u,p)p,p\big) \geq \alpha\,|p|^2 \quad \text{for } (t,u,p) \in I \times K;$$

also denote by $\|A\|$, the L^∞ norm of A on $I \times K$. Then one easily sees that (2.5) holds in $I \times K$ with $\gamma = \alpha/\|A\|$, that (2.6) is satisfied with $\delta(t) = U\,\|A\| = \text{Const.}$ and $\mu = 1$, and that (2.7) is verified with $\sigma(t) = \alpha$, $\nu = 1$. Finally, taking into account the observations just before the lemma of §2, together with the fact that (2.15) holds since $H(u,p) = \frac{1}{2}|p|^2$, we see that Theorem A is applicable to (1.5), with $m = 2$.

In turn, since $\nu = 1$, we get $q = (m+\nu)/(m-1) = 3$ in (3.1). Moreover, $\inf_I \sigma(t) = \alpha > 0$ and $\sup_n d_n = U\,\|A\| < \infty$, so (3.5) is satisfied. Corollary 3 then gives the criterion (1.6).

For Corollary 4 the hypotheses of Theorem B must be verified on the basis of the assumptions (1.7)–(1.9). Again fix a compact set K of $\mathbb{R}^N \times \mathbb{R}^N$. Then one easily sees that

$$\big(A(t,u,p)p,p\big) = \beta(t,u,p)\,|p|^2 + \big(B(t,u,p)p,p\big) \geq \beta_1\,|p|^2 \quad \text{on } J \times K,$$

by (1.8) and the fact that B is nonnegative definite. Hence we can take $\hat{\sigma}(t) = 1$ and $\varphi(u,p) = \beta_1\,|p|^2$ in (2.13). Also $\|A\| \leq \beta_2 + \|B\| < \infty$ by (1.9). Thus (2.5) holds in $I \times K$ with $\gamma = \beta_1/\|A\|$. As before we can take $\delta(t) = U\,\|A\|$ and $\mu = 1$.

Next (V_1) is satisfied with $h(t) = \beta_1 p_0^2$. Finally, for (V_2),

$$\big(Q(t,u,p),u\big) = -\beta(t,u,p)\,(p,u) - \big(B(t,u,p)p,u\big) \leq U\,\|B\|\,|p|,$$

when $|u| \leq U$ and $(p,u) \geq 0$. This gives the first part of (V_2) with $\varepsilon(p) = U\,\|B\|\,|p|$. The second part of (V_2) is automatic for (1.5) with $g(u) = \frac{1}{2}u^2$ and $C = 0$. Again taking into account the observations just before the lemma of §2, we see that Theorem B is applicable.

As before $\sup_n d_n = U\,\|A\| < \infty$. Corollary 4 then gives the criterion (1.10), since $m = 2$ and $\bar{q} = 2$ by (3.7).

4. Proof of Theorem 1. Recall that $I = \cup_1^\infty I_n$ where $I_n = [a_n,\ b_n]$. We begin with a simple lemma.

LEMMA. *Let k be a nonnegative measurable function such that $k(t) = 0$ for $t \in J \setminus I$ and for which*

$$d_n\,k_n^\mu \leq M_1 \tag{4.1}$$

and

$$\int_{I_n} k(t)dt \geq \frac{1}{M_2}\,|I_n|\,k_n, \tag{4.2}$$

where $k_n = \sup_{I_n} k(t)$ and M_1, M_2 are positive constants.

Then k satisfies condition (2.12) *with* $M = M_1 M_2$.

Proof. We have

$$\begin{aligned}\int_{I_n} \delta(t) k^{\mu+1}(t)dt &\le k_n^{\mu+1} \int_{I_n} \delta(t)dt = k_n |I_n|\, k_n^{\mu} d_n \\ &\le k_n |I_n| M_1 \quad \text{by (4.1)} \\ &\le M_1 M_2 \int_{I_n} k(t)dt \quad \text{by (4.2)}.\end{aligned}$$

Condition (2.12) now follows by summation over n and the fact that $R = 0$ on J/I.

Proof of Theorem 1. Recall that $x_n = \sigma_n d_n^{\ell}$, $\ell = 1/\mu$, $q = (\nu + m)/(m-1)$ if $\nu > m-2$ and $q = 2$ if $\nu \le m-2$.

We now construct a bounded piecewise smooth function $k = k(t)$ satisfying the assumptions $(2.9)_2$ and (2.10) of Theorem A. (Of course the functions σ, δ in Theorem A depend on U, so also k depends on U. As in Section 2 we do not specifically indicate this dependence.) In particular, let $k = 0$ on $J \setminus I$. To obtain k on the intervals I_n, we separately consider the two subcases

$$\text{(i)} \qquad |I_n|^{q-1} \le \frac{A}{B + x_n}$$

and

$$\text{(ii)} \qquad |I_n|^{q-1} > \frac{A}{B + x_n}.$$

Subcase (i). Let $I_n = [a_n, b_n]$ and put

$$k(t) = \begin{cases} C\,\sigma_n (t - a_n)^{q-1}, & a_n \le t \le \tfrac{1}{2}(a_n + b_n), \\ C\,\sigma_n (b_n - t)^{q-1}, & \tfrac{1}{2}(a_n + b_n) \le t \le b_n, \end{cases}$$

where $C = 2^{q-1} B/A$. Then $(2.10)_2$ holds on I_n with the Const. $= 2\,(q-1)\,(B/A)^{\lambda}$, independent of n; note that the exponent λ is defined in (2.11), and that $\lambda = 1/(q-1)$ by virtue of $(3.1)_2$. Next, letting $k_n = \max_{I_n} k(t)$ as in the lemma, we have

$$k_n = k\left(\frac{a_n + b_n}{2}\right) = C\,\sigma_n \left(\frac{|I_n|}{2}\right)^{q-1} \le 2^{1-q} C\,\sigma_n \frac{A}{B + x_n} = \frac{B\sigma_n}{B + x_n} \tag{4.3}$$

by (i) and the choice of C. In turn, since $x_n = \sigma_n d_n^{\ell} \ge c^{\ell}\, \sigma_n^{\ell+1}$ by (2.16), there follows

$$k_n \le \frac{B\sigma_n}{B + c^{\ell} \sigma_n{}^{\ell+1}} \le D, \tag{4.4}$$

where D is a constant depending only on B, c, and ℓ. Hence k is uniformly bounded on intervals I_n of type (i), with the bound independent of n. Moreover, again from (4.3), we have $k(t) \le \sigma_n \le \sigma(t)$, so $(2.10)_1$ holds on these intervals with Const. $= 1$.

Subcase (ii). Put

$$k(t) = \begin{cases} C\,\sigma_n (t - a_n)^{q-1}, & a_n \le t \le t_n, \\ \dfrac{B\sigma_n}{B + x_n}, & t_n < t < \bar{t}_n, \\ C\,\sigma_n (b_n - t)^{q-1}, & \bar{t}_n \le t \le b_n, \end{cases}$$

where t_n and $\bar{t}_n$ are chosen so that k is continuous on I_n. This can be done because of condition (ii).

As before k satisfies $(2.10)_2$ on I_n with the same Const. $= 2(q-1)(B/A)^{\lambda}$, since $k' = 0$ on $(t_n, \bar{t}_n)$. Moreover, we have

$$k_n = \frac{B\sigma_n}{B + x_n} \leq D, \tag{4.5}$$

as in (4.4). Therefore $(2.10)_1$ is satisfied and k is uniformly bounded on intervals I_n of type (ii).

We next show that k satisfies conditions (4.1) and (4.2) of the lemma. Indeed by (4.3) and (4.5), for each n,

$$d_n k_n^{\mu} \leq d_n \left(\frac{B\sigma_n}{B + x_n}\right)^{\mu} = B^{\mu} \left(\frac{x_n}{B + x_n}\right)^{\mu} \leq B^{\mu} = M_1.$$

Thus (4.1) is verified.

In case (i) an easy calculation and the use of (4.3) gives

$$\int_{I_n} k(t)\, dt = \frac{2}{q} C\sigma_n \left(\frac{|I_n|}{2}\right)^q = \frac{1}{q} |I_n| k_n,$$

while in case (ii)

$$\begin{aligned} \int_{I_n} k(t)\, dt &= \frac{(t_n - a_n)}{q} k(t_n) + (\bar{t}_n - t_n)\frac{B\sigma_n}{B + x_n} + \frac{(b_n - \bar{t}_n)}{q} k(\bar{t}_n) \\ &= \left\{\frac{1}{q}(t_n - a_n) + (\bar{t}_n - t_n) + \frac{1}{q}(b_n - \bar{t}_n)\right\} \frac{B\sigma_n}{B + x_n} \\ &> \frac{1}{q} |I_n| k_n, \end{aligned} \tag{4.6}$$

since $q > 1$. Hence (4.2) holds with $M_2 = q$.

The lemma now shows that condition (2.12) of Theorem A is satisfied. It remains only to verify the hypothesis $(2.9)_1$ to finish the proof. We already know that in case (i)

$$\int_{I_n} k(t)\, dt = \frac{2}{q} C\sigma_n \left(\frac{|I_n|}{2}\right)^q = \frac{1}{q}\frac{B}{A}\sigma_n |I_n|^q$$

by the choice of C, while in case (ii) by (4.5) and (4.6),

$$\int_{I_n} k(t)\, dt > \frac{1}{q}\sigma_n |I_n| \frac{B}{B + x_n}.$$

Consequently for each n,

$$\int_{I_n} k(t)\, dt \geq \frac{1}{q}\frac{B}{A}\sigma_n \min\left\{|I_n|^q,\ \frac{A}{B + x_n}|I_n|\right\}.$$

Since

$$\int_J k(t)\, dt = \sum_1^{\infty} \int_{I_n} k(t)\, dt,$$

condition (3.1) now shows that $k \notin L^1(J)$, as required in $(2.9)_1$.

This completes the proof.

5. Proof of Theorem 2. As in the proof of Theorem 1, we shall construct a bounded piecewise smooth function $k = k(t)$ satisfying the assumptions (2.9) and (2.14) of Theorem B. (As before, k depends on U in what follows.) In particular, we define k to be 0 on $J \setminus \cup_1^\infty I_n$ and, to obtain k on the intervals I_n, consider separately the two cases:

$$\text{(i)} \qquad |I_n|^{\bar{q}-1} \le \frac{A}{d_n^\ell}$$

and

$$\text{(ii)} \qquad |I_n|^{\bar{q}-1} > \frac{A}{d_n^\ell}.$$

Case (i). Put

$$k(t) = \begin{cases} C(t-a_n)^{\bar{q}-1}, & a_n \le t \le \frac{1}{2}(a_n+b_n), \\ C(b_n-t)^{\bar{q}-1}, & \frac{1}{2}(a_n+b_n) \le t \le b_n, \end{cases}$$

where $C = 2^{\bar{q}-1}/A$. Then recalling the definition of $\bar{q}$ in (3.7), we see that (2.14) holds on I_n with Const. $= (\bar{q}-1)C^{m-1}$ when $1 < m < 2$ and Const. $= C$ when $m \ge 2$.

Next

$$k_n = k\left(\tfrac{1}{2}(a_n+b_n)\right) = C\left(\tfrac{1}{2}|I_n|\right)^{\bar{q}-1}. \tag{5.1}$$

By (i) and the choice of C we obtain

$$k_n \le d_n^{-\ell} \quad \text{for each } n. \tag{5.2}$$

Now applying (2.17) in the lemma of §2 with $d = d(U, 1/2)$, we have $d_n \ge d$ whenever $|I_n| \ge 1/2$, so that $k_n \le d^{-\ell}$ for such I_n's. On the other hand, when $|I_n| \le 1/2$ we derive from (5.1) that $k_n \le 2^{1-\bar{q}}/A$. Hence k is bounded on each I_n of type (i), uniformly in n. Thus k is bounded on J, with

$$k(t) \le \max\{d^{-\ell},\ 2^{1-\bar{q}}/A\}, \qquad t \in J.$$

Case (ii). Put

$$k(t) = \begin{cases} C(t-a_n)^{\bar{q}-1}, & a_n \le t \le t_n, \\ d_n^{-\ell}, & t_n < t < \bar{t}_n, \\ C(b_n-t)^{\bar{q}-1}, & \bar{t}_n \le t \le b_n, \end{cases}$$

where t_n and $\bar{t}_n$ are chosen so that k is continuous on I_n. This can be done in virtue of (ii). As in case (i), we see that (2.14) is satisfied and that k is bounded on I_n uniformly in n, namely

$$k_n = d_n^{-\ell} \le \max\{d^{-\ell},\ 2^{1-\bar{q}}/A\},$$

since by (2.17) we have $d_n \ge d = d(U, 1/2)$ when $|I_n| \ge 1/2$, while by (ii) on the other hand, $d_n^{-\ell} \le 2^{1-\bar{q}}/A$ when $|I_n| \le 1/2$.

We next show that k satisfies conditions (4.1) and (4.2) of the lemma in §4. Indeed for each n,

$$d_n k_n^{\mu} \le d_n (d_n^{-\ell})^{\mu} = 1 = M_1,$$

by (5.2) in case (i) and by the definition of k in case (ii). Thus (4.1) is verified.

In case (i) by (5.1)

$$\int_{I_n} k(t)\, dt = 2\frac{C}{\bar{q}} \left(\frac{1}{2}|I_n|\right)^{\bar{q}} = \frac{1}{\bar{q}}\, |I_n|\, k_n,$$

while in case (ii), as in the proof of Theorem 1,

$$\int_{I_n} k(t)\, dt > \frac{1}{\bar{q}}\, |I_n|\, k_n.$$

Hence (4.2) holds with $M_2 = \bar{q}$.

The lemma now shows that condition (2.12) is satisfied, so that to apply Theorem B it remains only to verify $(2.9)_1$. Arguing as in the proof of Theorem 1, we obtain

$$\int_{I_n} k(t)\, dt \ge \frac{1}{\bar{q}A}\, \min\left\{|I_n|^{\bar{q}},\ \frac{A}{d_n^{\ell}}|I_n|\right\}.$$

Consequently (3.7) implies that $k \notin L^1(J)$, and this completes the proof.

6. Examples. The purpose of this section is to show that the exponents which appear in conditions (1.6) and (1.10) are best possible.

Consider the linear equation

$$u'' + a(t)u' + u = 0, \qquad t \in J = [0, \infty), \tag{6.1}$$

where $a\colon J \to \mathbb{R}$ is an *on–off* damping function of the form

$$a(t) = \begin{cases} 0 & \text{in } J \setminus \bigcup_1^{\infty} I_n, \\ 2 & \text{in } \bigcup_1^{\infty} I_n. \end{cases} \tag{6.2}$$

The following result shows that the exponent 3 in (1.6) is best possible.

PROPOSITION 1. *Let $\epsilon \in (0, 2]$ be fixed and let $(\lambda_n)_n$ be a sequence of positive real numbers such that*

$$\sum_1^{\infty} \lambda_n^{3-\epsilon} = \infty, \qquad \sum_1^{\infty} \lambda_n^3 < \infty. \tag{6.3}$$

Then there exists a sequence of disjoint intervals $I_n = [a_n,\ a_n + \lambda_n]$, with $a_n \to \infty$, such that $u = 0$ is not a global attractor for (6.1) *with the damping* (6.2).

Remarks. Since the damping function (6.2) is not continuous the corresponding solutions of (6.1) must be sought in the class $C^1(J)$. Of course, a smoothing procedure will obviously yield a corresponding result for (6.1) with continuous damping.

From the proof it will be clear that the sequence $(I_n)_n$ can be chosen to have arbitrarily large gaps, i.e., with $a_{n+1} - a_n$ unbounded.

Proof. We place the interval I_1 arbitrarily in $J = [0, \infty)$, and recursively determine the location of the successive intervals I_n for $n \geq 2$. In particular we shall impose the Cauchy conditions

$$u(a_n) = A_n, \qquad u'(a_n) = 0, \qquad A_1 \neq 0 \tag{6.4}$$

in order to construct a bounded solution u of (6.1)–(6.2) that does not approach 0 as $t \to \infty$. The values A_n will be recursively determined, along with the location of the intervals I_n.

From (6.4) and (6.2) it is clear that

$$u(t) = A_n(1 + t - a_n)e^{a_n - t} \quad \text{for } t \in I_n. \tag{6.5}$$

On the other hand, in the intervals $(a_n + \lambda_n,\ a_{n+1})$ between the sets I_n and I_{n+1}, we have

$$u(t) = \varphi_n(t) = B_n \cos(t + \theta_n) \tag{6.6}$$

for some constants B_n and θ_n, again by (6.2). Clearly (6.6) should join smoothly with (6.5) at the point $a_n + \lambda_n$ and also satisfy the conditions $\varphi_n(a_{n+1}) = A_{n+1}$ and $\varphi_n'(a_{n+1}) = 0$. These latter conditions take the specific form

$$B_n \cos(a_{n+1} + \theta_n) = A_{n+1}, \qquad B_n \sin(a_{n+1} + \theta_n) = 0,$$

so that $B_n^2 = A_{n+1}^2$. The former conditions are

$$B_n \cos(a_n + \lambda_n + \theta_n) = A_n(1 + \lambda_n)e^{-\lambda_n},$$
$$B_n \sin(a_n + \lambda_n + \theta_n) = A_n \lambda_n e^{-\lambda_n}.$$

Squaring and adding gives

$$A_{n+1}^2 = A_n^2\{(1 + \lambda_n)^2 + \lambda_n^2\}e^{-2\lambda_n} \equiv A_n^2 \Phi(\lambda_n). \tag{6.7}$$

This determines A_{n+1}^2 in terms of A_n^2 and λ_n. Because θ_n is not yet chosen, it is clear that B_n and A_{n+1} are so far determined only up to their signs. We can choose the sign of B_n as we wish, say sign $B_n =$ sign A_{n+1}. Then $\cos(a_{n+1} + \theta_n) = 1$, so without loss of generality $\theta_n = -a_{n+1}$. In turn

$$\tan(a_{n+1} - a_n - \lambda_n) = -\frac{\lambda_n}{1 + \lambda_n},$$

which determines a_{n+1} modulo π; indeed, if sign $A_{n+1} =$ sign A_n, then one sees that

$$a_{n+1} - a_n - \lambda_n \in \left(\tfrac{7}{4}\pi,\ 2\pi\right) \qquad \text{modulo } 2\pi,$$

while if sign $A_{n+1} = -$sign A_n, then

$$a_{n+1} - a_n - \lambda_n \in \left(\tfrac{3}{4}\pi,\ \pi\right) \qquad \text{modulo } 2\pi.$$

Clearly $a_{n+1} - a_n$ can be arbitrarily large, though not arbitrarily small; in fact since $\lambda_n \to 0$, there is a sequence $(k_n)_n$ of positive integers such that $\lim_n (a_{n+1} - a_n - k_n\pi) = 0$.

An easy calculation shows that

$$1 - \tfrac{4}{3}x^3 < \Phi(x) < 1 \quad \text{for } x > 0.$$

Hence $|A_{n+1}| < |A_n|$ by (6.7), so that

$$\limsup_{t\to\infty} |u(t)|^2 = \lim_n A_n^2 = A_1^2 \prod_1^\infty \Phi(\lambda_n).$$

We can assume without loss of generality that $\lambda_n^3 < 3/4$ for all n, in virtue of $(6.3)_2$. Therefore,

$$\limsup_{t\to\infty} |u(t)|^2 > A_1^2 \prod_1^\infty (1 - \tfrac{4}{3}\lambda_n^3) > 0,$$

where the last inequality is equivalent to $\sum_1^\infty \lambda_n^3 < \infty$. This completes the demonstration.

The proof sharply brings out the role of the exponent 3 in condition (1.6). Figure 1 shows a typical graph of u with all the $A_n > 0$ and with varying spacing between the I_n.

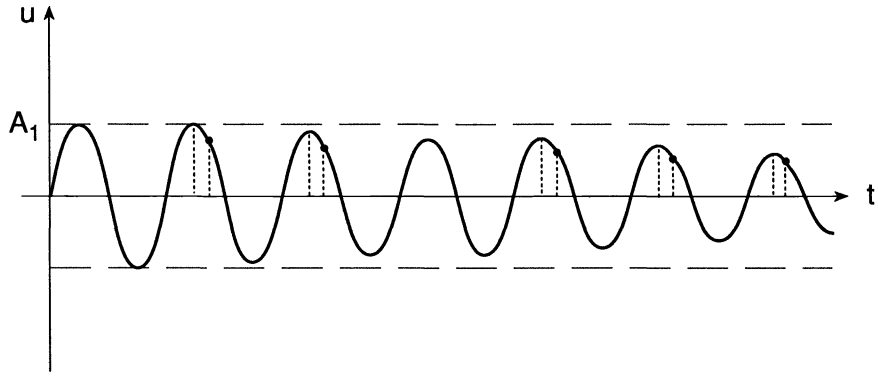


FIG. 1. *The solution u of Proposition* 1.

In Fig. 2 the heavy curve is the solution (6.5). The light curve is one arch of the cosine wave φ_n defined by (6.6), whose amplitude is A_{n+1}. The dashed curve is one arch of the cosine wave φ_{n-1}, whose amplitude is A_n.

The next result shows that (1.6) is *not* necessary for the global stability of the rest state of (6.1) when

$$\inf_n \sigma_n > 0 \quad \text{and} \quad \sup_n d_n < \infty.$$

It also indicates the extreme delicacy of the situation when one has *on–off* damping, that is, the exact switching times can be of great importance.

PROPOSITION 2. *Under the hypotheses of Proposition* 1 *there also exists a sequence of disjoint intervals* $I_n = [a_n, a_n + \lambda_n]$, *with* $a_{n+1} - a_n > \pi$, *such that* $u = 0$ *is a global attractor for* (6.1) *with the damping* (6.2).

Proof. We place I_1 arbitrarily. To construct the remaining intervals, choose some sequence $(k_n)_n$ of positive integers and define

$$a_{n+1} = a_n + \lambda_n + k_n\pi.$$

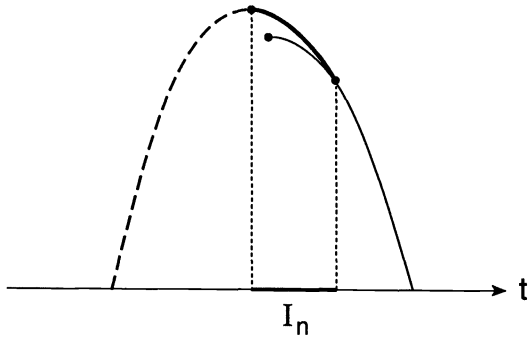


FIG. 2. *Behavior of u near an interval I_n.*

Now let u be any solution of (6.1)–(6.2) on J. Since (6.6) holds on the intervals $(a_n + \lambda_n, a_{n+1})$, whose lengths are multiples of π, it is evident that

$$\begin{aligned} u(a_{n+1}) &= (-1)^{k_n}\, u(a_n + \lambda_n), \\ u'(a_{n+1}) &= (-1)^{k_n}\, u'(a_n + \lambda_n). \end{aligned} \tag{6.8}$$

Now, for $\tau \in [0, \infty)$, we put

$$v(\tau) = (-1)^{j_n}\, u(a_n + \tau - \ell_n) \quad \text{if } \tau \in [\ell_n, \ell_{n+1}), \quad n = 1, 2, \ldots,$$

where $j_n = \sum_1^{n-1} k_i$ and $\ell_n = \sum_1^{n-1} \lambda_i$ when $n \geq 2$, and $j_1 = \ell_1 = 0$. By $(6.3)_1$ it is clear that $\ell_n \to \infty$ as $n \to \infty$, so v is well defined. Also v is of class C^1 in view of (6.8).

Directly from the definition of v we see that v is a solution of

$$v'' + 2v' + v = 0 \quad \text{in } [0, \infty), \tag{6.9}$$

so that v is certainly of class C^2. The equation (6.9) has characteristic values $r_1 = r_2 = -1$; hence $v(\tau) \to 0$ and $v'(\tau) \to 0$ as $\tau \to \infty$.

From this we get

$$\sup_{I_n} |u(t)| \to 0, \quad \sup_{I_n} |u'(t)| \to 0 \quad \text{as } n \to \infty;$$

it then follows at once that the coefficients B_n in (6.6) also approach 0 as $n \to \infty$. This completes the proof.

Remark. Similar conclusions can be obtained for the equation

$$u'' + a(t)\, |u'|^{\nu - 1} u' + u = 0, \qquad \nu > 0,$$

but we shall not pursue this here.

To investigate the exponent 2 in condition (1.10), we consider the equation (6.1) with the singular *on–off* damping

$$a(t) = \begin{cases} \infty & \text{in } J \setminus \bigcup_1^\infty I_n, \\ 2 & \text{in } \bigcup_1^\infty I_n. \end{cases} \tag{6.10}$$

Of course a is neither continuous nor has its values in the reals, so that equation (6.1) on $J \setminus \cup_1^\infty I_n$ must therefore be interpreted in terms of a family of equations in which the damping uniformly approaches ∞ on any compact (or even any bounded) subset of $J \setminus \cup_1^\infty I_n$. The corresponding solutions approach constants on the open intervals between the I_n's. In fact, the appropriate interpretation of a solution of the initial value problem $u(a_n + \lambda_n) = \alpha$, $u'(a_n + \lambda_n) = \beta$ on the interval $J_n = (a_n + \lambda_n,\ a_{n+1}]$ is $u(t) \equiv \alpha$. That is, if u_M is the solution of the initial value problem

$$\begin{cases} u'' + Mu' + u = 0, \\ u_M(a_n + \lambda_n) = \alpha, \qquad u'_M(a_n + \lambda_n) = \beta \end{cases}$$

on the interval $\overline{J_n}$, then $u_M(t) \to \alpha$ as $M \to \infty$ uniformly on $\overline{J_n}$, while $u'_M(t) \to 0$ uniformly on any compact subset of J_n. Accordingly, solutions of (6.1), (6.10) are to be interpreted as functions of class

$$C(J) \cap C^1(J \setminus \{a_n + \lambda_n,\ n = 1, 2, \ldots\})$$

which satisfy (6.1) with $a(t) = 2$ on each interval $[a_n,\ a_n + \lambda_n)$ and which are constant on each interval $\overline{J_n}$.

PROPOSITION 3. *Let $(I_n)_n$ be a sequence of disjoint intervals $I_n = [a_n,\ a_n + \lambda_n]$. Then $u = 0$ is a global attractor for* (6.1) *with the damping* (6.10) *if and only if*

$$\sum_1^\infty |I_n|^2 = \infty.$$

Proof. If $\sum_1^\infty |I_n|^2 = \infty$, then $a_n \to \infty$ and by Corollary 4 the rest state $u = 0$ is a global attractor.

Now assume that $\sum_1^\infty |I_n|^2 < \infty$ and, without loss of generality, that $\lambda_n^2 < 2$. For simplicity we first consider the case

$$a_n \nearrow \infty \quad \text{as } n \to \infty. \tag{6.11}$$

Then every solution of (6.1), (6.10) on J has the form

$$u(t) = \begin{cases} A_1 & \text{if } t \in [0, a_1], \\ A_n(1 + t - a_n)e^{a_n - t} & \text{if } t \in I_n, \\ A_{n+1} & \text{if } t \in J_n. \end{cases}$$

Furthermore, by continuity at the point $a_n + \lambda_n$ we have the recursive formula

$$A_{n+1} = A_n(1 + \lambda_n)e^{-\lambda_n}.$$

Obviously,

$$1 - \tfrac{1}{2}x^2 < (1 + x)e^{-x} < 1 \quad \text{for } x > 0.$$

It follows that $(|A_n|)_n$ is decreasing and so also $|u(t)|$ is decreasing on $[0, \infty)$. Thus

$$\lim_{t \to \infty} |u(t)| = |A_1| \cdot \prod_1^\infty (1 + \lambda_n)e^{-\lambda_n}.$$

Then if $A_1 \neq 0$

$$\lim_{t\to\infty} |u(t)| > |A_1| \cdot \prod_1^\infty (1 - \tfrac{1}{2}\lambda_n^2) > 0.$$

Consequently every solution, except the trivial one ($A_1 = 0$), approaches a nonzero limit at ∞. In particular, $u = 0$ is not a global attractor for (6.1), (6.10).

If (6.11) fails, then $a_n \nearrow$ finite a. The previous proof shows that $u(t) \to u_0 \neq 0$ as $t \nearrow a$, so that in turn

$$u(t) \equiv u_0 \quad \text{for } t \geq a;$$

the solution of course need not be smooth at the point $t = a$. This completes the proof.

The condition $\delta \notin L^1(J)$ is known to be a *necessary condition* for the rest state $u = 0$ to be a global attractor for the system (1.5); see [5, Cor. 1, §5]. (The notation used in §5 of [5] gives $\hat{\delta} = 2\delta/U$ and $\hat{\delta} \notin L^1(J)$: the multiplicative constant $2/U$ clearly does not affect the conclusion.) This result holds in particular for the linear equation

$$u'' + A(t)u' + u = 0, \tag{6.12}$$

where of course the necessary condition becomes $A \notin L^1(J)$.

On the other hand, *the condition $\sigma \notin L^1(J)$ alone is not sufficient for the rest state to be a global attractor*, even for equation (6.12), where $\sigma = \delta/U$, and even in the representative case in which A is bounded.

Proof. We must show that there exist equations of the form (6.12) with $A \in L^\infty(J)\backslash L^1(J)$ such that zero is not a global attractor. Indeed, consider equation (6.12) with $A = a$, and a given by (6.2). Moreover, suppose the intervals I_n in (6.2) satisfy (6.3). By $(6.3)_2$, eventually $|I_n| \leq 1$, say for all $n \geq K$, so that

$$\sum_K^\infty |I_n| \geq \sum_K^\infty |I_n|^{3-\varepsilon} = \infty$$

by $(6.3)_1$. Hence by (6.2)

$$\int_J A(t)dt = 2\sum_1^\infty |I_n| = \infty,$$

so $A \in L^\infty(J)\backslash L^1(J)$. On the other hand, by Propositon 1, if the intervals I_n are located properly, then $u = 0$ is *not* a global attractor.

A related result was shown in [6, Thm. 4.4], namely, that the condition $1/\sigma \notin L^1(J)$ is a *necessary condition* for the rest state to be a global attractor for (1.5). (In the notation of Theorem 4.4 of [6] the function here called σ is there called δ; moreover, in the application of Theorem 4.4 to (1.5) we have $q = k = 2$, $C = 0$, $\psi = 0$ and of course $H = |p|^2/2$ and $Q = -A(t,u,p)p$.)

On the other hand, *the condition $1/\delta \notin L^1(J)$ alone is not sufficient for the rest state to be a global attractor*, even for (6.12) and even when $1/A \in L^\infty(J)\backslash L^1(J)$.

Proof. We must show that there exist equations of the form (6.12) with $1/A \in L^\infty(J)\backslash L^1(J)$ such that zero is not a global attractor. Indeed consider (6.12) with

$A = a$, and a given by (6.10). Moreover, suppose the intervals I_n in (6.10) satisfy

$$\sum_1^\infty |I_n| = \infty, \qquad \sum_1^\infty |I_n|^2 < \infty.$$

Then

$$\int_J \frac{dt}{A(t)} = \frac{1}{2}\sum_1^\infty |I_n| = \infty,$$

so $1/A \in L^\infty(J)\backslash L^1(J)$. On the other hand, by Proposition 3 the rest state is *not* a global attractor.

Acknowledgment. The first author is a member of *Gruppo Nazionale di Analisi Funzionale e sue Applicazioni* of the *Consiglio Nazionale delle Ricerche.*

REFERENCES

[1] Z. ARTSTEIN AND E. F. INFANTE, *On the asymptotic stability of oscillators with unbounded damping*, Quart. Appl. Math., 35 (1976), pp. 195–199.

[2] L. HATVANI, *Nonlinear oscillation with large damping*, Dynamical Systems Appl., 1994, to appear.

[3] L. HATVANI AND V. TOTIK, *Asymptotic stability of the equilibrium of the damped oscillator*, Differential Integral Equations, 6 (1993), pp. 835–848.

[4] G. LEONI, *A note on a theorem of Pucci and Serrin*, J. Differential Equations, 1994 to appear.

[5] P. PUCCI AND J. SERRIN, *Precise damping conditions for global asymptotic stability for nonlinear second order systems*, Acta Math., 170 (1993), pp. 275–307.

[6] ———, *Continuation and limit behavior for damped quasi-variational systems*, in Degenerate Diffusions, IMA Vols. in Math. Appl. 47, W.-M. Ni, L. A. Peletier and J. L. Vazquez, eds., Springer-Verlag, New York, 1993, pp. 157–173.

[7] ———, *Precise damping conditions for global asymptotic stability for nonlinear second order systems*, II, J. Differential Equations, 1994, to appear.

[8] R. A. SMITH, *Asymptotic stability of* $x'' + a(t)x' + x = 0$, Quart. J. Math. Oxford, 12 (1961), pp. 123–126.

[9] L. H. THURSTON AND J. W. WONG, *On global asymptotic stability of certain second order differential equations with integrable forcing terms*, SIAM J. Appl. Math., 24 (1973), pp. 50–61.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 836–858, May 1994

002

ASYMPTOTIC ANALYSIS OF A MULTIDIMENSIONAL VIBRATING STRUCTURE*

CARLOS CONCA† AND ENRIKE ZUAZUA‡

Abstract. The aim of this paper is to describe the qualitative behavior of the eigenfrequencies and eigenmotions of a model problem that represents the vibrations of an elastic multidimensional body (or *multistructure*). The model studied here assumes that the multidimensional structure consists of two bodies: one of them is a bounded domain of $\mathbb{R}^N$ (N= 2 or 3 in practice), and the other a one-dimensional straight string (that is represented by a real interval). The bodies are elastically attached at a small neighborhood of a point of contact A on the boundary of the N-dimensional domain by one extreme of the string. When this structure undergoes impulses, both its parts vibrate. The result is the classical spectral problem for the Laplace operator in both regions of the *multistructure*, coupled with a special boundary condition, which models the junction between both bodies. It is a nonstandard eigenvalue system since the spectral problems corresponding to each part are linked through this junction condition. For a variety of reasons, there is interest in cases in which the junction region is very small. Thus one of the aims in this article is to study the asymptotic behavior of the spectrum of this eigenvalue problem when the junction region tends to disappear, and converges towards a set of Lebesgue measure zero containing the contact point. This is done in terms of the convergence of the Green's operator and the spectral family associated with this problem.

Key words. multistructures, eigenvalue problems, asymptotic analysis

AMS subject classifications. 35P10, 73K99, 35J25

Introduction. *Multistructures* are a very common occurrence in practice. They are found, for example, in the study of aerials, plates, and shells with stiffeners such as in solar panels, etc. However, despite their enormous practical importance, it is only recently that research in this field has begun from a rigorous point of view. One of the first theoretical studies of mathematical models in *multistructures* is the article by Ciarlet, Le Dret and Nzengwa (1989). Roughly speaking, the method that these authors propose for dealing with a junction of two bodies, say, one of dimension N and another of dimension $(N-1)$, consists of carrying out an asymptotic analysis of the structure. In the first place, the body of dimension $(N-1)$ is assimilated to another of dimension N, one of whose dimensions is (very) small compared to the others. Let us say that its size is ε while the $(N-1)$-dimensional volume of the original body is 1. Then a change is introduced to the scale of this body, so that the small dimension becomes of the same characteristic size as the rest, and one concludes by letting ε go to zero. In the general case, the idea is always the same, that is, of rescaling the different parts of the *multistructure* independently of each other and passing to the limit in all those dimensions which are small with respect to the others. It will be appreciated that in this approach the contact condition is dealt with implicitly, since the *multistructure* is approximated by a sequence of N-dimensional domains, and the

* Received by the editors January 25, 1993; accepted for publication (in revised form) April 5, 1993.

† Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 170/3 – Correo 3, Santiago, Chile (cconca@uchcecvm.bitnet). This work was partially supported by Consejo Nacional de Ciencia y Tecnología through Fondecyt grant 1201-91 and by D.T.I. grant E3099-9225.

‡ Departamento de Matemática Aplicada, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, 28040 Madrid, España (zuazua@mat.ucm.es). This work was partially supported by Project PB90-0245 of Dirección General de Investigatíon Científica y Tecnológica (Spain) and Eurhomogenization Project SC1-CT91-0732 of the EEC.

junction of both bodies thus comes about naturally. The paper of Ciarlet, Le Dret and Nzengwa (1989) has since been expanded, and its ideas extensively developed in the books by Ciarlet (1990) and by Le Dret (1991) [1]. The *analysis of multistructures* is an expanding field with many interesting contributions in the last few years. In the particular case of spectral problems we can cite the work of Bourquin and Ciarlet (1989), where the authors provide a mathematical justification of eigenvalue problems modeling junctions between bodies, and the paper by Le Dret (1990), which studies the vibrations of a folded plate.

The technique that we shall use here to deal with the multidimensional structure is totally different from that above. We shall employ a contact condition recently proposed by Puel and Zuazua (1991), (1992), in which the junction is dealt with explicitly, by means of a special boundary condition coupling one of the extremes of the one-dimensional interval with the boundary of the N-dimensional domain. As mentioned above, this is a local condition not a punctual one because not only A intervenes in it, but an open neighborhood γ of A that we have called the junction region does as well (see Fig. 1). For reasons of a practical nature, and in order to understand better the limit of validity of this way of modeling the junction, we study a simple spectral problem: the asymptotic behavior of the eigenvalues and eigenfunctions as γ goes to a set of Lebesgue measure zero. To this end we begin by introducing an abstract functional framework, in which a precise meaning is given to the convergence of the contact region γ, and we show the existence of a *limit-spectral problem* describing the asymptotic behavior mentioned above. As one might expect a priori, this problem can be separated out into two independent boundary subproblems, one of which is associated with the N-dimensional body, and the other with the string. Both of these are none other than the usual spectral problem for the Laplace operator. The role played by the junction between the bodies can only be appreciated at the level of the boundary condition at the extreme A of the string.

The precise meaning of the convergence is described later in this article. Let us for the moment simply mention that this convergence is formulated in terms of Green's operators and spectral families. In other words, we prove that the Green's operator associated with the original problem converges uniformly (that is, in norm) to that of the *limit-spectral problem*, and that the corresponding spectral families converge almost everywhere in $\mathbb{R}$. Neither of these results sheds much light on the convergence of the eigenvalues and eigenfunctions, and we shall therefore deal explicitly with this by means of proving some corollaries from these results.

Let us now say a few words about the methods used in the proof of our results of convergence. We shall employ general theorems of spectral and perturbation theory of linear operators, but in particular, we shall use two classical results of this theory, which can be referred to in the books by Sánchez-Hubert and Sánchez-Palencia (1989) and by Kato (1980). These are Theorems V.9.10 and V.11.1 on pages 205 and 211 in the first of these books, and Theorem VIII.1.14 on page 431 of the second. We first reduce our differential eigenvalue problem to that of finding the characteristic values of an operator T_ε (ε being a measure of the size of the junction region). In a similar way, we associate an operator T to the *limit-spectral problem*. The method for proving convergence has two steps: The first is to prove that the sequence $\{T_\varepsilon\}$ converges in norm to T. The next is to translate the above result into a convergence result for the resolvent map of the original problem and to apply the general theorems quoted above.

[1] For related topics the reader is referred to the references quoted in these books.

We feel that at this point it is appropriate to mention some possible generalizations of our work. The most obvious, which could be implemented without any difficulties beyond those outlined in the present article are, in the first place, the fact of considering *multistructures* with more than one string, joined to the N-dimensional body at different contact points, and secondly, working with second-order elliptic operators, more general than the Laplace operator. Another possible generalization, but one which could no longer be dealt with in the context of the techniques we introduce here, is the consideration of a junction between an N-dimensional body with a bounded domain of $\mathbb{R}^n$, $1 < n < N$.

To conclude this Introduction let us briefly discuss the content of this paper, section by section. In §1.1, we give a precise formulation of our model eigenvalue problem and introduce the abstract hypothesis that allows us to define the convergence of γ. Section 1.2 is devoted to introducing the Green's operator T_ε and to proving existence of eigenvalues and eigenfunctions. Next, in §1.3 we give a precise description of the *limit-spectral problem* and of its spectrum. In §1.4 we include the main results of convergence. Finally, in §2 we include some complementary results of convergence and we prove the main results of §1.

1. Formulation of the problem and convergence results.

1.1. Formulation of the problem. Let Ω be a bounded domain of $\mathbb{R}^N$ (N= 2 or 3 in the applications) with a locally Lipschitz boundary Γ. Let us denote by $\omega = (A, B)$ a one-dimensional interval of length $\ell \equiv |B - A|$, which we shall assume to be attached to the boundary of Ω by its lowest extreme A, as shown in Fig. 1.

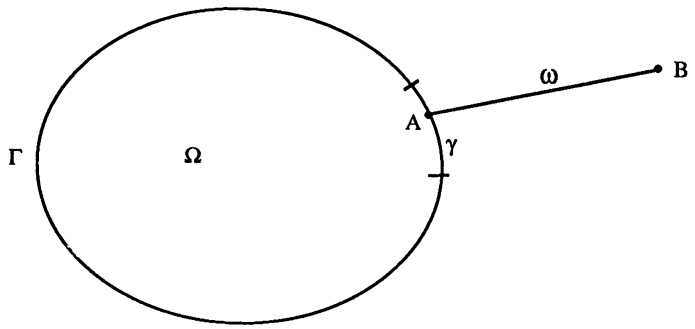


FIG. 1. *Graphical view of the multistructure.*

Several ways have been proposed of modeling the contact condition between the N-dimensional body Ω and the straight string ω. Our purpose in this article is to study some asymptotic properties of one of these models. This is the model recently introduced by Puel and Zuazua (1991), (1992). The admissible deformations of the multidimensional body are defined as follows: Let θ be a smooth function (say of class C^1), defined on Γ, with a compact support concentrated in a neighborhood of point A of Γ, and such that $\theta(A) = 1$, $\theta(x) \geq 0$ for all $x \in \Gamma$. A pair of independent deformations (u_Ω, u_ω) of Ω and ω, respectively, represents an admissible deformation of the multidimensional structure ($\Omega \cup \omega$) if the following contact condition is satisfied:

$$u_\Omega(x) = \theta(x) u_\omega(A) \quad \forall\, x \in \text{ supp}(\theta). \tag{1.1}$$

We shall call the function θ and its support, the function (or profile) and the contact region between Ω and ω, respectively.

By studying a simple spectral problem in $(\Omega \cup \omega)$, we want to analyze the limit behavior of this type of condition when the supports of the contact functions or profiles converge towards a set of measure zero which contains A.

To be more precise, let $\varepsilon > 0$ be a small parameter, intended to converge to zero, and let $\{\theta_\varepsilon\}_{\varepsilon>0}$ be a family of contact functions which converge to zero with ε, in a sense that we shall define below. We shall parametrize the interval $\omega = (A, B)$ using a real variable s in $(0, \ell)$, so that the value $s = 0$ will correspond to the contact point A, and the value $s = \ell$ to the point B. Given these conditions, we shall study the limit behavior, as $\varepsilon \to 0$, of the following sequence of eigenvalue problems in $\Omega \cup (0, \ell)$: Find $\lambda_\varepsilon \in \mathbb{R}$ for which there exist two functions $u_\varepsilon : \Omega \to \mathbb{R}$ and $v_\varepsilon : (0, \ell) \to \mathbb{R}$, not both identically zero, such that

$$-\Delta u_\varepsilon = \lambda_\varepsilon u_\varepsilon \quad \text{in } \Omega, \tag{1.2a}$$

$$-v_\varepsilon'' = \lambda_\varepsilon v_\varepsilon \quad \text{in } (0, \ell), \tag{1.2b}$$

$$u_\varepsilon = \theta_\varepsilon v_\varepsilon(0) \quad \text{on } \Gamma, \tag{1.2c}$$

$$v_\varepsilon'(\ell) = 0, \tag{1.2d}$$

$$v_\varepsilon'(0) = \int_\Gamma \frac{\partial u_\varepsilon}{\partial n} \theta_\varepsilon(x) d\Gamma(x), \tag{1.2e}$$

where $'$ denotes the derivation with respect to s[2], and $\frac{\partial}{\partial n}$ denotes the derivative with respect to the outward normal vector of Γ.

Note that the boundary condition (1.2c) can in fact be separated out into two conditions. On the support of θ_ε, (1.2c) is the same as the contact condition (1.1), written for $\theta = \theta_\varepsilon$ and identifying ω with $(0, \ell)$. On the rest of the boundary of Ω, (1.2c) reduces simply to the Dirichlet condition:

$$u_\varepsilon(x) = 0 \quad \forall x \in (\Gamma \setminus \operatorname{supp}(\theta_\varepsilon)). \tag{1.3}$$

With regard to the family of contact profiles, we shall assume that for every $\varepsilon > 0$, the functions θ_ε satisfy the following conditions:

$$\theta_\varepsilon \in H^{1/2}(\Gamma) \cap \mathcal{C}^0(\bar{\Gamma}), \tag{1.4a}$$

$$\theta_\varepsilon(A) = 1, \tag{1.4b}$$

$$\operatorname{supp}(\theta_\varepsilon) \subset \gamma, \tag{1.4c}$$

where γ is a fixed neighborhood (independent of ε) of A in Γ, such that the surface measure of $(\Gamma \setminus \bar{\gamma})$ is positive.

We shall also make the following hypotheses about its asymptotic behavior as $\varepsilon \to 0$:

$$\theta_\varepsilon(x) \longrightarrow 0 \quad \text{for almost all } x \in \Gamma, \text{ and} \tag{1.5a}$$

There exists a constant $c \geq 0$ such that

$$\|\nabla \tilde{\theta}_\varepsilon\|^2_{L^2(\Omega)^N} \longrightarrow c \quad \text{as } \varepsilon \to 0, \tag{1.5b}$$

[2] This derivative will sometimes also be denoted as -d.

where $\tilde{\theta}_\varepsilon \in H^1(\Omega)$ denotes the harmonic extension of θ_ε to Ω. Since Ω is a domain with a locally Lipschitz boundary and θ_ε satisfies (1.4a), such an extension is well defined, and can be characterized as the (unique) solution of the following nonhomogeneous Dirichlet boundary-value problem in Ω:

$$\Delta\tilde{\theta}_\varepsilon = 0 \quad \text{in } \Omega, \tag{1.6a}$$

$$\tilde{\theta}_\varepsilon = \theta_\varepsilon \quad \text{on } \Gamma. \tag{1.6b}$$

Hypotheses (1.5a) and (1.5b) might seem rather strange at first glance and there is no doubt that some comments and further explanations are needed. We shall present them in §1.5. There we shall also show some model examples of families $\{\theta_\varepsilon\}_\varepsilon$, which converge to zero in the sense given by these equations.

As we shall see below, the constant c of condition (1.5b) plays a crucial role in the results of convergence. In particular, they will be substantially different depending on whether $c > 0$ or $c = 0$.

Before moving on to study these results, and with a view to presenting the variational formulation of our original problem (1.2), we shall now introduce some notation. First we introduce the functional space

$$V_\varepsilon = \{(\varphi, \psi) \in H^1(\Omega) \times H^1(0,\ell) \mid \varphi = \theta_\varepsilon \psi(0) \text{ on } \Gamma\},$$

which we equip with the standard scalar product of $H^1(\Omega) \times H^1(0,\ell)$. Let us use $\|\cdot\|_{m,D}$ as the general notation for the usual norm of the spaces $H^m(D)^r$, $m \geq 0$ and $r = 1$, N or N^2, where D is any open set of $\mathbb{R}^N$ or $\mathbb{R}$, whether bounded or not. Applying standard arguments, it is a straightforward matter to prove that the seminorm

$$|(\varphi,\psi)|_{1,\Omega\times(0,\ell)} = \{\|\nabla\varphi\|^2_{0,\Omega} + \|\psi'\|^2_{0,(0,\ell)}\}^{1/2}$$

defines a norm in V_ε, equivalent to the norm induced by $H^1(\Omega) \times H^1(0,\ell)$ on V_ε. Equipped with this norm, V_ε is a Hilbert space.

For every function $v \in L^2(0,\ell)$, we shall denote the mean value of v by

$$\bar{v} = \frac{1}{\ell}\int_0^\ell v(s)ds,$$

and we shall designate by $L^2_0(0,\ell)$ the space of functions $L^2(0,\ell)$, with zero mean-value in $(0,\ell)$. For $v \in L^2(0,\ell)$, $\overset{\circ}{v}$ will denote the projection of v on $L^2_0(0,\ell)$, that is,

$$\overset{\circ}{v} = v - \bar{v}.$$

Let (φ,ψ) be any element of V_ε. Multiplying (1.2a) by φ and (1.2b) by ψ, and integrating by parts in Ω and in $(0,\ell)$, respectively, it can easily be proved, applying (1.2c,d,e) that the variational formulation of (1.2) is

(1.7a) Find $\lambda_\varepsilon \in \mathbb{R}$, $(u_\varepsilon, v_\varepsilon) \in V_\varepsilon$, $(u_\varepsilon, v_\varepsilon) \neq (0,0)$ such that

$$\int_\Omega \nabla u_\varepsilon \cdot \nabla\varphi dx + \int_0^\ell v'_\varepsilon \psi' ds = \lambda_\varepsilon \left\{ \int_\Omega u_\varepsilon \varphi dx + \int_0^\ell v_\varepsilon \psi ds \right\} \quad \forall (\varphi,\psi) \in V_\varepsilon. \tag{1.7b}$$

1.2. Existence result. Our starting point for the study of (1.2) (or (1.7)) is the question of existence of eigenvalues and eigenvectors. Our strategy for approaching it consists of defining an operator T_ε whose characteristic values coincide with the eigenvalues of (1.2). On the basis of the properties of T_ε, and applying general results of spectral theory of linear operators, we shall infer a theorem of existence for (1.2).

The operator T_ε that we will associate with (1.2) acts from the space $L^2(\Omega) \times L^2(0,\ell)$ into itself, and is simply defined as the Green's function corresponding to problem (1.7), that is,

$$T_\varepsilon(f,g) = (u_\varepsilon, v_\varepsilon) \quad \forall (f,g) \in L^2(\Omega) \times L^2(0,\ell), \tag{1.8}$$

where $(u_\varepsilon, v_\varepsilon)$ is the (unique) solution of the following variational problem in V_ε:

$$\text{Find } (u_\varepsilon, v_\varepsilon) \in V_\varepsilon \text{ such that} \tag{1.9a}$$

$$\int_\Omega \nabla u_\varepsilon \cdot \nabla \varphi dx + \int_0^\ell v_\varepsilon' \psi' ds = \int_\Omega f\varphi dx + \int_0^\ell g\psi ds \quad \forall (\varphi,\psi) \in V_\varepsilon. \tag{1.9b}$$

The existence and uniqueness of $(u_\varepsilon, v_\varepsilon)$ follow easily from the Lax–Milgram lemma, since the left side of (1.9b) defines a coercive bilinear form on V_ε, and its right side is a well-defined continuous linear form on V_ε. T_ε is therefore well defined. Furthermore, applying equally classical arguments, among which is included the fact that the canonical embedding of V_ε in $L^2(\Omega) \times L^2(0,\ell)$ is compact (which holds true, since Ω and $(0,\ell)$ are bounded), it is a straightforward matter to prove the following result.

PROPOSITION 1.1. *The operators T_ε satisfy*

(i) *T_ε is a continuous, self-adjoint, and compact operator on $L^2(\Omega) \times L^2(0,\ell)$;*

(ii) *T_ε is nonnegative definite in the following sense:*

$$(T_\varepsilon(f,g),(f,g))_{L^2(\Omega)\times L^2(0,\ell)} = \|\nabla u\|^2_{0,\Omega} + \|v'\|^2_{0,(0,\ell)} \geq 0 \quad \forall (f,g) \in L^2(\Omega) \times L^2(0,\ell).$$

As a consequence of Proposition 1.1, we deduce that the spectrum of T_ε consists of a countable infinite sequence of strictly positive real numbers converging to zero:

$$\mu_{1\varepsilon} \geq \mu_{2\varepsilon} \geq \cdots \geq \mu_{n\varepsilon} \geq \cdots \longrightarrow 0, \tag{1.10}$$

where, from now on, the eigenvalues will always be numbered in such a way that the same value is repeated as often as its (geometric) multiplicity. The corresponding eigenvectors will be denoted $(u_{1\varepsilon}, v_{1\varepsilon}), \ldots, (u_{n\varepsilon}, v_{n\varepsilon}), \ldots$. We assume that they have been chosen, forming an orthonormal basis of $L^2(\Omega) \times L^2(0,\ell)$.

Given that zero cannot be a characteristic value of T_ε, and that $\lambda_\varepsilon = 0$ cannot be a solution of (1.7), it is now clear that if μ_ε is a characteristic value of T_ε with $(u_\varepsilon, v_\varepsilon)$ as the corresponding eigenfunction, then $(1/\mu_\varepsilon, (u_\varepsilon, v_\varepsilon))$ is a solution of (1.7), and vice versa. Thus, if we define

$$\lambda_{j\varepsilon} = \frac{1}{\mu_{j\varepsilon}}, \qquad j = 1, \ldots, \tag{1.11}$$

grouping together the above results, we can establish the following theorem.

THEOREM 1.2. *The triplets* $\{\lambda_{j\varepsilon}, (u_{j\varepsilon}, v_{j\varepsilon})\}_{j\geq 1}$ *satisfy*
(i) *For each* $j \geq 1$, $\{\lambda_{j\varepsilon}, (u_{j\varepsilon}, v_{j\varepsilon})\}$ *is a solution of* (1.2) *and* (1.7);
(ii) *The set* $\{(u_{j\varepsilon}, v_{j\varepsilon})\}_{j\geq 1}$ *forms a Hilbert basis of* $L^2(\Omega) \times L^2(0,\ell)$;
(iii) $\lambda_{j\varepsilon} \longrightarrow +\infty$, *as* $j \to \infty$;
(iv) *These are all the solutions of* (1.2) *in the following sense: If* $\{\lambda_\varepsilon, (u_\varepsilon, v_\varepsilon)\}$ *is any solution of* (1.2), *then there exists* $j \geq 1$ *such that* $\lambda_\varepsilon = \lambda_{j\varepsilon}$ *and* $(u_\varepsilon, v_\varepsilon)$ *can be written as a linear combination of all* $(u_{j\varepsilon}, v_{j\varepsilon})$ *for which* $\lambda_{j\varepsilon} = \lambda_\varepsilon$.

1.3. Description of the limit problem and of its spectrum. As we shall see, under the hypotheses (1.4) and (1.5), the limit behavior of problem (1.2) is governed by two independent eigenvalue problems, one of them being formulated in the domain Ω, and the other in the interval $(0,\ell)$. In other words, the sequence of problems (1.2) converges, when $\varepsilon \to 0$, towards a *limit-spectral problem* in $\Omega \cup (0,\ell)$, that breaks down into two independent subproblems concerning each region of the multistructure. As one might expect a priori, the limit problem in Ω is none other than the spectral problem associated with the Laplace operator, with a homogeneous Dirichlet boundary condition on the boundary of Ω. The problem governing the limit behavior of (1.2) in $(0,\ell)$ is a little more unusual. This is the standard eigenvalue problem for the operator $-d^2$, but it has a boundary condition of the third type in $s = 0$, in which the constant c of condition (1.5b) is explicitly involved. If $c = 0$, this boundary condition is reduced quite simply to a homogeneous Neumann-type condition. At the other end of the interval, in $s = \ell$, the limit problem always retains the same boundary condition as (1.2), whether $c > 0$ or $c = 0$.

To be more precise, the spectral problem that governs the asymptotic behavior of (1.2) in Ω is: Find $\nu \in \mathbb{R}$ for which there exists $u \in H^1(\Omega), u \not\equiv 0$, such that

$$-\Delta u = \nu u \quad \text{in } \Omega, \tag{1.12a}$$

$$u = 0 \quad \text{on } \Gamma. \tag{1.12b}$$

The spectrum of this problem are the eigenvalues of $-\Delta$ in Ω with a homogeneous Dirichlet condition on Γ. We shall denote them as

$$0 < \nu_1 < \nu_2 \leq \cdots \leq \nu_m \leq \cdots \longrightarrow +\infty. \tag{1.13}$$

On the other hand, we shall prove that in $(0,\ell)$, the asymptotic behavior of (1.2) is governed by the following eigenvalue problem: Find $\mu \in \mathbb{R}$ for which there exists $v \in H^1(0,\ell), v \not\equiv 0$, such that

$$-v'' = \mu v \quad \text{in } (0,\ell), \tag{1.14a}$$

$$v'(0) = cv(0), \tag{1.14b}$$

$$v'(\ell) = 0. \tag{1.14c}$$

It is not difficult to check that the spectrum of (1.14) also consists of a sequence of nonnegative eigenvalues of finite multiplicity, which converge to $+\infty$. Furthermore, given that (1.14) is a second-order ordinary boundary-value problem, it is possible to calculate them almost explicitly. Indeed, if $c > 0$, by explicitly solving (1.14) one can prove that the eigenvalues are exactly the same as the positive roots of the nonlinear equation: $\sqrt{x}tg[\ell\sqrt{x}] = c$. There is therefore one eigenvalue in each of the following intervals: $(0, \frac{\pi}{2\ell}), (\frac{\pi}{2\ell}, \frac{3\pi}{2\ell}), (\frac{3\pi}{2\ell}, \frac{5\pi}{2\ell}), \ldots$ .They are all simple. Furthermore, if $c = 0$, then (1.14) can simply be reduced to the Neumann spectral problem, and in this case, the

eigenvalues are the set $\{k^2\pi^2/\ell^2\}_{k\geq 0}$. In both instances, we shall use the following notation for the eigenvalues of (1.14):

$$0 \leq \mu_1 < \mu_2 < \cdots < \mu_m < \cdots \longrightarrow +\infty. \tag{1.15}$$

Now, the spectral limit problem in $\Omega \cup (0,\ell)$ is merely these two problems put together, i.e., the problem of finding $\lambda \in \mathbb{R}$, for which there exists a pair $(u,v) \in H^1(\Omega) \times H^1(0,\ell)$ such that either $u \not\equiv 0$ and (λ, u) is a solution of (1.12), or $v \not\equiv 0$ and (λ, v) is a solution of (1.14). Of course, if λ is simultaneously an eigenvalue for (1.12) and (1.14) we may have both $u \not\equiv 0$ and $v \not\equiv 0$.

It is clear by definition that the spectrum of the boundary problem in $\Omega \cup (0,\ell)$ is formed exclusively of eigenvalues and, what is more, it is nothing other than the union of the sets $\{\nu_j\}_{j\geq 1}$ and $\{\mu_j\}_{j\geq 1}$. Before reordering the eigenvalues in increasing order, we shall denote the elements of this union by $\{\lambda_j\}$; thus

$$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m \leq \cdots \longrightarrow \infty.$$

The corresponding eigenfunctions are denoted by $(u_1, v_1), (u_2, v_2), \cdots$. Given that one can choose the eigenfunctions of (1.12) and (1.14) so that they form a Hilbert basis of $L^2(\Omega)$ and $L^2(0,\ell)$, respectively, it is also clearly possible to choose the pairs $\{(u_j, v_j)\}_{j\geq 1}$ to form a Hilbert basis of $L^2(\Omega) \times L^2(0,\ell)$. These functions thus automatically verify the following orthonormalization criterion:

$$\int_\Omega u_i u_j dx + \int_0^\ell v_i v_j ds = \delta_{ij} \quad \forall i,j \geq 1. \tag{1.16}$$

1.4. Results of convergence. Now that the notations and preliminary results have been established, we can proceed to present an initial result on the convergence of problems (1.2) to the boundary problems (1.12) and (1.14).

THEOREM 1.3. *Assume that the functions $\{\theta_\varepsilon\}_\varepsilon$ fulfill hypotheses* (1.4) *and* (1.5). *Let I be an open interval of $\mathbb{R}$ that, including multiplicities, contains m eigenvalues of the limit problem* (1.12), (1.14); *assume, let us say, that $\{\lambda_j\}_{j=p,p+m-1} \subset I$ for some $p \geq 1$. Then, for sufficiently small values of ε, $\{\lambda_{j\varepsilon}\}_{j=p,p+m-1} = \{\lambda_{j\varepsilon}\}_{j\geq 1} \cap I$.*

Let us denote as $P_{I\varepsilon} \in \mathcal{L}(L^2(\Omega) \times L^2(0,\ell))$ the orthogonal projection on the subspace spanned by the eigenfunctions associated with the eigenvalues $\{\lambda_{j\varepsilon}\}_j$ contained in I, and let $P_I \in \mathcal{L}(L^2(\Omega) \times L^2(0,\ell))$ be the projector on the eigenspace associated with the eigenvalues $\{\lambda_j\}_{j=p,p+m-1}$. Then $P_{I\varepsilon}$ converges uniformly (that is, in norm) to P_I. Thus,

$$\|P_{I\varepsilon} - P_I\|_{\mathcal{L}(L^2(\Omega)\times L^2(0,\ell))} \longrightarrow 0 \quad \text{as } \varepsilon \to 0. \tag{1.17}$$

In terms of the eigenfunctions and eigenvalues of (1.2), this result shows in particular that the eigenfunctions converge in $L^2(\Omega) \times L^2(0,\ell)$ strongly. More precisely, let us fix $j \geq 1$ and, so as not to complicate the analysis below, let us assume that λ_j is a simple eigenvalue of the limit problem. A first consequence of Theorem 1.3 is that for all ε, sufficiently small, $\lambda_{j\varepsilon}$ is also a simple eigenvalue of (1.2), and

$$\lambda_{j\varepsilon} \longrightarrow \lambda_j \quad \text{as } \varepsilon \to 0. \tag{1.18a}$$

Furthermore, if $(u_j, v_j) \in L^2(\Omega) \times L^2(0,\ell)$ denotes an eigenfunction corresponding to λ_j, normalized in accordance with (1.16), that is,

$$\|u_j\|^2_{0,\Omega} + \|v_j\|^2_{0,(0,\ell)} = 1, \tag{1.18b}$$

then from (1.17) it can be deduced that

$$\alpha_{j\varepsilon}(u_{j\varepsilon}, v_{j\varepsilon}) \longrightarrow (u_j, v_j) \quad \text{in } L^2(\Omega) \times L^2(0,\ell) \text{ strongly} \quad \text{as } \varepsilon \to 0, \tag{1.19a}$$

where

$$\alpha_{j\varepsilon} = \int_\Omega u_{j\varepsilon} u_j \, dx + \int_0^\ell v_{j\varepsilon} v_j ds, \tag{1.19b}$$

and $(u_{j\varepsilon}, v_{j\varepsilon})$ is an eigenfunction of (1.2) corresponding to $\lambda_{j\varepsilon}$, with norm 1. Since (u_j, v_j) also has norm 1, it follows from (1.19a) that $|\, \alpha_{j\varepsilon} \,| \longrightarrow 1$; accordingly, if the sign of the eigenfunction $(u_{j\varepsilon}, v_{j\varepsilon})$ is suitably chosen, then

$$(u_{j\varepsilon}, v_{j\varepsilon}) \longrightarrow (u_j, v_j) \quad \text{in } L^2(\Omega) \times L^2(0,\ell) \text{ strongly} \quad \text{as } \varepsilon \to 0. \tag{1.20}$$

In this sense, we can conclude that the simple eigenfunctions of (1.2) converge towards the corresponding eigenfunctions of the limit problem in $L^2(\Omega) \times L^2(0,\ell)$ strongly. If the corresponding eigenvalue is not simple, the eigensubspace will converge in the sense of the uniform convergence of the projections.

In §2.2, we shall prove a complementary property of convergence for (1.2), which will allow us to establish much more accurately the way in which the eigenfunctions converge. This is Lemma 2.3, which used together with Theorem 1.3 and the reasoning above makes it easy to deduce the following corollary.

COROLLARY 1.4. *Assume that the hypotheses of Theorem* 1.3 *are fulfilled. Let* $(u_j, v_j) \in L^2(\Omega) \times L^2(0,\ell)$ *be an eigenfunction of the limit problem* (1.12) and (1.14), *associated with an eigenvalue* λ_j. *Then there is a sequence* $\{(u_\varepsilon, v_\varepsilon)\}_\varepsilon$ *of eigenfunctions of* (1.2), *corresponding to the eigenvalue* $\lambda_{j\varepsilon}$, *such that, as* $\varepsilon \to 0$,

$$u_\varepsilon \longrightarrow u_j \quad \textit{in } H^1(\Omega) \textit{ weakly and } L^2(\Omega) \textit{ strongly}, \tag{1.21a}$$

$$v_\varepsilon \longrightarrow v_j \quad \textit{in } H^1(0,\ell) \textit{ strongly}. \tag{1.21b}$$

If the hypothesis (1.5b) *holds with* $c = 0$, *then the sequence* $\{(u_\varepsilon, v_\varepsilon)\}_\varepsilon$ *can be selected so that*

$$(u_\varepsilon, v_\varepsilon) \longrightarrow (u_j, v_j) \quad \textit{in } H^1(\Omega) \times H^1(0,\ell) \textit{ strongly} \quad \textit{as } \varepsilon \to 0. \tag{1.22}$$

It should be observed that when Lemma 2.3 is applied, different conclusions are reached depending on whether $c > 0$ or $c = 0$. Indeed, when $c > 0$ convergence (1.21a) does not hold in the strong topology of $H^1(\Omega)$. This is the main difference between the two cases.

1.5. Model examples of contact profile sequences. In this paragraph we shall present some model examples of functions $\{\theta_\varepsilon\}$ verifying hypotheses (1.4) and (1.5). In all the examples to be studied, it will be assumed that the boundary of the domain Ω, in addition to being locally Lipschitz, is at least of class C^1 in a neighborhood of point A, of contact between Ω, and the straight string ω. Therefore,

in what follows, we shall assume the existence of an open neighborhood U in $\mathbb{R}^N$ of A, and of an invertible mapping $\xi : x \to z = \xi(x)$, one time continuously differentiable from U onto the unit ball of $\mathbb{R}^N$ (that we shall denote B) such that

(1.23a) ξ^{-1} is one time continuously differentiable from B onto U,

(1.23b) $\xi(A) = 0$,

(1.23c) $\xi(U \cap \Omega) = \{z = (z', z_N) \in \mathbb{R}^N \mid |z| < 1,\ z_N > 0\}$,

(1.23d) $\xi(U \cap \Gamma) = \{z = (z', z_N) \in \mathbb{R}^N \mid |z'| < 1,\ z_N = 0\}$.

As is customary, we shall refer to the pair (U, ξ) as the system of local coordinates which define Γ in U.

Model Example 1. Let $\gamma_z \subset B'$ (the unit ball of $\mathbb{R}^{N-1}$) be an open neighborhood of the origin in $\mathbb{R}^{N-1}$ and let $\theta : \gamma_z \to \mathbb{R}$ be a $\mathcal{C}^1(\bar{\gamma}_z; \mathbb{R})$ function, with compact support in γ_z, such that $\theta(0) = 1$. Using the local system of coordinates (U, ξ), we define $\gamma \equiv \xi^{-1}(\{(z', 0) \mid z' \in \gamma_z\})$, and for each $\varepsilon > 0$ (small), $\gamma_\varepsilon \equiv \xi^{-1}(\{(z', 0) \mid z' \in \varepsilon\gamma_z\})$ (see Fig. 2). Based on this function θ, we shall define a family of contact functions $\{\theta_\varepsilon\}$ by the following rule:

$$x \in \Gamma \longrightarrow \theta_\varepsilon(x) = \begin{cases} \theta(z'/\varepsilon) & \text{if } x \in \gamma_\varepsilon, \\ 0 & \text{if } x \in (\Gamma \setminus \overline{\gamma_\varepsilon}), \end{cases} \tag{1.24}$$

where $(z', 0) = \xi(x)$. By construction, it is clear that the functions $\{\theta_\varepsilon\}$ satisfy the hypotheses (1.4a,b,c) and (1.5a). Let us show that (1.5b) holds, too. To this end, and to simplify the calculations, let us begin with the case in which the boundary Γ is plane in a neighborhood of A. Without loss of generality, we can assume in this case that the local coordinates coincide with the usual coordinates, that is, that $U = B, A = 0$, $\xi =$ identity map, $\gamma = \gamma_z$, $\gamma_\varepsilon = \varepsilon\gamma_z$. Thus $\Gamma \cap U = \{(x', x_N) \mid |x'| < 1,\ x_N = 0\}$ and the definition of θ_ε is reduced to

$$\theta_\varepsilon(x', 0) = \begin{cases} \theta(x'/\varepsilon) & \text{if } (x', 0) \in \gamma_\varepsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{1.25}$$

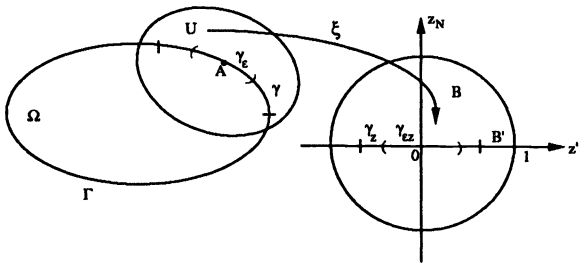


FIG. 2

Since $\text{supp}(\theta_\varepsilon) \subset \gamma_\varepsilon \subset \gamma$, in the nomenclature of Lions and Magenes (1978), the function θ_ε belongs to the space $H_{00}^{1/2}(\gamma)$, which is defined as the interpolated space (of exponent 1/2) between $L^2(\gamma)$ and $H_0^1(\gamma)$. Let us denote by $\|\cdot\|_{00,\gamma}$ the standard norm of interpolation in $H_{00}^{1/2}(\gamma)$. It is known that the mapping $\psi \in H_{00}^{1/2}(\gamma) \to \|\bar{\psi}\|_{1/2,\Gamma}$,

where $\bar{\psi}$ is the extension by zero outside γ, defines a norm in $H_{00}^{1/2}(\gamma)$, equivalent to $\|\cdot\|_{00,\gamma}$ (see Stephan (1987); this norm will be referred to as the norm induced by $H^{1/2}(\Gamma)$ in $H_{00}^{1/2}(\gamma)$).

To check hypothesis (1.5b), we will estimate $\|\theta_\varepsilon\|_{00,\gamma}$, and to this end we will explicitly compute the quantities $\|\theta_\varepsilon\|_{0,\gamma}$ and $|\ \theta_\varepsilon\ |_{1,\gamma}$, where $|\ \theta_\varepsilon\ |_{1,\gamma}$ denotes the norm $L^2(\gamma)^{N-1}$ of the (tangential) gradient of θ_ε. If the change of variables $y' = x'/\varepsilon$ is now introduced, we have

$$\|\theta_\varepsilon\|_{0,\gamma}^2 = \varepsilon^{N-1} \int_\gamma |\ \theta(y')\ |^2\, dy' = \varepsilon^{N-1} \|\theta\|_{0,\gamma}^2 \tag{1.26a}$$

and

$$|\ \theta_\varepsilon\ |_{1,\gamma}^2 = \varepsilon^{N-3} \int_\gamma |\ \nabla_{y'}\theta(y')\ |^2\, dy' = \varepsilon^{N-3} |\ \theta\ |_{1,\gamma}^2, \tag{1.26b}$$

and it therefore follows, interpolating between $L^2(\gamma)$ and $H_0^1(\gamma)$, that

$$\|\theta_\varepsilon\|_{00,\gamma}^2 \le \varepsilon^{N-2} \|\theta\|_{0,\gamma}\ |\ \theta\ |_{1,\gamma}\ . \tag{1.27}$$

Now, we know that the mapping $\psi \in H^{1/2}(\Gamma) \to \|\nabla\tilde{\psi}\|_{0,\Omega}$, where $\tilde{\psi}$ is the harmonic extension of ψ to Ω, defines a norm in $H^{1/2}(\Gamma)$, equivalent to the usual norm. Thus, equipping $H^{1/2}(\Gamma)$ with this norm, (1.27) proves that the sequence $\{\theta_\varepsilon\}$ satisfies (1.5b). It additionally fulfills this hypothesis with $c = 0$ if $N \ge 3$. Since the norm induced by $H^{1/2}(\Gamma)$ in $H_{00}^{1/2}(\gamma)$ is equivalent to the usual norm, there exists constants $d_1, d_2 > 0$ such that

$$d_1\|\psi\|_{00,\gamma} \le \|\nabla\tilde{\bar{\psi}}\|_{0,\Omega} \le d_2\|\psi\|_{00,\gamma} \quad \forall \psi \in H_{00}^{1/2}(\gamma),$$

and therefore (1.5b) is fulfilled with $c \le d_2^2\|\theta\|_{0,\gamma}\ |\ \theta\ |_{1,\gamma}$, if $N = 2$. The exact value of c depends on θ and Ω, and it has to be computed in each particular case. However, if $N = 2$, it is a straightforward matter to prove that c is strictly positive. Effectively, let us explicitly compute the norm $H_{00}^{1/2}(\gamma)$ of θ_ε. Using its definition, we have

$$\|\theta_\varepsilon\|_{00,\gamma}^2 = \int_{\mathbb{R}^{N-1}} (1 + |\tau'|^2)^{1/2} |\hat{\bar{\theta}}_\varepsilon(\tau', 0)|^2 d\tau',$$

where $\hat{\bar{\theta}}_\varepsilon(\tau', 0)$ is the Fourier transform of the extension by zero of θ_ε outside γ (τ' denotes the variable dual to x' in Fourier transform). Using (1.25), it can be easily checked that $\hat{\bar{\theta}}_\varepsilon(\tau', 0) = \varepsilon^{N-1}\hat{\bar{\theta}}(\varepsilon\tau')$, where $\bar{\theta}$ is the extension by zero of θ outside γ. Then,

$$\|\theta_\varepsilon\|_{00,\gamma}^2 = \varepsilon^{N-2} \int_{\mathbb{R}^{N-1}} (\varepsilon^2 + |\tau'|^2)^{1/2} |\hat{\bar{\theta}}(\tau')|^2 d\tau'.$$

Thus, if $N = 2$, it follows that

$$\lim_{\varepsilon\to 0} \|\theta_\varepsilon\|_{00,\gamma}^2 = \int_{\mathbb{R}} |\tau'| |\hat{\bar{\theta}}(\tau')|^2 d\tau' > 0.$$

Therefore, we can finally conclude that $\|\nabla\tilde{\theta}_\varepsilon\|^2_{0,\Omega}$ has a limit as $\varepsilon \to 0$, and

$$\lim_{\varepsilon\to 0} \|\nabla\tilde{\theta}_\varepsilon\|^2_{0,\Omega} \geq d_1^2 \int_{\mathbb{R}} |\tau'||\hat{\tilde{\theta}}(\tau')|^2 d\tau',$$

which implies $c > 0$.

To handle the general case, in which Γ is not necessarily plane in a neighborhood of A, all that is needed is to introduce the change of variables $x = \xi^{-1}(z)$ into the computation of the quantities $\|\theta_\varepsilon\|^2_{0,\gamma}$ and $|\ \theta_\varepsilon\ |^2_{1,\gamma}$. In fact a brief calculation, using the definition of θ_ε (see (1.25)) and the system of local coordinates, shows that

$$\|\theta_\varepsilon\|^2_{0,\gamma} = \int_{\varepsilon\gamma_z} \left|\theta\left(\frac{z'}{\varepsilon}\right)\right|^2 g(z')dz' = \varepsilon^{N-1} \int_{\gamma_z} |\ \theta(y')\ |^2\ g(\varepsilon y')dy', \tag{1.28a}$$

where $g(z') = \sqrt{|\det A^t(z')A(z')|}$ and $A(z')$ is the N by $(N-1)$ matrix whose entries are $[(\partial\xi_i^{-1}/\partial z_j)(z',0)]$.

By analogy we see that

(1.28b)

$$|\ \theta_\varepsilon\ |^2_{1,\gamma} = \int_{\varepsilon\gamma_z} \left|B^t(z')\nabla_{z'}\theta\left(\frac{z'}{\varepsilon}\right)\right|^2 g(z')dz' = \varepsilon^{N-3} \int_{\gamma_z} |\ B^t(\varepsilon y')\nabla_{y'}\theta(y')\ |^2\ g(\varepsilon y')dy',$$

where $B(z')$ is the $(N-1)$ by N matrix whose entries are $[(\partial\xi_i/\partial x_j)(\xi^{-1}(z',0))]$.

Given that the change of variables is invertible, using standard estimates it can be easily proved from (1.28) that there exist constants $C_1, C_2 > 0$, such that

$$\|\theta_\varepsilon\|^2_{0,\gamma} \ \leq\ C_1^2\varepsilon^{N-1}$$

and

$$|\ \theta_\varepsilon\ |^2_{1,\gamma} \leq\ C_2^2\varepsilon^{N-3}.$$

Thus interpolating we deduce

$$\|\theta_\varepsilon\|^2_{00,\gamma} \ \leq\ C_1C_2\varepsilon^{N-2}$$

and accordingly, in dimension $N \geq 3$, the sequence $\{\theta_\varepsilon\}$ satisfies (1.5b) with $c = 0$, and in dimension 2, as before, it can be easily deduced that $\{\theta_\varepsilon\}$ satisfies (1.5b) with $0 < c \leq d_2^2C_1C_2$. In this model example, it can be observed that the support of θ_ε converges towards the singleton set whose sole element is the contact point A.

Model example 2. The second example we will look at is very like the foregoing. The only difference is that this time the functions $\{\theta_\varepsilon\}$ will be constructed by only rescaling the first $(N-2)$ variables of θ. To be more precise, assuming that $N \geq 3$, the family of functions $\{\theta_\varepsilon\}$ will in this example be defined as follows:

$$x \in \Gamma \longrightarrow \theta_\varepsilon(x) = \begin{cases} \theta(z''/\varepsilon, z_{N-1}) & \text{if } x \in \gamma_\varepsilon, \\ 0 & \text{if } x \in (\Gamma \setminus \overline{\gamma_\varepsilon}), \end{cases} \tag{1.29}$$

where γ_ε is now defined by $\gamma_\varepsilon \equiv \xi^{-1}(\{(\varepsilon y'', y_{N-1}, 0) \mid (y'', y_{N-1}, 0) \in \gamma\})$ and $(z'', z_{N-1}, 0) = \xi(x)$. Once again it is easy to prove that θ_ε, thus defined, satisfies (1.4a,b,c) and (1.5a). For the sake of simplicity, we shall restrict ourselves in verifying

(1.5b) to the case in which Γ is plane in a neighborhood of A. Without loss of generality, let us assume accordingly that $\Gamma \cap U = \{(x'', x_{N-1}, x_N) \mid \mid (x'', x_{N-1}) \mid < 1,\ x_N = 0\}$ and that θ_ε is defined by

$$\theta_\varepsilon(x'', x_{N-1}, x_N) = \begin{cases} \theta(x''/\varepsilon, x_{N-1}) & \text{if } (x'', x_{N-1}, 0) \in \gamma_\varepsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{1.30}$$

If we use the change of variables $y' = (x''/\varepsilon, x_{N-1})$ in the definition of the quantities $\|\theta_\varepsilon\|_{0,\gamma}$ and $\mid \theta_\varepsilon \mid_{1,\gamma}$, we have

$$\|\theta_\varepsilon\|^2_{0,\gamma} = \int_{\gamma_\varepsilon} \left| \theta\left(\frac{x''}{\varepsilon}, x_{N-1}\right) \right|^2 dx'' dx_{N-1} = \varepsilon^{N-2} \|\theta\|^2_{0,\gamma} \tag{1.31a}$$

and

$$\mid \theta_\varepsilon \mid^2_{1,\gamma} = \varepsilon^{N-4} \int_\gamma \mid \nabla_{y''} \theta(y') \mid^2 dy' + \varepsilon^{N-2} \int_\gamma \left| \frac{\partial \theta}{\partial y_{N-1}}(y') \right|^2 dy'. \tag{1.31b}$$

Interpolating between these two identities, it can finally be concluded that

$$\|\theta_\varepsilon\|^2_{00,\gamma} \leq \varepsilon^{N-3} \|\theta\|_{0,\gamma} \left(\int_\gamma \mid \nabla_{y''} \theta(y') \mid^2 dy' \right)^{1/2} + O(\varepsilon^{N-2}).$$

Thus, if $N \geq 4$, condition (1.5b) holds with $c = 0$, and if $N = 3$, as in the previous examples, one can easily verify that θ_ε fulfills this hypothesis with a constant c such that $0 < c \leq d_2^2 \|\theta\|_{0,\gamma} (\int_\gamma \mid \nabla_{y''} \theta(y') \mid^2 dy')^{1/2}$. In this second model example, the support of θ_ε converges to a set of measure zero that is not reduced to a sole point.

2. Convergence of Green's operators and spectral families. In this section we shall prove Theorem 1.3 and other convergence results for the sequence of problems (1.2). This proof falls into two clearly defined parts. Firstly, it will be proved that the Green's operators associated with (1.2) converge uniformly towards the Green's operator of the limit problem (1.12) and (1.14). Secondly, a conclusion is reached based on some general results of perturbation theory of linear operators. In this particular instance, we shall use Theorem V.9.10 of Sánchez-Hubert and Sánchez-Palencia's book (1989, p. 205). It would, however, be equally possible to use Theorem VIII.1.14 in Kato's book (1980, p. 431). We shall also use general results of spectral theory to prove that the spectral families associated with the Green's operators converge strongly to the spectral family of the limit Green's operator.

2.1. A new family of Green's operators for (1.2). For technical reasons that will become clear below, we shall reformulate the problem in our study of (1.2), introducing the change of variable:

$$\kappa_\varepsilon = \lambda_\varepsilon + 1, \tag{2.1}$$

and replacing the operators $-\Delta$ and $-d^2$ by $(-\Delta + I)$ and $(-d^2 + I)$, respectively, in (1.2a) and (1.2b). Note that problem (1.2) remains exactly the same despite these changes.

Let S_ε denote the Green's operator associated with (1.2), but relative to the operators $(-\Delta + I)$ and $(-d^2 + I)$. That is, S_ε acts from $L^2(\Omega) \times L^2(0, \ell)$ into

itself, and it is defined by the following rule: $S_\varepsilon(f,g) = (u_\varepsilon, v_\varepsilon)$, where $(u_\varepsilon, v_\varepsilon)$ is the (unique) solution of the following variational problem:

(2.2a) Find $(u_\varepsilon, v_\varepsilon) \in V_\varepsilon$, such that

$$\text{(2.2b)} \quad \int_\Omega (\nabla u_\varepsilon \cdot \nabla \varphi + u_\varepsilon \varphi) dx + \int_0^\ell (v_\varepsilon' \psi' + v_\varepsilon \psi) ds = \int_\Omega f\varphi + \int_0^\ell g\psi ds \quad \forall (\varphi, \psi) \in V_\varepsilon.$$

By construction, it is clear that the set of characteristic values of S_ε is simply $\{\kappa_\varepsilon = 1 + \lambda_{j\varepsilon}\}_{j \geq 1}$. Furthermore, as for T_ε, it is a classical proof that S_ε is bounded, compact, and self-adjoint.

Exactly analogously, in the limit problems (1.12) and (1.14), let us change ν to $(\nu+1)$, μ to $(\mu+1)$ and let $-\Delta$ be replaced by $(-\Delta + I)$ and $-d^2$ by $(-d^2 + I)$. The Green's operator associated with the limit problem in $\Omega \cup (0,\ell)$, relative to these new operators, will thus be defined by the following rule:

$$S : L^2(\Omega) \times L^2(0,\ell) \longrightarrow L^2(\Omega) \times L^2(0,\ell),$$

$$\text{(2.3)} \quad S(f,g) = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix} = \begin{bmatrix} S_1 f \\ S_2 g \end{bmatrix},$$

where S_1 y S_2 are operators in $L^2(\Omega)$ and $L^2(0,\ell)$, respectively. S_1 is the Green's operator of $(-\Delta + I)$, with a homogeneous Dirichlet condition on Γ, and accordingly for $f \in L^2(\Omega)$, $S_1 f \equiv u$ is the (unique) solution of the following problem:

$$\text{(2.4a)} \quad -\Delta u + u = f \quad \text{in } \Omega,$$

$$\text{(2.4b)} \quad u = 0 \quad \text{on } \Gamma.$$

In turn, $S_2 : L^2(0,\ell) \to L^2(0,\ell)$ is defined by $S_2 g \equiv v$, where v is the (unique) solution of the following boundary-value problem in $(0,\ell)$:

$$\text{(2.5a)} \quad -v'' + v = g \quad \text{in } \Omega,$$

$$\text{(2.5b)} \quad v'(0) = cv(0),$$

$$\text{(2.5c)} \quad v'(\ell) = 0.$$

Since the operator $(-d^2 + I)$ appears in (2.5a) instead of $-d^2$, problem (2.5) admits a unique solution, even if $c = 0$. This is why we reformulated problem (1.2) in this way. Otherwise it would have been necessary to differentiate the cases $c > 0$ and $c \neq 0$ in the limit problem itself. It will be seen that this other approach allows both cases to be dealt with in a unified way.

As is usual, the solutions to (2.4) and (2.5) should be interpreted in a weak sense; in this case, as solutions of the following variational problems:

(2.6a) Find $u \in H_0^1(\Omega)$ such that

$$\text{(2.6b)} \quad \int_\Omega (\nabla u \cdot \nabla \varphi + u\varphi) dx = \int_\Omega f\varphi dx \quad \forall \varphi \in H_0^1(\Omega)$$

and

(2.7a) Find $v \in H^1(0,\ell)$ such that

$$\int_0^\ell (v'\psi' + v\psi)ds + cv(0)\psi(0) = \int_0^\ell g\psi ds \quad \forall \psi \in H^1(0,\ell). \tag{2.7b}$$

By applying Rellich's compactness theorem, the Lax–Milgram lemma, and the symmetry of the bilinear forms occurring in (2.6) and (2.7), it is easy to check that S belongs to $\mathcal{L}(L^2(\Omega) \times L^2(0,\ell))$, and that it is compact and self-adjoint. Finally it should be observed that, by construction, the set of characteristic values of S is nothing but $\{1+\lambda_j\}_{j\geq 1}$, where the λ_j are the eigenvalues of the limit problem (1.12) and (1.14).

2.2. Convergence of Green's operators. (The first part of the proof of Theorem 1.3.) We shall begin by examining the asymptotic behavior of the sequence of contact functions $\{\theta_\varepsilon\}_\varepsilon$. To this end, let us accept for the moment the following result, whose simple proof we present at the end of §2.2.

PROPOSITION 2.1. *Assume that the sequence $\{\theta_\varepsilon\}_\varepsilon$ satisfies hypotheses* (1.4) *and* (1.5). *Then, as $\varepsilon \to 0$, we have*

$$\tilde{\theta}_\varepsilon \rightharpoonup 0 \quad \text{in } H^1(\Omega) \text{ weakly}, \tag{2.8}$$

where $\tilde{\theta}_\varepsilon$ is the harmonic extension of θ_ε to Ω, as in (1.6).

Moreover, if condition (1.5b) *holds with $c = 0$, then*

$$\tilde{\theta}_\varepsilon \longrightarrow 0 \quad \text{in } H^1(\Omega) \text{ strongly} \quad \text{as } \varepsilon \to 0. \tag{2.9}$$

The proof of Theorem 1.3 is essentially based on the following result, which explains the convergence of the Green's operators.

THEOREM 2.2. *Suppose that the functions $\{\theta_\varepsilon\}_\varepsilon$ fulfill hypotheses* (1.4) *and* (1.5). *Then the sequence of operators $\{S_\varepsilon\}_\varepsilon$ converges uniformly towards the operator S, that is,*

$$\|S_\varepsilon - S\|_{\mathcal{L}(L^2(\Omega)\times L^2(0,\ell))} \longrightarrow 0 \quad \text{as } \varepsilon \to 0. \tag{2.10}$$

In turn, the proof of Theorem 2.2 requires the following lemma.

LEMMA 2.3. *Assume the hypotheses of Theorem* 2.2 *hold true. Let $\{(f_\varepsilon, g_\varepsilon)\}_\varepsilon$ be a sequence of functions in $L^2(\Omega) \times L^2(0,\ell)$, satisfying*

$$f_\varepsilon \rightharpoonup f \quad \text{in } L^2(\Omega) \text{ weakly} \quad \text{as } \varepsilon \to 0, \tag{2.11a}$$

$$g_\varepsilon \rightharpoonup g \quad \text{in } L^2(0,\ell) \text{ weakly} \quad \text{as } \varepsilon \to 0, \tag{2.11b}$$

and let us denote $(u_\varepsilon, v_\varepsilon) = S_\varepsilon(f_\varepsilon, g_\varepsilon)$ and $(u,v) = S(f,g)$. Then, as $\varepsilon \to 0$, we have

$$u_\varepsilon \longrightarrow u \quad \text{in } H^1(\Omega) \text{ weakly} \quad \text{and} \quad L^2(\Omega) \text{ strongly}, \tag{2.12a}$$

$$v_\varepsilon \longrightarrow v \quad \text{in } H^1(0,\ell) \text{ strongly}. \tag{2.12b}$$

Moreover, there exists a rest function $r_\varepsilon \in H^1(\Omega)$ *such that*

$$u_\varepsilon = u + v_\varepsilon(0)\tilde{\theta}_\varepsilon + r_\varepsilon, \tag{2.13a}$$

$$r_\varepsilon \longrightarrow 0 \quad \text{in } H^1(\Omega) \text{ strongly} \quad \text{as } \varepsilon \to 0. \tag{2.13b}$$

If, on the other hand, hypothesis (1.5b) *holds with* $c = 0$, *then*

$$u_\varepsilon \longrightarrow u \quad \text{in } H^1(\Omega) \text{ strongly} \quad \text{as } \varepsilon \to 0. \tag{2.14}$$

Proof. By definition, $(u_\varepsilon, v_\varepsilon)$ is the solution of (2.2) (with $f = f_\varepsilon$ and $g = g_\varepsilon$). So, $(u_\varepsilon, v_\varepsilon) \in V_\varepsilon$ and

$$\int_\Omega (\nabla u_\varepsilon \cdot \nabla \varphi + u_\varepsilon \varphi) dx + \int_0^\ell (v_\varepsilon' \psi' + v_\varepsilon \psi) ds = \int_\Omega f_\varepsilon \varphi dx + \int_0^\ell g_\varepsilon \psi ds \quad \forall (\varphi, \psi) \in V_\varepsilon. \tag{2.15}$$

Let us take $(\varphi, \psi) = (u_\varepsilon, v_\varepsilon)$ as test function in (2.15); we obtain

$$\|u_\varepsilon\|^2_{1,\Omega} + \|v_\varepsilon\|^2_{1,(0,\ell)} = \int_\Omega f_\varepsilon u_\varepsilon dx + \int_0^\ell g_\varepsilon v_\varepsilon ds.$$

Since $\{(f_\varepsilon, g_\varepsilon)\}_\varepsilon$ is bounded in $L^2(\Omega) \times L^2(0,\ell)$, using Cauchy–Schwarz's inequality and standard arguments, we deduce from this latter identity an estimate of the following form:

$$\|u_\varepsilon\|^2_{1,\Omega} + \|v_\varepsilon\|^2_{1,(0,\ell)} \leq C,$$

where C is a constant, independent of ε. It is then possible to extract subsequences, which we shall continue to denote as $\{u_\varepsilon\}$ and $\{v_\varepsilon\}$ respectively, and there are functions $(u, v) \in H^1(\Omega) \times H^1(0,\ell)$, such that

$$u_\varepsilon \longrightarrow u \quad \text{in } H^1(\Omega) \text{ weakly and } L^2(\Omega) \text{ strongly} \quad \text{as } \varepsilon \to 0, \tag{2.16a}$$

$$v_\varepsilon \longrightarrow v \quad \text{in } H^1(0,\ell) \text{ weakly and } L^2(0,\ell) \text{ strongly} \quad \text{as } \varepsilon \to 0. \tag{2.16b}$$

The next step of the proof consists in passing to the limit in (2.15), and proving that u and v are solutions of (2.4) and (2.5), respectively. To this end, let φ be any function of $\mathcal{C}_0^\infty(\Omega)$, and let us take as test function in (2.15) the pair $(\varphi, 0)$ (which is possible, since $\varphi = 0$ on Γ, and therefore $(\varphi, 0) \in V_\varepsilon$). It follows that

$$\int_\Omega (\nabla u_\varepsilon \cdot \nabla \varphi + u_\varepsilon \varphi) dx = \int_\Omega f_\varepsilon \varphi dx.$$

But this identity passes trivially to the limit under the convergence (2.11a) and (2.16a). It is thus proved that u is a solution of (2.6b), since φ is an arbitrary function of $\mathcal{C}_0^\infty(\Omega)$, which is a dense subspace of $H_0^1(\Omega)$. On the other hand, passing punctually to the limit for almost every $x \in \Gamma$, it follows from (1.5a) that

$$u(x) = \lim_{\varepsilon \to 0} v_\varepsilon(0)\theta_\varepsilon(x) = 0 \quad \text{for almost all } x \in \Gamma,$$

since $\{v_\varepsilon(0)\}_\varepsilon$ is bounded in $\mathbb{R}$, thanks to (2.16b). Thus $u \in H_0^1(\Omega)$ and u is therefore the (weak) solution of (2.4).

Let us now prove that v is the solution of (2.5). To do this, let us start by taking as test function in (2.15) a pair of the form $(0,\psi)$, with ψ an arbitrary function in the space $H \equiv \{\psi \in H^1(0,\ell) \mid \psi(0)=0\}$. We obtain the identity

$$\int_0^\ell (v_\varepsilon'\psi' + v_\varepsilon\psi)ds = \int_0^\ell g_\varepsilon\psi ds,$$

which, using (2.11b) and (2.16b), passes to the limit and allows us to deduce that v satisfies

$$\int_0^\ell (v'\psi' + v\psi)ds = \int_0^\ell g\psi ds \quad \forall\psi \in H.$$

Therefore, v is a weak solution of (2.5a) and (2.5c). Now, if we integrate (2.5a) between 0 and ℓ, it follows that

$$v'(0) + \ell\bar{v} = \ell\bar{g}. \tag{2.17}$$

Let us, on the other hand, take $(\varphi,\psi) = (\tilde{\theta}_\varepsilon, 1)$ as test function in (2.15) (which is possible, since θ_ε verifies (1.4b) and therefore $(\tilde{\theta}_\varepsilon,1) \in V_\varepsilon$). We obtain

$$\int_\Omega (\nabla u_\varepsilon \cdot \nabla\tilde{\theta}_\varepsilon + u_\varepsilon\tilde{\theta}_\varepsilon)dx + \ell\bar{v}_\varepsilon = \int_\Omega f_\varepsilon\tilde{\theta}_\varepsilon dx + \ell\bar{g}_\varepsilon.$$

From Proposition 2.1 we know that $\{\tilde{\theta}_\varepsilon\}$ converges to zero in $H^1(\Omega)$ weakly and in $L^2(\Omega)$ strongly. Thus the second term on the left and first on the right of this identity will converge to zero, as $\varepsilon \to 0$. On the other hand, $\bar{g}_\varepsilon \to \bar{g}$ and $\bar{v}_\varepsilon \to \bar{v}$. It can therefore be concluded that the sequence $\{\int_\Omega \nabla u_\varepsilon \cdot \nabla\tilde{\theta}_\varepsilon dx\}_\varepsilon$ has a limit, and that furthermore

$$\lim_{\varepsilon\to 0} \int_\Omega \nabla u_\varepsilon \cdot \nabla\tilde{\theta}_\varepsilon dx + \ell\bar{v} = \ell\bar{g}. \tag{2.18a}$$

Now, if $c=0$, then the term on the left side converges to zero and as a consequence, $\ell\bar{g} = \ell\bar{v}$. Comparing with (2.17) we deduce that if $c=0$, then $v'(0)=0$, and the proof that v is a solution of (2.5) is completed.

Now let us look at the case $c>0$ and prove that $v'(0) = cv(0)$. To do this, let us multiply (1.6a) by u_ε, and integrate by parts in Ω. Using the contact condition $u_\varepsilon = v_\varepsilon(0)\theta_\varepsilon$ on Γ, it follows that

$$\int_\Omega \nabla\tilde{\theta}_\varepsilon \cdot \nabla u_\varepsilon dx = v_\varepsilon(0) \int_\Gamma \frac{\partial\tilde{\theta}_\varepsilon}{\partial n}\theta_\varepsilon d\Gamma(x)$$

but

$$\int_\Gamma \frac{\partial\tilde{\theta}_\varepsilon}{\partial n}\theta_\varepsilon d\Gamma(x) = \frac{1}{2}\int_\Gamma \frac{\partial}{\partial n}(\tilde{\theta}_\varepsilon)^2 d\Gamma(x) = \frac{1}{2}\int_\Omega \Delta(\tilde{\theta}_\varepsilon)^2 dx,$$

and as $\tilde{\theta}_\varepsilon$ is harmonic in Ω, the following identity holds:

$$\int_\Gamma \frac{\partial \tilde{\theta}_\varepsilon}{\partial n}\theta_\varepsilon d\Gamma = \|\nabla \tilde{\theta}_\varepsilon\|^2_{0,\Omega}.$$

Thus we finally conclude

$$\int_\Omega \nabla \tilde{\theta}_\varepsilon \cdot \nabla u_\varepsilon dx = v_\varepsilon(0)\|\nabla \tilde{\theta}_\varepsilon\|^2_{0,\Omega}, \tag{2.18b}$$

which in fact, as will be appreciated, (2.18b) is also valid if $c = 0$. Combining (2.18a) with (2.18b) and using (1.5b), if $c > 0$, we can conclude that the sequence $\{v_\varepsilon(0)\}_\varepsilon$ is convergent, and that its limit verifies the following identity:

$$c \lim_{\varepsilon \to 0} v_\varepsilon(0) = \ell\bar{g} - \ell\bar{v}.$$

Now, since the canonical embedding of $H^1(0,\ell)$ into $C^0((\overline{0,\ell}))$ is compact, then it follows from (2.16b) that $v_\varepsilon(0) \to v(0)$ as $\varepsilon \to 0$. Thus it holds that if $c > 0$, then

$$cv(0) = \ell\bar{g} - \ell\bar{v}. \tag{2.19}$$

Comparing (2.19) with (2.17), it can be deduced that v satisfies (2.5b), and the proof of the fact that v is the (weak) solution of (2.5) is concluded. As problems (2.4) and (2.5) admit only one solution, (u, v) is the only possible cluster point of $\{(u_\varepsilon, v_\varepsilon)\}$; in (2.16) the whole sequence will then converge.

Let us now prove that in (2.16b) the convergence takes place in $H^1(0,\ell)$ strongly. To this end we shall study the convergence of the energy. Since $(u_\varepsilon, v_\varepsilon)$ is a solution of (2.15), v is solution of the following equations:

$$-v_\varepsilon'' + v_\varepsilon = g_\varepsilon \quad \text{in } (0,\ell), \tag{2.20a}$$

$$v_\varepsilon'(\ell) = 0. \tag{2.20b}$$

Multiplying (2.20a) by v_ε and integrating by parts between 0 and ℓ, it follows that

$$\|v_\varepsilon\|^2_{1,(0,\ell)} + v_\varepsilon'(0)v_\varepsilon(0) = \int_0^\ell g_\varepsilon v_\varepsilon ds.$$

But a simple integration of (2.20a) between 0 and ℓ shows that $v_\varepsilon'(0) + \ell\bar{v}_\varepsilon = \ell\bar{g}_\varepsilon$, and therefore

$$\|v_\varepsilon\|^2_{1,(0,\ell)} = \int_0^\ell g_\varepsilon v_\varepsilon ds - v_\varepsilon(0)(\ell\bar{g}_\varepsilon - \ell\bar{v}_\varepsilon),$$

and passing to the limit (using (2.11b) and (2.16b)), we obtain

$$\lim_{\varepsilon \to 0} \|v_\varepsilon\|^2_{1,(0,\ell)} = \int_0^\ell gv ds - v(0)(\ell\bar{g} - \ell\bar{v}).$$

In other words, if we use (2.17),

$$\lim_{\varepsilon\to 0}\|v_\varepsilon\|^2_{1,(0,\ell)} = \int_0^\ell gvds - v(0)v'(0). \tag{2.21a}$$

On the other hand, taking $\psi = v$ in (2.7b), we have

$$\begin{aligned}\|v\|^2_{1,(0,\ell)} &= \int_0^\ell gvds - cv(0)^2\\ \text{(and by (2.5b))} \ &= \int_0^\ell gvds - v'(0)v(0).\end{aligned} \tag{2.21b}$$

Combining (2.21a) and (2.21b), we conclude that

$$\lim_{\varepsilon\to 0}\|v_\varepsilon\|^2_{1,(0,\ell)} = \|v\|^2_{1,(0,\ell)}.$$

As in addition $v_\varepsilon \rightharpoonup v$ in $H^1(0,\ell)$ weakly, this completes the proof of (2.12b).

Let us now prove that the expression $v_\varepsilon(0)\tilde\theta_\varepsilon$ is a first-order corrector for the sequence $\{u_\varepsilon\}$, i.e., let us prove (2.13). To this end, let us begin by pointing out that all that is needed is to prove that

$$(u_\varepsilon - v_\varepsilon(0)\tilde\theta_\varepsilon) \longrightarrow u \quad \text{in } H^1(\Omega) \text{ strongly} \quad \text{as } \varepsilon\to 0. \tag{2.22}$$

Now, $(u_\varepsilon - v_\varepsilon(0)\tilde\theta_\varepsilon) \rightharpoonup u$ in $H^1(\Omega)$ weakly as $\varepsilon\to 0$. To prove (2.22) it is therefore enough to prove the convergence of the norms. Let us then consider the identity

$$\|u_\varepsilon - v_\varepsilon(0)\tilde\theta_\varepsilon\|^2_{1,\Omega} = \|u_\varepsilon\|^2_{1,\Omega} + v^2_\varepsilon(0)\|\tilde\theta_\varepsilon\|^2_{1,\Omega} - 2v_\varepsilon(0)(u_\varepsilon,\tilde\theta_\varepsilon)_{1,\Omega}. \tag{2.23}$$

Using hypothesis (1.5b), Proposition 2.1, and the identities (2.18a) and (2.18b), one can pass to the limit in the last two terms on the right side. We obtain

$$v^2_\varepsilon(0)\|\tilde\theta_\varepsilon\|^2_{1,\Omega} \longrightarrow v^2(0)c \quad \text{as } \varepsilon\to 0 \quad \text{and} \tag{2.24a}$$

$$2v_\varepsilon(0)(u_\varepsilon,\tilde\theta_\varepsilon)_{1,\Omega} \longrightarrow 2v^2(0)c \quad \text{as } \varepsilon\to 0. \tag{2.24b}$$

Taking $(u_\varepsilon, v_\varepsilon(0))$ as test function in (2.15), on the other hand, it follows that

$$\|u_\varepsilon\|^2_{1,\Omega} + v_\varepsilon(0)\ell\bar v_\varepsilon = \int_\Omega f_\varepsilon u_\varepsilon dx + v_\varepsilon(0)\ell\bar g_\varepsilon.$$

Replacing in (2.23) and passing to the limit (using (2.24), (2.11), and (2.16)),

$$\lim_{\varepsilon\to 0}\|u_\varepsilon - v_\varepsilon(0)\tilde\theta_\varepsilon\|^2_{1,\Omega} = \int_\Omega fudx - v(0)(cv(0) - (\ell\bar g - \ell\bar v)).$$

But $cv(0) = v'(0) = \ell\bar g - \ell\bar v$ (by virtue of (2.5b) and (2.18a)) and $\|u\|^2_{1,\Omega} = \int_\Omega fudx$, because u is a solution of (2.4). Hence, we have

$$\lim_{\varepsilon\to 0}\|u_\varepsilon - v_\varepsilon(0)\tilde\theta_\varepsilon\|^2_{1,\Omega} = \|u\|^2_{1,\Omega},$$

which finishes the proof of (2.13).

Finally, (2.14) is a direct consequence of (2.13) and of the result of convergence (2.9) of Proposition 2.1. This completes the proof of Lemma 2.3. □

Proof of Theorem 2.2. To prove Theorem 2.2 by means of Lemma 2.3, we shall use contradictory reasoning. Let us suppose then that (2.10) does not hold. Then there exist $\delta > 0$ and a sequence $\{(f_\varepsilon, g_\varepsilon)\}_\varepsilon$ in $L^2(\Omega)\times L^2(0,\ell)$, with $\|f_\varepsilon\|^2_{0,\Omega}+\|g_\varepsilon\|^2_{0,(0,\ell)} = 1$, such that

$$\|(S_\varepsilon - S)(f_\varepsilon, g_\varepsilon)\|_{L^2(\Omega)\times L^2(0,\ell)} \geq \delta. \tag{2.25}$$

However, since $L^2(\Omega) \times L^2(0,\ell)$ is a Hilbert space, by extracting a subsequence we can assume

$$(f_\varepsilon, g_\varepsilon) \rightharpoonup (f, g) \quad \text{in } L^2(\Omega) \times L^2(0,\ell) \text{ weakly} \quad \text{as } \varepsilon \to 0.$$

But, using Lemma 2.3 and the fact that S is compact, it follows that

$$(S_\varepsilon - S)(f_\varepsilon, g_\varepsilon) \longrightarrow (0,0) \quad \text{in } L^2(\Omega) \times L^2(0,\ell) \text{ strongly} \quad \text{as } \varepsilon \to 0,$$

which is clearly in contradiction with (2.25). To finish the proof of Theorem 2.2, it remains to prove Proposition 2.1. □

Proof of Proposition 2.1. Hypothesis (1.5b) implies in particular that the sequence $\{\|\nabla\tilde{\theta}_\varepsilon\|_{0,\Omega}\}$ remains bounded as $\varepsilon \to 0$. Since $\tilde{\theta}_\varepsilon$ vanishes on a fixed part of Γ (see (1.4c)), if we apply the generalized Poincare's inequality, it follows that $\{\tilde{\theta}_\varepsilon\}$ is bounded in $H^1(\Omega)$. A subsequence can therefore be extracted, still denoted $\{\tilde{\theta}_\varepsilon\}$, which converges in $H^1(\Omega)$ weakly. Say

$$\tilde{\theta}_\varepsilon \rightharpoonup \theta \quad \text{in } H^1(\Omega) \text{ weakly} \quad \text{as } \varepsilon \to 0. \tag{2.26}$$

Given that $\tilde{\theta}_\varepsilon$ is harmonic in Ω, θ will be too, and passing to the limit in (1.6b), applying (1.5a), it follows that $\theta = 0$ on Γ. Thus, $\theta \equiv 0$, and in (2.26) the whole sequence converges.

Finally, if $c = 0$, hypothesis (1.5b) and the generalized Poincare's inequality prove that the convergence of the functions $\tilde{\theta}_\varepsilon$ to zero take place in $H^1(\Omega)$ strongly. Proposition 2.1 is thus proved. □

Proof of Theorem 1.3 (second part). As mentioned in the introduction to Chapter 2, Theorem 1.3 is a consequence of Theorem 2.2 and of Theorem V.9.10 in Sánchez-Hubert and Sánchez-Palencia's book (1989, p. 205). With a view to applying these results, let us now begin by identifying our starting system (1.2) with the spectral problem associated with an unbounded operator A_ε in $L^2(\Omega) \times L^2(0,\ell)$.

Let the domain $D(A_\varepsilon) \subset L^2(\Omega) \times L^2(0,\ell)$ be defined as follows: $(u,v) \in D(A_\varepsilon)$ if and only if

$$\Delta u \in L^2(\Omega), \tag{2.27a}$$

$$v'' \in L^2(0,\ell), \tag{2.27b}$$

$$u = v(0)\theta_\varepsilon \quad \text{on } \Gamma, \tag{2.27c}$$

$$v'(0) = \int_\Gamma \frac{\partial u}{\partial n}\theta_\varepsilon(x)d\Gamma(x), \tag{2.27d}$$

$$v'(\ell) = 0. \tag{2.27e}$$

Let us observe, in passing, that (2.27b) implies that v does indeed belong to $H^2(0,\ell)$. Thus $v(0), v'(0)$, and $v'(\ell)$ are well defined. On the other hand, since $\theta_\varepsilon \in H^{\frac{1}{2}}(\Gamma)$, it is clear from (2.27a) and (2.27c) that u belongs to $H^1(\Omega)$. It is worth recalling that if $u \in H^1(\Omega)$ and $\Delta u \in L^2(\Omega)$, then the traces $\{u, \frac{\partial u}{\partial n}\} \mid_\Gamma$ are both well defined elements of $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$, respectively. Therefore, it is possible to interpret the right-hand side of (2.27d) as the standard duality bracket between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$.

We can now define the operator A_ε by the following rule:

$$A_\varepsilon : D(A_\varepsilon) \longrightarrow L^2(\Omega) \times L^2(0,\ell),$$

(2.28)

$$A_\varepsilon(u,v) = (-\Delta u, -v'').$$

One can easily verify that the spectral problem of A_ε (formally) coincides with system (1.2). The main properties of A_ε are summarized in the following proposition.

PROPOSITION 2.4. *A_ε enjoys the following properties*:

(i) *A_ε is densely defined in $L^2(\Omega) \times L^2(0,\ell)$*;

(ii) *A_ε is closed*;

(iii) *$(A_\varepsilon + I)^{-1} = S_\varepsilon$, where S_ε is defined in* §2.1.

Proof. (i) The proof is an immediate consequence of the fact that $\mathcal{D}(\Omega) \times \mathcal{D}(0,\ell)$ is contained in $D(A_\varepsilon)$. To prove (ii), let us take a sequence $\{(u_n, v_n)\}_n$ in $D(A_\varepsilon)$ such that, as $n \to \infty$,

$$(u_n, v_n) \longrightarrow (u,v) \quad \text{in } L^2(\Omega) \times L^2(0,\ell), \tag{2.29a}$$

$$(-\Delta u_n, -v_n'') \longrightarrow (w,z) \quad \text{in } L^2(\Omega) \times L^2(0,\ell). \tag{2.29b}$$

The aim is to show $(u,v) \in D(A_\varepsilon)$ and $(w,z) = A_\varepsilon(u,v)$. First, it is clear that $w = -\Delta u$, $g = -v''$ and so it suffices to prove that (u,v) belongs to $D(A_\varepsilon)$. Next, we observe that $u_n \to u$ in the space $H(\Omega,\Delta) \equiv \{u \in L^2(\Omega) \mid \Delta u \in L^2(\Omega)\}$ and $v_n \to v$ in $H((0,\ell),d^2) \equiv \{v \in L^2(0,\ell) \mid d^2 v \in L^2(0,\ell)\}$. Since the maps $u \to u \mid_\Gamma$ and $v \to v(0)$ are continuous from $H(\Omega,\Delta)$ onto $H^{-1/2}(\Gamma)$ and from $H((0,\ell),d^2)$ onto $\mathbb{R}$, respectively, we can pass to the limit in the contact condition:

$$u_n = v_n(0)\theta_\varepsilon \quad \text{on } \Gamma. \tag{2.30}$$

Then, it follows that (u,v) verifies (2.27c). Furthermore, this also implies that $u_n \mid_\Gamma \to u \mid_\Gamma$ in $H^{1/2}(\Gamma)$, as $n \to \infty$. Thus, using (2.29), it is straightforward to prove that $\nabla u_n \to \nabla u$ in $L^2(\Omega)^N$ strongly. Since the trace map $\nabla u \to \nabla u \cdot n = \frac{\partial u}{\partial n} \mid_\Gamma$ is continuous from $H(\mathrm{div},\Omega)$ onto $H^{-1/2}(\Gamma)$ we see that

$$\int_\Omega \frac{\partial u_n}{\partial n}\theta_\varepsilon d\Gamma(x) \longrightarrow \int_\Gamma \frac{\partial u}{\partial n}\theta_\varepsilon d\Gamma(x) \quad \text{as } n \to \infty.$$

Analogously, the trace maps $v \to (v'(0), v'(\ell))$ are continuous from $H((0,\ell),d^2)$ onto $\mathbb{R} \times \mathbb{R}$. Therefore, we can pass to the limit in the following boundary conditions:

$$v_n'(0) = \int_\Gamma \frac{\partial u_n}{\partial n}\theta_\varepsilon d\Gamma(x) \quad \text{and} \quad v_n'(\ell) = 0,$$

and prove that (u,v) verifies (2.27d,e). Thus $(u,v) \in D(A_\varepsilon)$ and the proof of (ii) is completed.

To prove (iii), let $(f,g) \in L^2(\Omega) \times L^2(0,\ell)$ be given and let us consider the equation: Find $(u_\varepsilon, v_\varepsilon) \in D(A_\varepsilon)$ such that

$$(A_\varepsilon + I)(u_\varepsilon, v_\varepsilon) = (f,g). \tag{2.31}$$

From the definitions of A_ε and S_ε, (2.31) is clearly equivalent to the variational problem (2.2), and for this reason allows only one solution, which is none other than $S_\varepsilon(f,g)$. S_ε is thus the inverse (in the sense of nonbounded operators, i.e., by the left) of $(A_\varepsilon + I)$. Proposition 2.4 is therefore proved. □

Wholly analogously, one can identify the limit problem (2.4) and (2.5) with the spectral problem of a nonbounded operator A in $L^2(\Omega) \times L^2(0,\ell)$. This operator has the following domain: $(u,v) \in D(A)$ if and only if

$$u \in H^2(\Omega) \cap H_0^1(\Omega), \tag{2.32a}$$

$$v \in H^2(0,\ell), \tag{2.32b}$$

$$v'(0) = cv(0), \tag{2.32c}$$

$$v'(\ell) = 0, \tag{2.32d}$$

and it is defined as follows:

$$\begin{aligned} &A : D(A) \longrightarrow L^2(\Omega) \times L^2(0,\ell), \\ &A(u,v) = (-\Delta u, -v''). \end{aligned} \tag{2.33}$$

Applying analogous arguments to those used in proving Proposition 2.4, it is easily shown that A is a closed, densely defined operator, and that $(A+I)^{-1} = S$.

Once the operators A_ε and A have been introduced, all that is needed to apply Theorem V.9.10 of Sánchez-Hubert and Sánchez-Palencia (1989, p. 205) is to observe that the convergence result (2.10) of Theorem 2.2 is equivalent to the fact that the resolvent mapping of A_ε in -1 converges in norm (i.e., uniformly) towards the resolvent of A in -1. In accordance with the theorem quoted above, it is sufficient to conclude Theorem 1.3 and, what is more, it also allows us to conclude that if $\mu \in \rho(A)$ (resolvent set of A), then we also have $\mu \in \rho(A_\varepsilon)$ for sufficiently small ε, and

$$\|(A_\varepsilon - \mu)^{-1} - (A - \mu)^{-1}\|_{\mathcal{L}(L^2(\Omega)\times L^2(0,\ell))} \longrightarrow 0 \quad \text{as } \varepsilon \to 0. \tag{2.34}$$

2.3. Convergence of spectral families. Let $E(S_\varepsilon, \cdot) : \mathbb{R} \longrightarrow \mathcal{L}(L^2(\Omega)\times L^2(0,\ell))$ and $E(S,\cdot) : \mathbb{R} \longrightarrow \mathcal{L}(L^2(\Omega) \times L^2(0,\ell))$ denote the spectral families associated with the operators S_ε and S, respectively. If we apply Rellich's theorem (which is referred to in Sánchez-Hubert and Sánchez-Palencia (1989, p. 211, Theorem V.11.1)) to the sequence $\{S_\varepsilon\}$, the following result is deduced, which shows the convergence of the family of projections $E(S_\varepsilon,\cdot)$ towards $E(S,\cdot)$.

THEOREM 2.5. *Under the hypotheses of Theorem* 1.3, *if* λ *is not an eigenvalue of* S, *then* $E(S_\varepsilon,\lambda)$ *converges strongly to* $E(S,\lambda)$ *as* $\varepsilon \to 0$, *i.e., for all* $(u,v) \in L^2(\Omega) \times L^2(0,\ell)$, *we have*

$$E(S_\varepsilon,\lambda)(u,v) \longrightarrow E(S,\lambda)(u,v) \quad \textit{in } L^2(\Omega) \times L^2(0,\ell) \textit{ strongly}. \tag{2.35}$$

The foregoing result of convergence involves the spectral families of S_ε and S. However, this can be easily translated in terms of the families of projections of A_ε and A if we take the following relations into account:

$$E(A_\varepsilon, \lambda) = I - E\left(S_\varepsilon, \frac{1}{(\lambda+1)}\right)^* \tag{2.36a}$$

$$E(A, \lambda) = I - E\left(S, \frac{1}{(\lambda+1)}\right)^*, \tag{2.36b}$$

where the asterisk $(*)$ indicates that the spectral family is taken to be strongly continuous from the left, instead of the classical convention "strongly continuous from the right."

Acknowledgments. Part of this research was carried out while C. Conca was visiting the Centre de Mathématiques Appliquées de l'Ecole Polytechnique de Paris with the support of the Service de Coopération du Gouvernement Français au Chili. He is grateful to J.-C. Nédélec for having made this visit possible and pleasant. We wish both to thank J.-P. Puel for comments and fruitful discussions.

REFERENCES

F. Bourquin and Ph. Ciarlet (1989), *Modeling and justification of eigenvalue problems for junctions between elastic structures*, J. Funct. Anal., 87, pp. 392–427.

Ph. Ciarlet, H. Le Dret, and R. Nzengwa (1989), *Junctions between* 3d *and* 2d *linearly elastic structures*, J. Math. Pures Appl., 68, pp. 261–295.

Ph. Ciarlet (1990), *Plates and Junctions in Elastic Multistructures. An Asymptotic Analysis*, RMA 14, Masson, Paris.

T. Kato (1980), *Perturbation Theory for Linear Operators* 2nd ed., Springer-Verlag, New York.

H. Le Dret (1990), *Vibrations of a folded plate*, Math. Modeling Numer. Anal., 24, pp. 501–521.

——— (1991), *Problèmes Variationnels dans les Multi-Domaines. Modélisation des Jonctions et Applications*, RMA 19, Masson, Paris.

J.-L. Lions and E. Magenes (1978), *Problèmes aux Limites Non Homogènes*, Vol. 1, Masson, Paris.

J.-P. Puel and E. Zuazua (1993), *Exact controllability for a model of multidimensional flexible structure*, Proc. Roy. Soc. Edinburgh Sect. A, 123A, pp. 323–344.

——— (1992), *Contrôlabilité exacte et stabilisation d'un modèle de structure vibrante multidimensionnelle*, C. R. Acad. Sci. Paris Sér. I, t314, pp. 121–125.

J. Sánchez-Hubert and E. Sánchez-Palencia (1989), *Vibration and Coupling of Continuous Systems. Asymptotic Methods*, Springer-Verlag, Berlin.

E.-P. Stephan (1987), *Boundary integral equations for screen problems in* $\mathbb{R}^3$, Integral Equations Operator Theory, 10, pp. 236–257.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 859–875, May 1994

003

ON CHARACTERIZATION OF SOLUTIONS OF SOME NONLINEAR DIFFERENTIAL EQUATIONS AND APPLICATIONS*

MARIE-FRANÇOISE BIDAUT-VERON† AND MUSTAPHA BOUHAR†

Abstract. This paper gives a complete classification of the solutions of the equation $\omega'' - \mu\omega + |\omega|^{q-1}\omega = 0$ on $S^1 = \mathbb{R}/2\pi\mathbb{Z}$, where $\mu, q \in \mathbb{R}$ with $q > 1$, and compares their different levels of energy. Thus the asymptotical behavior and all the possible orbits are found for the parabolic one-dimensional equation $u_t = u_{xx} + |u|^{q-1}u - \mu u$ and the elliptic two-dimensional equation $\Delta u - c|x|^{-2}\, u + |u|^{q-1}u = 0$, where $c > 0$.

Key words. nonlinear differential equations, periodic solutions, levels of energy, asymptotic behavior, elliptic and parabolic equations

AMS subject classifications. 34C, 35J, 35K

Introduction. In this paper we study the following nonlinear eigenvalue problem on $S^1 = \mathbb{R}/2\pi\mathbb{Z}$:

$$\omega'' - \mu\omega + |\omega|^{q-1}\omega = 0, \tag{0.1}$$

where $\mu, q \in \mathbb{R}$, with $q > 1$. We give the complete structure of the set of solutions of this equation and study their respective levels of energy, according to the possible values of μ. We give some extensions to more general functions of ω than $|\omega|^{q-1}\omega$.

The main interest of (0.1) lies in the fact that it describes the equilibrium states for some parabolic or elliptic semilinear equations.

Consider first the one-dimensional well-known parabolic equation with a source term:

$$u_t = u_{xx} + |u|^{q-1}u - \mu u. \tag{0.2}$$

The study of (0.1) allows us to give the precise, large-time behavior of solutions of (0.2) with periodic, Dirichlet, or Neumann boundary conditions, when applying the convergence results of Cazenave and Lions [11] and Matano [21]; we also find some results of Matano [22] concerning global solutions.

Now consider the two-dimensional elliptic equation with an inverse square potential:

$$\Delta u - c\frac{u}{|x|^2} + |u|^{q-1}u = 0, \tag{0.3}$$

where $c > 0$ is the most interesting case. Then we can get the behavior of $|x|^{2/(q-1)}u$ near the origin or infinity with the help of (0.1). Let us make the classical transformation in polar coordinates:

$$u(r,\theta) = r^{-\delta}v(t,\theta), \quad t = -\mathrm{Log}\, r, \quad r > 0, \quad \theta \in S^1, \tag{0.4}$$

where

$$\delta = 2/(q-1). \tag{0.5}$$

* Received by the editors April 27, 1992; accepted for publication (in revised form) January 25, 1993.

† Département de Mathématiques, Faculté des Sciences, Parc de Grandmont, 37200 Tours, France.

The function v satisfies an elliptic equation in a cylinder:

$$v_{tt} + 2\delta v_t + v_{\theta\theta} - (c - \delta^2)v + |v|^{q-1}v = 0; \tag{0.6}$$

therefore, (0.1) describes the equilibrium states of (0.6) when $\mu = \delta^2 - c$.

In the N-dimensional case with $c = 0$, (0.3) can be reduced in the radial case to Emden's equation [14]; it was solved by Fowler [15] many years ago. In the nonradial case, Brezis and Lions [8], [20] gave the local behavior of u when $c = 0$ and $N = 2$ or $N \geq 3$ and $q < N/(N-2)$; when $q = N/(N-2)$ the problem was solved by Aviles [3] and by Caffarelli, Gidas, and Spruck [10], [16] when $q \leq (N+2)/(N-2)$. Those results were recently improved by Bidaut-Veron and Veron [5] who consider a larger range for q and give estimates for any c when q is undercritical. We shall see that when $N = 2$ and $c > (2/(q-1))^2$, the interference between the Laplacian, the potential, and the nonlinearity is very strong, and the situation is almost quite as rich as in the case $N \geq 3$ without any potential.

Notice that (0.1) is more difficult to study than the equation with the other sign:

$$-\omega'' - \mu\omega + |\omega|^{q-1}\omega = 0, \tag{0.7}$$

which was first considered in the classical paper of Chafee and Infante [12]; it is linked to the parabolic problem with an absorption term:

$$u_t = u_{xx} + \mu\omega - |\omega|^{q-1}\omega, \tag{0.8}$$

and to the elliptic equation in $\mathbb{R}^2/\{0\}$:

$$-\Delta u - c\frac{u}{|x|^2} + |u|^{q-1}u = 0. \tag{0.9}$$

Equation (0.9) was investigated by Chen, Matano, and Veron [13] when $c = 0$, in the N-dimensional case with $c = 0$ by Veron [29], [30], then for general c by Guerch and Veron [18].

Concerning (0.1), our main result is the following.

THEOREM 0.1. *For any real μ, let E_μ be the set of solutions of* (0.1) *on* S^1:

$$E_\mu = \{0\} \cup E_\mu^+ \cup E_\mu^- \cup \tilde{E}_\mu,$$

where E_μ^+ is the set of positive solutions and $E_\mu^- = -E_\mu^+$, and $\tilde{E}_\mu$ is the set of changing sign solutions. Then

(i) *$\tilde{E}_\mu$ has an infinity of one-dimensional connected components*:

$$\tilde{E}_\mu = \bigcup_{k=\tilde{k}(\mu)}^{+\infty} \tilde{C}_k,$$

where $\tilde{k}(\mu)$ is the smallest positive integer such that $k^2 + \mu > 0$, and $\tilde{C}_k$ is generated by a function $\tilde{\omega}_k$ with least period $2\pi/k$: $\tilde{C}_k = \{\tilde{\omega}_k(\cdot + \phi) | \phi \in S^1\}$;

(ii) *If $\mu \leq 0, E_\mu^+ = \emptyset$. If $\mu \in (0, 1/(q-1)), E_\mu^+ = \{\mu^{1/(q-1)}\}$. If $\mu > 1/(q-1)$, then E_μ^+ has a finite number of one-dimensional connected components*:

$$E_\mu^+ = \{\mu^{1/(q-1)}\} \cup \left(\bigcup_{k=1}^{k_+(\mu)} C_k^+ \right),$$

where $k_+(\mu)$ *is the largest integer smaller than* $((q-1)\mu)^{1/2}$, *and* C_k^+ *is generated by a positive function* $\omega_{+,k}$ *with least period* $2\pi/k$.

This gives us the complete bifurcation diagram for (0.1): bifurcations near 0 at the eigenvalues $\mu = -n^2, n \in \mathbb{N}/\{0\}$, bifurcations near the constant solutions at the eigenvalues $\mu = n^2/(q-1), n \in \mathbb{N}/\{0\}$.

Now consider the energy function $\mathbf{E}$ defined for any $\omega \in E_\mu$ by

$$\mathbf{E}(\omega) = \int_{S^1} \left(\frac{1}{2}\omega'^2 + \frac{\mu}{2}\omega^2 - \frac{|\omega|^{q+1}}{q+1} \right) d\theta = \frac{q-1}{2(q+1)} \int_{S^1} |\omega|^{q+1} d\theta; \tag{0.10}$$

we prove in particular that the functions $k \to \mathbf{E}(\omega_k^+)$ and $k \to \mathbf{E}(\tilde{\omega}_k)$ are increasing ones; therefore, the energies of positive solutions (or changing sign solutions) are classified by the connected components to which they belong. This way we can know what connecting orbits might exist for global solutions of (0.2) and (0.3). Determining what connections do exist is still an open question. Notice that this difficult problem was solved for (0.8) by Brunovsky and Fiedler [9], Henry [19], and for (0.9) by Matano [23].

Our paper is organized as follows:

(1) The eigenvalue problem on S^1;

(2) Application to the parabolic problem;

(3) Application to the elliptic problem.

1. The eigenvalue problem on S^1. Here we deal with equation (0.1) on S^1, and more generally with the ordinary differential equation on whole $\mathbb{R}$:

$$\omega'' - \mu\omega + g(\omega) = 0, \tag{1.1}$$

where $\mu \in \mathbb{R}$, and g satisfies the assumptions

$$\begin{cases} g \in C^1(\mathbb{R}) \cap C^3(\mathbb{R}/\{0\}), \quad g(0) = g'(0) = 0; \\ rg''(r) > 0 \quad \text{for any } r \neq 0; \quad \text{and} \quad \lim\limits_{r \to \pm\infty} g(r)/r = +\infty. \end{cases} \tag{1.2}$$

This equation can be viewed as a Hamiltonian equation with the potential $U/2$, where

$$U(\omega) = 2G(\omega) - \mu\omega^2, \qquad G(r) = \int_0^r g(s)ds. \tag{1.3}$$

If $\mu \leq 0, U$ is a nonnegative convex function with $U(0) = 0$. If $\mu > 0$, U is decreasing on $(-\infty, a_-) \cup (0, a_+)$, increasing on $(a_-, 0) \cup (a_+, +\infty)$, and $U(b_-) = U(b_+) = 0$, where a_+, a_-, b_+, b_- are defined by

$$\begin{cases} g(a_\pm) = \mu a_\pm, \qquad 2G(b_\pm) = \mu b_\pm^2, \\ b_- < a_- < 0 < a_+ < b_+. \end{cases} \tag{1.4}$$

Obviously the constant solutions of (1.1) are 0, and a_+, a_- when $\mu > 0$. From standard phase-plane methods, when $\mu \leq 0$ all the solutions are periodic, all nontrivial solutions are changing sign functions, and their extremal values can be any nonzero real number. When $\mu > 0$, all the solutions are periodic but one (the function $t \mapsto ((q+1)\mu/2\cosh^2((q-1)\sqrt{\mu}\, t/2))^{1/(q-1)}$ when $g(r) = |r|^{q-1}r$) and its translated ones. There is still a family of changing sign functions, with any maximal points in

$(b_+, +\infty)$ and minimal ones in $(-\infty, b_-)$; there is also a family of positive solutions, oscillating around the constant solution $\omega \equiv a_+$, with any maximal values in (a_+, b_+) and minimal ones in $(0, a_+)$; in the same way there is a family of negative solutions oscillating around $\omega \equiv a_-$.

Among all those solutions we look for 2π-periodic ones. By translation we can suppose that $\omega(0) = 0$ if ω is a changing sign function, $\omega(0) = a_+$ if ω is positive.

First we study how the least period of the changing sign solutions of (1.1) depends on the initial slope.

LEMMA 1.1. *For any $\alpha > 0$, let $\tilde{\omega}(.,\alpha)$ be the solution of* (1.1) *on $\mathbb{R}$ such that $\tilde{\omega}(0,\alpha) = 0$ and $\tilde{\omega}'(0,\alpha) = \alpha$, and let $\tilde{P}(\alpha)$ be its least period. Then, under the assumption* (1.2), *$\tilde{P}$ is decreasing from $(0, +\infty)$ to $(0, 2\pi/\sqrt{-\mu})$ if $\mu < 0$, from $(0, +\infty)$ to $(0, +\infty)$ if $\mu \geq 0$.*

Proof. For any $s > 0$, let $\tilde{z}(s)$ be the positive solution of the equation $U(\tilde{z}(s)) = s^2$, and $\tilde{y}(s)$ be the negative one. Since $\tilde{\omega}(.,\alpha)$ satisfies the equation $\omega'^2 + U(\omega) = \alpha^2$, we can write, following an idea of [1],

$$\tilde{P}(\alpha) = 2(Q(\tilde{z}(\alpha)) - Q(\tilde{y}(\alpha))), \tag{1.5}$$

where

$$Q(\xi) = \int_0^{\xi} \frac{dv}{\sqrt{U(\xi) - U(v)}} = \xi \int_0^1 \frac{d\tau}{\sqrt{U(\xi) - U(\tau\xi)}}. \tag{1.6}$$

We claim that Q is decreasing, on $\mathbb{R}/\{0\}$ if $\mu \leq 0$, on $(-\infty, b_-) \cup (b_+, +\infty)$ if $\mu > 0$. Indeed we get by differentiation, valid since $U'(\xi) \neq 0$,

$$Q'(\xi) = \int_0^1 \frac{\Theta(\xi) - \Theta(\tau\xi)}{(U(\xi) - U(\tau\xi))^{3/2}} d\tau, \tag{1.7}$$

where

$$\Theta(\xi) = U(\xi) - \xi U'(\xi)/2; \tag{1.8}$$

now $\Theta''(\xi) = -\xi U^{(3)}(\xi)/2 = -\xi g''(\xi)$; therefore, from (1.2) Θ is a concave nonpositive function with $\Theta(0) = 0$; henceforth, $Q'(\xi) < 0$. From (1.5), $\tilde{P}$ is decreasing, since $\tilde{z}'(\alpha) > 0$ and $\tilde{y}'(\alpha) < 0$.

Let us now look at the limits near zero and infinity. Suppose first that $\mu \leq 0$; then we have

$$\tilde{P}(\alpha) = 2 \int_0^{\alpha} (\tilde{z}' - \tilde{y}')(s) \frac{ds}{\sqrt{\alpha^2 - s^2}} = 2 \int_0^1 (\tilde{z}' - \tilde{y}')(\alpha t) \frac{dt}{\sqrt{1 - t^2}} \tag{1.9}$$

and $\tilde{z}'(s) = 2s/U'(\tilde{z}(s)), \tilde{y}'(s) = 2s/U'(\tilde{y}(s))$. We easily get $\lim_{s\to+\infty} \tilde{z}'(s) = \lim_{s\to+\infty} \tilde{y}'(s) = 0$ and $\lim_{s\to 0} \tilde{z}'(s) = -\lim_{s\to 0} \tilde{y}'(s) = 1/\sqrt{-\mu}$ ($+\infty$ if $\mu = 0$). Thus we obtain the limits of $\tilde{P}$ from the Lebesgue theorem when $\mu < 0$; when $\mu = 0$ we verify that $\tilde{z}' - \tilde{y}'$ is increasing and conclude by the Beppo–Levi theorem. Suppose now that $\mu > 0$; then we have

$$\tilde{P}(\alpha) = 2 \int_{b_-}^{b_+} \frac{dv}{\sqrt{\alpha^2 - U(v)}} + 2 \int_0^1 (\tilde{z}' - \tilde{y}')(\alpha t) \frac{dt}{\sqrt{1 - t^2}}, \tag{1.10}$$

and $\lim_{s\to+\infty} \tilde{z}'(s) = \lim_{s\to+\infty} \tilde{y}'(s) = \lim_{s\to 0} \tilde{z}'(s) = \lim_{s\to 0} \tilde{y}'(s) = 0$; hence we get the limits of $\tilde{P}$, since $1/\sqrt{-U(v)}$ is nonintegrable near the origin. □

From the change of concavity of g at 0, the two functions $\alpha \mapsto Q(\tilde{z}(\alpha))$ and $\alpha \mapsto -Q(\tilde{y}(\alpha))$ are both decreasing; that means that the length of positive arches and the length of negative ones vary in the same way. The contrary holds for one sign solutions, for example, positive ones that oscillate around the constant solution $\omega \equiv a_+$, and the question is harder. In the next lemma we establish that the length of the upper arches (above a_+) and the length of the lower ones (under a_+) vary in opposite ways; and this is because the concavity of g does not change at a_+. To conclude we need an additional assumption on function g.

LEMMA 1.2. *Suppose that $\mu > 0$. For any $\beta \in (0, \sqrt{-U(a_+)})$ let $\omega_+(.,\beta)$ be the positive solution of* (1.1) *on $\mathbb{R}$ such that $\omega_+(0,\beta) = a_+$ and $\omega'_+(0,\beta) = \beta$, and let $P_+(\beta)$ be its least period. Then under* (1.2) *we have*

$$\lim_{\beta \to 0} P_+(\beta) = 2\pi/\sqrt{g'(a_+) - \mu}; \qquad \lim_{\beta \to \sqrt{-U(a_+)}} P_+(\beta) = +\infty. \tag{1.11}$$

Moreover, suppose that

$$r \mapsto (g'(r) - \mu)^{-2/3} \quad \text{is convex on } ((g')^{-1}(\mu), +\infty) \tag{1.12}$$

(for example, $g(r) = |r|^{q-1}r, q > 1$); then P_+ is increasing on $(0, \sqrt{-U(a_+)})$.

Proof. For any $s \in (0, \sqrt{-U(a_+)})$, let $z(s)$ be the solution of the equation $U(z(s)) - U(a_+) = s^2$ greater than a_+, and let $y(s)$ be the solution smaller than a_+. Since $\omega_+(.,\beta)$ satisfies the equation $\omega'^2 + U(\omega) - U(a_+) = \beta^2$, we can write

$$\begin{aligned} P_+(\beta) &= 2\int_{a_+}^{z(\beta)} \frac{dv}{\sqrt{\beta^2 - U(v) + U(a_+)}} - 2\int_{a_+}^{y(\beta)} \frac{dv}{\sqrt{\beta^2 - U(v) + U(a_+)}} \\ &= 2\int_0^1 (z' - y')(\beta t)\frac{dt}{\sqrt{1-t^2}}; \end{aligned} \tag{1.13}$$

in particular, $P_+(\beta) \geq 2\int_{y(\beta)}^{a_+} dv/\sqrt{-U(v)}$. Hence $\lim_{\beta \to \sqrt{-U(a_+)}} P_+(\beta) = +\infty$, since $\sqrt{-U}$ is nonintegrable at 0. On the other hand, we easily get $\lim_{s \to 0}(z' - y')(s) = 2\sqrt{2/U''(a_+)} = 2/\sqrt{g'(a_+) - \mu}$; therefore, $\lim_{\beta \to 0} P_+(\beta) = 2\pi/\sqrt{g'(a_+) - \mu}$ by the Lebesgue theorem.

Suppose now (1.12); for any $s \in (0, \sqrt{-U(a_+)})$ we have

$$z''(s) = 2\frac{U'^2 - 2U''(U - U(a_+))}{U'^3}(z(s)), \tag{1.14}$$

and similarly for $y(s)$. But for any $v \in (0, b^+)$,

$$(U'^2 - 2U''(U - U(a_+)))'(v) = -2U^{(3)}(v)(U(v) - U(a_+)) \leq 0,$$

since $U^{(3)}(v) = 2g''(v) > 0$. Then $z''(s)$ and $y''(s)$ are both nonpositive, since $U'(y(s)) < 0 < U'(z(s))$. By l'Hôpital rule, applied twice, we get $\lim_{s \to 0} z''(s) = \lim_{s \to 0} y''(s) = -2U^{(3)}(a_+)/3U''^2(a_+)$. Let us prove that

$$z''(s) \geq -2U^{(3)}(a_+)/3U''^2(a_+) \geq y''(s) \quad \text{in } (0, \sqrt{-U(a_+)}). \tag{1.15}$$

By using (1.14) it is enough to prove that the function

$$F = U'^2 - 2(U - U(a_+))U'' + (U^{(3)}(a_+)/3U''^2(a_+))U'^3 \tag{1.16}$$

is nonnegative on $(0, b_+)$. Let $c = g'^{-1}(\mu)$; then from (1.2), (1.3), U'' is negative on $(0, c)$, and

$$F' = -2(U - U(a_+))U^{(3)} + (U^{(3)}(a_+)/U''^2(a_+))U'^2U'';$$

therefore, F is decreasing on $(0, c)$. Now consider the function $H = F/U''$ on $(c, +\infty)$; we get $H' = U'^2U^{(3)}K/U''^2$, where

$$K = -1 + (U^{(3)}(a_+)/3U''^2(a_+))(3U''^2/U^{(3)} - U').$$

Then $K(a_+) = 0$ and

$$(3U''^2(a_+)/U^{(3)}(a_+))U''(U^{(3)})^{-2}K' = 5(U^{(3)})^2 - 3U^{(4)}U''. \tag{1.17}$$

From (1.12), $(U'')^{-2/3}$ is convex on $(c, +\infty)$; therefore, K' is nonnegative, H is nonincreasing on (c, a_+), nondecreasing on $(a_+, +\infty)$, then nonnegative on $(c, +\infty)$. Henceforth F is nonnegative on $(0, +\infty)$, positive on $(0, c)$, so we get (1.15) with $z''(s) > y''(s)$ when $s^2 > U(c) - U(a_+)$. Finally, P_+ is an increasing function. □

Remark. Assumption (1.12) is equivalent to

$$5g''^2(r) - 3g^{(3)}(r)(g'(r) - g'(c)) \geq 0 \quad \forall r \in (c, +\infty), \tag{1.18}$$

where $c = g'^{-1}(\mu)$, and in fact it must be assumed only on (c, b_+). Many common functions satisfy (1.18): not only $g(r) = |r|^{q-1}r (q > 1)$ but also $g(r) = (\cosh r - 1)$ sgn $r, g(r) = r$ Log$(1 + |r|)$ and any function g such that $g^{(3)} \leq 0$ on $\mathbb{R}^+$.

Theorem 0.1 is a consequence of the following theorem. Here we use Lemmas 1.1 and 1.2, similar results for negative α, β, and the mean value theorem (notice that $\tilde{\omega}(., -\alpha)$ and $\tilde{\omega}(., \alpha)$ are in the same connected component, even if g is not odd).

THEOREM 1.1. *For any real μ, let $E_\mu = \{0\} \cup E_\mu^+ \cup E_\mu^- \cup \tilde{E}_\mu$ be the set of solutions of* (1.1) *on S^1, with assumption* (1.2); *E_μ^+ (respectively, E_μ^-) is the subset of positive (respectively, negative) ones, $\tilde{E}_\mu$ the subset of changing sign solutions. Then*

(i) *$\tilde{E}_\mu = \cup_{k=\tilde{k}(\mu)}^{+\infty} \tilde{C}_k$, where $\tilde{k}(\mu)$ is the smallest positive integer such that $k^2 + \mu > 0$;*

(ii) *If $\mu \leq 0, E_\mu^+ = E_\mu^- = \emptyset$. If $\mu > 0$, then*

$$E_\mu^\pm \supset \{a_\pm\} \cup \left(\bigcup_{k=1}^{k_\pm(\mu)} C_k^\pm \right),$$

where $k_\pm(\mu)$ is the largest integer smaller than $(g'(a_+) - \mu)^{1/2}$, and the inclusion is an equality under assumption (1.12);

(iii) *Each connected component $\tilde{C}_k^+, \tilde{C}_k^-, C_k^+, C_k^-$ is one-dimensional, generated by rotation of a function with least period $2\pi/k$.*

Let us now consider the energy function associated to any $\omega \in E_\mu$:

$$\mathbf{E}(\omega) = \int_{S^1} \left(\frac{\omega'^2}{2} + \mu\frac{\omega^2}{2} - G(\omega) \right) d\theta = \frac{\pi}{P} \int_0^P (\omega'^2 - U(\omega)) d\theta, \tag{1.19}$$

where P is the least period of ω.

Since $\omega'^2 + U(\omega) = (\omega'(0))^2 + U(\omega(0))$, we get

$$\mathbf{E}(\omega) = \frac{\pi}{P}\int_0^P (2\omega'^2 - (\omega'(0))^2 - U(\omega(0)))d\theta. \tag{1.20}$$

By multiplying (1.1) by ω and integrating it on P, we also have

$$\mathbf{E}(\omega) = \frac{\pi}{P}\int_0^P (\omega g(\omega) - 2G(\omega))d\theta. \tag{1.21}$$

More generally, we shall study the function $\mathbf{E}$ defined by (1.20), (1.21) for any periodic solution ω of (1.1) on whole $\mathbb{R}$. Let us notice that $\mathbf{E}$ is nonnegative, since from (1.2) the function $r \mapsto H(r) = r\,g(r) - G(r)$ is convex on $\mathbb{R}$, with $H(0) = H'(0) = 0$.

First we consider the case of the changing sign solutions of (1.1).

LEMMA 1.3. *Under the assumptions of Lemma* 1.1, *the function* $\alpha \mapsto \mathbf{E}(\tilde{\omega}(.,\alpha))$ *is increasing from* $(0,+\infty)$ *to* $(0,+\infty)$.

Proof. With the notation of Lemma 1.1, we get

$$\mathbf{E}(\tilde{\omega}(.,\alpha)) = \pi\tilde{\phi}(\alpha)/\tilde{P}(\alpha) \tag{1.22}$$

from (1.20), where, for any $\alpha > 0$,

$$\tilde{\phi}(\alpha) = 2\int_0^{\tilde{P}(\alpha)} (\tilde{\omega}'(\theta,\alpha))^2 d\theta - \alpha^2\tilde{P}(\alpha). \tag{1.23}$$

Hence

$$\tilde{\phi}(\alpha) = 4\int_{\tilde{y}(\alpha)}^{\tilde{z}(\alpha)} \sqrt{\alpha^2 - U(v)}dv - \alpha^2\tilde{P}(\alpha). \tag{1.24}$$

Let

$$\tilde{I}(\alpha) = \int_0^1 (\tilde{z}' - \tilde{y}')(\alpha t)\frac{dt}{\sqrt{1-t^2}}, \tag{1.25}$$

$$\tilde{J}(\alpha) = \int_0^1 (\tilde{z}' - \tilde{y}')(\alpha t)\sqrt{1-t^2}dt. \tag{1.26}$$

Suppose first that $\mu > 0$. Then from (1.10) we know that

$$\tilde{P}(\alpha)/2 = \int_{b_-}^{b_+} \frac{dv}{\sqrt{\alpha^2 - U(v)}} + \tilde{I}(\alpha); \tag{1.27}$$

and from (1.23),

$$\begin{aligned}\tilde{\phi}(\alpha)/2 &= 2\int_{b_-}^{b_+} \sqrt{\alpha^2 - U(v)}dv + \alpha^2(2\tilde{J}(\alpha) - \tilde{P}(\alpha)/2) \\ &= \int_{b_-}^{b_+} \frac{\alpha^2 - 2U(v)}{\sqrt{\alpha^2 - U(v)}}dv + \alpha^2(2\tilde{J}(\alpha) - \tilde{I}(\alpha)).\end{aligned} \tag{1.28}$$

Now let us define

$$\tilde{K}(\alpha) = \int_0^1 (\tilde{z} - \tilde{y})(\alpha t) t \, \frac{dt}{\sqrt{1-t^2}}; \tag{1.29}$$

integrating by parts we get

$$\tilde{K}(\alpha) = \left[-(\tilde{z} - \tilde{y})(\alpha t)\sqrt{1-t^2} \right]_0^1 + \alpha \int_0^1 (\tilde{z}' - \tilde{y}')(\alpha t)\sqrt{1-t^2} dt,$$

which means

$$\tilde{K}(\alpha) = b_- - b_+ + \alpha \tilde{J}(\alpha). \tag{1.30}$$

By differentiation we get

$$\tilde{K}'(\alpha) + \tilde{J}(\alpha) = \int_0^1 (\tilde{z}' - \tilde{y}')(\alpha t) \left(\frac{t^2}{\sqrt{1-t^2}} + \sqrt{1-t^2} \right) dt.$$

In other terms

$$\alpha \tilde{J}'(\alpha) + 2\tilde{J}(\alpha) = \tilde{I}(\alpha). \tag{1.31}$$

From (1.28) and (1.31) we deduce

$$\tilde{\phi}(\alpha)/2 = \int_{b_-}^{b_+} \frac{\alpha^2 - 2U(v)}{\sqrt{\alpha^2 - U(v)}} dv - \alpha^3 \tilde{J}'(\alpha), \tag{1.32}$$

and we get by differentiation of (1.31) and (1.32) the relation

$$\tilde{\phi}'(\alpha)/2 = \alpha^3 \int_{b_-}^{b_+} \frac{dv}{(\alpha^2 - U(v))^{3/2}} - \alpha^2 \tilde{I}'(\alpha). \tag{1.33}$$

Now by differentiating (1.27) we get the relation

$$\tilde{\phi}'(\alpha) = -\alpha^2 \tilde{P}'(\alpha) \quad \forall \alpha > 0. \tag{1.34}$$

From Lemma 1.1, $\tilde{P}$ is a decreasing function; hence $\tilde{\phi}$ is a positive increasing function on $(0, +\infty)$, since $\mathbf{E}$ is nonnegative. Then $\alpha \mapsto \mathbf{E}(\tilde{\omega}(.,\alpha))$ is increasing on $(0, +\infty)$. Suppose now that $\mu \leq 0$; then from (1.9), the proof is the same with b_-, b_+ replaced by 0.

Let us look at the limits of $\mathbf{E}$ near zero and infinity. When $\mu \geq 0$, we have $\lim_{\alpha \to 0} \tilde{P}(\alpha) = +\infty$; the function $\tilde{\phi}$ has a finite nonnegative limit at 0, hence $\lim_{\alpha \to 0} \mathbf{E}(\tilde{\omega}(.,\alpha)) = 0$. When $\mu < 0$, from Lemma 1.1 we have $\lim_{\alpha \to 0} \tilde{I}(\alpha) = \pi/\sqrt{-\mu}$, we also get $\lim_{\alpha \to 0} \tilde{J}(\alpha) = \pi/2\sqrt{-\mu}$; since $\tilde{\phi}(\alpha) = 2\alpha^2(2\tilde{J}(\alpha) - \tilde{I}(\alpha))$ and $\tilde{P}(\alpha) = 2\tilde{I}(\alpha)$ we get more precisely $\lim_{\alpha \to 0} \alpha^{-2}\mathbf{E}(\tilde{\omega}(.,\alpha)) = 0$. For any real μ, we have $\lim_{\alpha \to +\infty} \tilde{P}(\alpha) = 0$, and $\tilde{\phi}$ is increasing; hence $\lim_{\alpha \to +\infty} \mathbf{E}(\tilde{\omega}(.,\alpha)) = +\infty$. □

Remark. When $\mu = 0$ we can do explicit computation and easily get the following properties:

$$\tilde{\omega}(\theta, \alpha) = \alpha^{2/(q+1)} \tilde{\omega}(1, \alpha^{(q-1)/(q+1)}\theta) \quad \forall (\theta, \alpha) \in \mathbb{R} \times \mathbb{R}^+,$$
$$\tilde{P}(\alpha) = \alpha^{-(q-1)/(q+1)} \tilde{P}(1),$$
$$E(\tilde{\omega}(.,\alpha)) = \alpha^2 E(\tilde{\omega}(.,1)),$$

which give a new proof of the monotonicity of the period and energy functions.

Let us now consider the case of the positive solutions of (1.1).

LEMMA 1.4. *Under the assumptions of Lemma* 1.2 *with condition* (1.12), *the function* $\beta \mapsto \mathbf{E}(\omega_+(.,\beta))$ *is decreasing from* $(0, \sqrt{-U(a_+)})$ *to* $(0, \mathbf{E}(a_+)), \mathbf{E}(a_+) = -\pi U(a_+)$.

Proof. With the notation of Lemma 1.2 we now get

$$\mathbf{E}(\omega_+(.,\beta)) = \pi \frac{\phi(\beta)}{P_+(\beta)} - \pi U(a_+), \tag{1.35}$$

where, for any $\beta \in (0, \sqrt{-U(a_+)})$,

$$\phi(\beta) = 4 \int_{y(\beta)}^{z(\beta)} \sqrt{\alpha^2 - U(v) + U(a_+)} dv - \beta^2 P_+(\beta). \tag{1.36}$$

As above, let

$$I(\beta) = \int_0^1 (z' - y')(\beta t) \frac{dt}{\sqrt{1-t^2}} = P_+(\beta)/2, \tag{1.37}$$

$$J(\beta) = \int_0^1 (z' - y')(\beta t) \sqrt{1-t^2} dt, \tag{1.38}$$

$$K(\beta) = \int_0^1 (z - y)(\beta t) t \frac{dt}{\sqrt{1-t^2}}; \tag{1.39}$$

then

$$\phi(\beta)/2 = \beta^2 (2J(\beta) - I(\beta)). \tag{1.40}$$

We again get the relations $K(\beta) = \beta J(\beta)$, $\beta J'(\beta) + 2J(\beta) = I(\beta)$, and then

$$\phi'(\beta) = -\beta^2 P'_+(\beta) \quad \forall \beta \in (0, \sqrt{-U(a_+)}). \tag{1.41}$$

Here we cannot take into account the sign of ϕ, so we differentiate (1.35) and get after simplification, from (1.37) and (1.40),

$$(\phi/P_+)'(\beta) = -4\beta^2 J(\beta) P'_+(\beta)/P_+^2(\beta). \tag{1.42}$$

From Lemma 1.2, P_+ is increasing under condition (1.12); since J is a positive function we deduce that $\beta \mapsto \mathbf{E}(\omega_+(.,\beta))$ is decreasing on $(0, \sqrt{-U(a_+)})$.

Moreover, we have $\lim_{\beta \to 0} I(\beta) = \pi / \sqrt{g'(a_+) - \mu}$ and also get $\lim_{\beta \to 0} J(\beta) = \pi / 2\sqrt{g'(a_+) - \mu}$; then from (1.35) and (1.40), $\lim_{\beta \to 0} \mathbf{E}(\omega_+(.,\beta)) = -\pi U(a_+) = \mathbf{E}(a_+)$, which is the energy of the constant solution $\omega_o \equiv a_+$. More precisely, we have $\lim_{\beta \to 0} \beta^{-2}(\mathbf{E}(\omega_+(.,\beta)) - \mathbf{E}(a_+)) = 0$. On the other hand, $\lim_{\beta \to \sqrt{-U(a_+)}} P_+(\beta) = +\infty$. From (1.21) we have also

$$P_+(\beta)\mathbf{E}(\omega_+(.,\beta))/2\pi = \int_0^1 ((zg(z) - 2G(z))z' - (yg(y) - 2G(y))y')(\beta t) \frac{dt}{\sqrt{1-t^2}}; \tag{1.43}$$

we easily get

$$\lim_{s\to 0}((zg(z)-2G(z))z'-(yg(y)-2G(y))y')(s)=2(a_+g(a_+)-G(a_+))/\sqrt{g'(a_+)-\mu},$$

$$\lim_{s\to\sqrt{-U(a_+)}}((zg(z)-2G(z))z')(s)=\sqrt{-U(a_+)}(b_+g(b_+)-2G(b_+))/(g(b_+)-\mu b_+),$$

$$\lim_{s\to\sqrt{-U(a_+)}}((yg(y)-2G(y))y')(s)=\sqrt{-U(a_+)}\lim_{y\to 0}\frac{yg(y)-2G(y)}{g(y)-\mu y}=0$$

from (1.2); hence the integral (1.43) remains bound near $\sqrt{-U(a_+)}$; then $\lim_{\beta\to\sqrt{-U(a_+)}}$ $\mathbf{E}(\omega_+(.,\beta))=0$. □

Now we can classify the energy of the solutions by the connected components to which they belong.

COROLLARY 1.1. *Under the assumptions of Theorem* 1.1, *for any function* g *satisfying* (1.2), *the function* $k\mapsto \mathbf{E}(\tilde{\omega}_k)$ *is increasing from* $\mathbb{N}\cap[\tilde{k}(\mu),+\infty)$ *to* $(0,+\infty)$. *When* g *satisfies* (1.12), *and* $g'(a_+)-\mu>1$, *the function* $k\mapsto \mathbf{E}(\omega_{+k})$ *is increasing from* $\mathbb{N}\cap[1,k_+(\mu)]$ *to* $(0,\mathbf{E}(a_+))$.

Remark. When $g(r)=|r|^{q-1}r, q>1$, the condition $g'(a_+)-\mu>1$ is equivalent to $\mu>1/(q-1)$.

Remark. When $g(r)=r^3$ or $g(r)=|r|r$, J. R. Licois has pointed out that the period and energy functions can be expressed in terms of elliptic integrals.

When $g(r)=r^3$, all the periodic solutions of (1.1) are given explicitly: denoting by $sn(.,k), cn(.,k), dn(.,k)$ with $k\in(0,1)$ the classical Jacobian elliptic functions (see [31]), we get the following: the positive nonconstant solutions (when $\mu>0$) are given by

$$\theta\mapsto\omega_+(\theta)=\sqrt{\frac{2\mu}{2-k^2}}dn\left(\sqrt{\frac{\mu}{2-k^2}}(\theta-\theta_o),k\right),\qquad \theta_o\in\mathbb{R}, k\in(0,1).$$

The changing sign solutions with $\mu\neq 0$ are given by

$$\theta\mapsto\tilde{\omega}(\theta)=\sqrt{\frac{2\mu k^2}{2k^2-1}}cn\left(\sqrt{\frac{\mu}{2k^2-1}}(\theta-\theta_o),k\right),\qquad \theta_o\in\mathbb{R},$$
$$k\in(1/\sqrt{2},1)\quad \text{if } \mu>0, k\in(0,1/\sqrt{2})\quad \text{if } \mu<0.$$

With those formulas we can find again the monotonicity of the period function of changing sign solutions by differentiating with respect to the parameter k; this kind of proof does not work for positive solutions.

When $\mu=0$ we get the functions

$$\theta\mapsto\tilde{\omega}(\theta)=p\ cn(p(\theta-\theta_o),1/2),\qquad \theta_o\in\mathbb{R},\ p>0.$$

2. Application to the parabolic problem. Here we apply results of §1 to the parabolic equation

$$u_t=u_{xx}+|u|^{q-1}u-\mu u, \tag{2.1}$$

where $q>1$, with periodic conditions on $(-\pi,+\pi)$,

$$u(t)(-\pi)=u(t)(\pi),\qquad u_t(t)(-\pi)=u_t(t)(\pi),\quad \text{with } \mu>0; \tag{2.2}$$

or Neumann conditions on $(0,\pi)$,

$$u_x(t)(0)=u_x(t)(\pi)=0,\quad \text{with } \mu>0; \tag{2.3}$$

or Dirichlet conditions on $(0,\pi)$,

$$u(t)(0) = u(t)(\pi) = 0, \quad \text{with } \mu \geq 0. \tag{2.4}$$

COROLLARY 2.1. *Let u be any (smooth enough) solution of* (2.1), (2.2) *or* (2.1), (2.3) *on* $[0,+\infty)$; *then u converges at infinity to one of the solutions of* (0.1): *either to a changing sign function $\tilde{\omega}_k$, with least period $2\pi/k, k \in \mathbb{N}/\{0\}$; or to* 0 *or* $\pm\mu^{1/(q+1)}$; *or to a function $\pm\omega_{+k}$ with least period $2\pi/k$, with $k^2 < (q-1)\mu$, when $\mu > 1/(q-1)$. Any solution of* (2.1), (2.4) *converges to* 0 *or to a function $\tilde{\omega}_k, k \in \mathbb{N}/\{0\}$.*

Proof. From Cazenave and Lions [11] we know that any solution of (2.1), (2.3) defined on $[0,+\infty)$ is bounded: $\sup_{t>0} \|u(t)\|_{L^\infty} < +\infty$; the proof is the same for Neumann or periodic conditions, since $\mu > 0$. Hence from Matano [21] it converges precisely to one of the solutions of the stationary problem (0.1), and we apply Theorem 0.1. □

Remark. Now consider the global problem for (2.1): let u be any solution of (2.1), (2.2) (or similarly (2.3), (2.4)), bounded at $-\infty$; then u converges at $+\infty$ (respectively, $-\infty$) towards a solution $\omega_{+\infty}$ (respectively, $\omega_{-\infty}$) of (0.1). The energy relation of the problem is

$$\frac{d}{dt}\mathbf{E}(u(t)) = -\int_{-\pi}^{+\pi} u_t^2 \, dx, \tag{2.5}$$

where $\mathbf{E}$ has been defined in (0.10); hence

$$\mathbf{E}(\omega_{+\infty}) < \mathbf{E}(\omega_{-\infty}), \quad \text{or } u \equiv \omega_+ \equiv \omega_-. \tag{2.6}$$

From Corollary 1.1 we see that if $\omega_{+\infty}$ and $\omega_{-\infty}$ both have a constant sign (or both a changing sign), then $\omega_{-\infty}$ has more points of extremum than $\omega_{+\infty}$. This is conformal to the results of Matano [22]. Consider in particular any positive nonconstant global solution; then $\omega_{+\infty}$ is nonconstant or identically 0, and $\omega_{-\infty} = \mu^{1/(q-1)}$, $\omega_{+\infty} = 0$ if $\mu < 1/(q-1)$.

The problem of the existence of connecting orbits is still open, even for positive functions: for given solutions $\omega_{+\infty}$, $\omega_{-\infty}$ of (0.1) such that $\mathbf{E}(\omega_{+\infty}) < \mathbf{E}(\omega_{-\infty})$, does there exist a global solution u of (2.1), (2.2) with $\lim_{t\to\pm\infty} u(t) = \omega_{\pm\infty}$?

3. Application to the elliptic problem. Here we apply our results to find the behavior near the origin or near infinity of the solutions of the elliptic equation in $\mathbb{R}^2/\{0\}$:

$$\Delta u - c\frac{u}{|x|^2} + |u|^{q-1}u = 0, \tag{3.1}$$

where $c, q \in \mathbb{R}$, $c > 0$, and $q > 1$.

The first question for such an equation is to find an a priori estimate in $|x|^{-\delta}(\delta = 2/(q-1))$. This problem is quite difficult and not completely solved in N-space dimension with $N \geq 3$ (see [5], [10], [16]) and still open for changing sign solutions. Nevertheless in the two-dimensional case Bouhar and Veron [6] give a nice proof for the case of positive solutions, using semigroup techniques for the function v defined by (0.4). Set $B_1 = \{x \in \mathbb{R}^2 |\ |x| < 1\}$. Then we have the following.

THEOREM 3.1. *Let u be any (smooth enough) positive solution of* (3.1) *in $B_1/\{0\}$ (respectively, in $\mathbb{R}^2/\overline{B}_1$). Then u satisfies the estimate*

$$|x|^\delta u \in L^\infty_{\text{loc}}(B_1) \quad (\textit{respectively, } L^\infty_{\text{loc}}(\mathbb{R}^2/\overline{B}_1)). \tag{3.2}$$

Here we give another proof of this result, which works when q is not too large.

Proof in the case $1 < q < 2(1+\sqrt{3})^2$. We follow the proof of the estimates given by Bidaut-Veron and Veron in [5] when $N \geq 3$, obtained by using Bochner–Lichnerowicz formula. It lies on Lemma 6.2 of [5], which asserts that when $N \geq 3$ and $1 < q < (N+2)/(N-2)$, there exist $d \in \mathbb{R}$ and $y \in \mathbb{R}/\{1\}$ such that

$$\begin{cases} -2\dfrac{N-1}{N}y^2 + 2dy - d^2 + d > 0, \qquad 2(N-1)q/(N+2) < d; \\ d + 2q - 2y > N(q-1)/2, \quad d + 2 - 2y > 0, \quad d + q - 2y > 0. \end{cases} \tag{3.3}$$

When $N = 2$ and $q < 2(1+\sqrt{3})^2$ such conditions are still satisfied; hence we get the estimate near the origin as in [5, Thm. 6.3], and in a similar way at infinity. □

Remark. By this way we get a majorization of $\sup_{0<|x|<1/2} |x|^{\delta}u(x)$ by a constant C that depends on q and c, but does not depend on u. In the proof of [6] for general q, C may depend on u.

Now we study the question of convergence for the function v defined by (0.4) as in [5].

THEOREM 3.2. *Let u be any solution of* (3.1) *in* $B_1/\{0\}$ (*respectively,* $\mathbb{R}^2/\overline{B}_1$). *If u satisfies* (3.2), *then* $|x|^{\delta}u(|x|,.)$ *converges in* $C^3(S^1)$ *to some connected component of the set* E_{δ^2-c} *of the solutions of the equation on* S^1:

$$\omega_{\theta\theta} - (c - \delta^2)\,\omega + |\omega|^{q-1}\omega = 0. \tag{3.4}$$

Moreover, $|x|^{\delta}u(|x|,.)$ *converges precisely to one element of* E_{δ^2-c} *when u is positive, or when q is an odd integer.*

Proof. Under assumption (3.2), the function v defined on $(0,+\infty)\times S^1$ (respectively, $(-\infty,0)\times S^1$) is bounded on $[1,+\infty)\times S^1$ (respectively, $(-\infty,-1]\times S^1$); hence with elliptic equations theory, the orbit of v is relatively compact in $C^3(S^1)$. As in [5] and [13], the ω-limit set (respectively, the α-limit set) is compact, connected, and contained in the set E_{δ^2-c} of the solutions of the stationary problem. When u is positive, or q is odd, from Simon's analyticity results [27] we deduce as in [5, Thms. 3.2, 5.1] that v converges precisely to some $\omega \in E_{\delta^2-c}$ in $C^3(S^1)$. □

If we look for asymptotics of the positive solutions of (3.1), the richest case is $c > \delta^2$: then (3.4) admits a positive constant solution. Hence

$$u(x) = \lambda|x|^{-\delta}, \quad \text{with } \lambda = (c-\delta^2)^{1/(q-1)} \tag{3.5}$$

is a solution of (3.1); the nonlinear effect of the power ω^q and the linear effect of the Laplacian interfere. Notice that the linear equation associated to (3.1),

$$\Delta\psi - c\frac{\psi}{|x|^2} = 0, \tag{3.6}$$

admits two linear independent positive radial solutions: $|x|^{\pm\sqrt{c}}$.

Our result in this supercritical case is the following.

THEOREM 3.3. *Assume that* $c > \delta^2 = (2/(q-1))^2$ *and u is a positive solution of* (3.1) *in* $B_1/\{0\}$ (*respectively,* $\mathbb{R}^2/\overline{B}_1$). *Then* (i) *either there is a* $\gamma > 0$ (*respectively,* $\overline{\gamma} > 0$) *such that*

$$\lim_{x\to 0} |x|^{-\sqrt{c}}u(x) = \gamma \quad \left(\textit{respectively,} \lim_{|x|\to+\infty} |x|^{\sqrt{c}}u(x) = \overline{\gamma}\right); \tag{3.7}$$

(ii) *or there is a positive solution* ω *(respectively,* $\overline{\omega}$*) of* (3.4) *on* S^1 *such that*

$$\lim_{x\to 0} |x|^{\delta} u(|x|,.) = \omega(.) \quad (\textit{respectively}, \lim_{|x|\to+\infty} |x|^{\delta} u(|x|,.) = \overline{\omega}\ (.)) \tag{3.8}$$

in $C^3(S^1)$;

(iii) *if* $\delta^2 < c \le \delta^2 + 1/(q-1)$, *then*

$$\omega \equiv \lambda = (c-\delta^2)^{1/(q-1)} \quad (\textit{respectively, } \overline{\omega} \equiv \lambda); \tag{3.9}$$

(iv) *if* $c > \delta^2 + 1/(q-1)$, *then* $\omega \equiv \lambda$ *(respectively,* $\overline{\omega} \equiv \lambda$*) or* ω *(respectively,* $\overline{\omega}$*) oscillates around* λ *with least period* $2\pi/k$ *for some* $k < ((q-1)(c-\delta^2))^{1/2}$.

Proof. From Theorems 3.1 and 3.2 above, $|x|^{\delta}u(|x|,.)$ converges to some nonnegative solution of (3.4); we get (i) if it is zero, (ii) if not, from Theorems 3.2 and 3.3 of [5], adapted to the case $N = 2$. Then we apply Theorem 0.1 with $\mu = c - \delta^2$ to get (iii) and (iv). □

Remark. There do exist solutions u of (3.1) and (3.7), for example, radial ones, since $(0,0)$ is a saddle point of the linearized equation $w_{tt} + 2\delta w_t - (c-\delta^2)w = 0$; and for any $\gamma, \overline{\gamma} > 0$ by scaling. On the contrary, all the functions u satisfying (3.8) at $+\infty$ are necessarily nonradial but $u(x) \equiv \lambda|x|^{-\delta}$, since $(\lambda, 0)$ is totally instable for the linearized equation $y_{tt} + 2\delta y_t + (q-1)\mu y = 0$.

Otherwise, when $c > \delta^2 + 1/(q-1)$, the nonradial functions $x \to |x|^{-\delta}\omega(.), \omega \in E_{c-\delta^2}/\{\lambda\}$ obviously give $k_+(c-\delta^2)$ circles of solutions of (3.1).

The study of the energy of the solutions of (3.4), developed in §1, allows us to select the possible connections for global positive solutions of (3.1).

THEOREM 3.4. *Assume that* $c > \delta^2$ *and* u *is a positive solution of* (3.1) *in* $\mathbb{R}^2/\{0\}$. *Then* (i) *either* u *is singular at* 0 *and regular at infinity:*

$$\lim_{x\to 0} |x|^{\delta} u(|x|,.) = \omega(.) \quad \textit{and} \quad \lim_{|x|\to+\infty} |x|^{\sqrt{c}} u(x) = \overline{\gamma} \tag{3.10}$$

for some positive solution ω *of* (3.4) *and some* $\overline{\gamma} > 0$;

(ii) *or* u *is singular at* 0 *and infinity:*

$$\lim_{x\to 0} |x|^{\delta} u(|x|,.) = \omega(.) \quad \lim_{|x|\to+\infty} |x|^{\delta} u(|x|,.) = \overline{\omega}(.), \tag{3.11}$$

where $\omega, \overline{\omega} > 0$ *satisfy* (3.4). *Henceforth, if* $c < \delta^2 \le \delta^2 + 1/(q-1)$, *then* $u(x) \equiv \lambda|x|^{-\delta}$.

If $c > \delta^2 + 1/(q-1)$, *then*

— *either* $u(x) \equiv \lambda|x|^{-\delta}$;

— *or* $\omega \equiv \lambda$ *and* $\overline{\omega}$ *has a least period* $2\pi/\overline{k}$, *with* $\overline{k} < ((q-1)(c-\delta^2))^{1/2}$;

— *or* ω *and* $\overline{\omega}$ *have least periods* $2\pi/k$, $2\pi/\overline{k}$, *with* $\overline{k} \le k < ((q-1)(c-\delta^2))^{1/2}$;

and if $k = \overline{k}$ *then* $\omega \equiv \overline{\omega}$ *and* $u(x) \equiv |x|^{-\delta}\omega(.)$.

Proof. From Theorems 3.1 and 3.2 there are two functions $\omega, \overline{\omega}$ in E_{δ^2-c} such that $\lim_{t\to+\infty} v(t,.) = \omega$ and $\lim_{t\to-\infty} v(t,.) = \overline{\omega}$ in $C^3(S^1)$, where v is defined by (0.4). From (0.6) the energy relation for function v is the following:

$$\frac{d}{dt}\left(\mathbf{E}(v(t)) - \frac{1}{2}\int_{S^1} v_t^2(t)\, d\theta\right) = 2\delta \int_{S^1} v_t^2(t)\, d\theta, \tag{3.12}$$

where $\mathbf{E}$ is defined in (0.10) with $\mu = c - \delta^2$; hence

$$\mathbf{E}(\omega) > \mathbf{E}(\overline{\omega}) \quad \text{or} \quad v \equiv \omega \equiv \overline{\omega}; \tag{3.13}$$

we obtain the conclusions with Theorem 0.1 and Corollary 1.1. □

Remark. As for the parabolic problem, we can extend a part of the results of Theorems 3.2 and 3.3 to any solution of (3.1) satisfying the boundedness condition (3.2) by using Corollary 1.1 for changing sign solutions.

Now we briefly give the asymptotics in the critical or undercritical case for positive solutions. Then only linear singularities appear because $E^+_{c-\delta^2} = \emptyset$ from Theorem 0.1.

First we notice that no positive solutions of (3.1) can exist in the case of the exterior problem, as in the case $c = 0$; see [4] and [24].

PROPOSITION 3.1. *Assume* $0 < c \leq \delta^2$. *Then any nonnegative solution* u *of* (3.1) *in* $\mathbb{R}^2/\overline{B}_1$ *is identically zero.*

Proof. Suppose that v defined by (0.4) is nonnegative on $(-\infty, 0) \times S^1$. Then the mean value function $\overline{v}(t) = \int_{S^1} v(t,\theta)\, d\theta$ satisfies the inequality

$$\overline{v}_{tt} + 2\delta\overline{v}_t + (\delta^2 - c)\overline{v} + \overline{v}^q \leq 0;$$

let $\overline{v}(t) = \eta(s)$, with $s = e^{-2\delta t}$. Then

$$4\delta^2 s^2 \eta_{ss} + (\delta^2 - c)\eta + \eta^q \leq 0,$$

and hence $4\delta^2 s^2 \eta_{ss} + \eta^q \leq 0$. If η is nonidentically 0, then for some $s_o > 1$ it is positive, concave, and nondecreasing on $(s_o, +\infty)$. For any $\sigma \geq s \geq s_o$ we get

$$4\delta^2 \eta_s(\sigma) \leq 4\delta^2 \eta_s(s) - \int_s^\sigma \eta^q t^{-2} dt \leq 4\delta^2 \eta_s(s) - \frac{\eta^q(s)}{s} + \frac{\eta^q(s)}{\sigma}.$$

If, for some $s \geq s_o$, $4\delta^2\eta_s(s) - \eta^q(s)/s < 0$, then $\eta_s(\sigma) < 0$ for large σ; by contradiction we have $4\delta^2\eta_s(s) - \eta^q(s)/s \geq 0$ for any $s \geq s_o$, and $s \mapsto \mathrm{Log}\, s + 4\delta^2 z^{1-q}(s)/(q-1)$ is nonincreasing, which is impossible. □

Therefore we are reduced to study the behavior near the origin.

THEOREM 3.5. *Assume that* $c = \delta^2$ *and* u *is a positive solution of* (3.1) *in* $B_1/\{0\}$. *Then*

(i) *either there is a* $\gamma > 0$ *such that*

$$\lim_{x \to 0} |x|^{-\delta} u(x) = \gamma; \tag{3.14}$$

(ii) *or*

$$\lim_{x \to 0} |x|^{\delta} (-\mathrm{Log}|x|)^{\delta/2} u(x) = \delta^\delta. \tag{3.15}$$

Proof. From [5, Cor. 6.5], we have the estimate

$$|x|^{\delta} (-\mathrm{Log}|x|)^{\delta/2} u(x) \in L^\infty_{\mathrm{loc}}(B_1), \tag{3.16}$$

obtained from (3.2) and Harnack inequality. We make the transformation

$$u(r,\theta) = (-\mathrm{Log}\, r)^{-\delta/2} r^{-\delta} \zeta(t,\theta), \quad t = -\mathrm{Log}\, r, \quad r > 0, \quad \theta \in S^1; \tag{3.17}$$

we get

$$\zeta_{tt} + \delta\left(2 - \frac{1}{t}\right)\zeta_t + \zeta_{\theta\theta} + \frac{q}{(q-1)^2}\frac{\zeta}{t^2} - \delta^2\frac{\zeta}{t} + \frac{\zeta^q}{t} = 0. \tag{3.18}$$

Then, following the techniques used by Aviles [3] for the equation $\Delta u + u^{N/(N-2)} = 0$ with $N \geq 3$, we prove that

$$\lim_{t\to+\infty}\zeta(t)=l \quad \text{with } l=0 \text{ or } \delta^{\delta}. \tag{3.19}$$

When $l\neq 0$ we get (3.15). Now suppose $l=0$; then for any $\varepsilon\in(0,1]$ there is a $r(\varepsilon)\in(0,\varepsilon)$ such that

$$-\Delta u+(\delta^2-\varepsilon|\mathrm{Log}\ r|^{-1})r^{-2}u\leq 0 \quad \text{when } r=|x|\leq r(\varepsilon). \tag{3.20}$$

Let us look at radial solutions $\phi(r)$ of the equation

$$-\Delta\phi+(\delta^2-|\mathrm{Log}\ r|^{-1})r^2\phi=0; \tag{3.21}$$

define $f(t)=\phi(r)$, with $t=-\mathrm{Log}\ r$. Then f satisfies

$$f_{tt}(t)-\delta^2(1-t^{-1})f(t)=0. \tag{3.22}$$

From [7], (3.22) has two independent solutions f_1, f_2 such that $\lim_{t\to+\infty}e^{-\delta t}t^{\delta/2}f_1(t)=1$ and $\lim_{t\to+\infty}e^{\delta t}t^{-\delta/2}f_2(t)=1$, and (3.21) has two corresponding solutions ψ_1,ψ_2 such that $\lim_{r\to 0}r^{\delta}(-\mathrm{Log}\ r)^{\delta/2}\psi_1(r)=1$ and $\lim_{r\to 0}r^{-\delta}(-\mathrm{Log}\ r)^{-\delta/2}\psi_2(r)=1$. From (3.19) there is a $\overline{r}(\varepsilon)\in(0,r(\varepsilon))$ such that $u\leq\varepsilon\psi_1$ when $r\leq\overline{r}(\varepsilon)$; from the maximum principle we get $u\leq\varepsilon\psi_1+a\psi_2$ when $\overline{r}(\varepsilon)\leq r\leq r(1)$, where $a=(\max_{|x|=r(1)}u(x)/\psi_2(r(1)))$. Then we get the estimate

$$|x|^{-\delta}(-\mathrm{Log}\ |x|)^{-\delta/2}u(x)\in L^{\infty}_{\mathrm{loc}}(B_1); \tag{3.23}$$

in particular, the singularity is removable. Moreover, the function $v(t)$ defined by (0.4) decreases exponentially in $C^o(S^1)$ near $+\infty$; hence its behavior is essentially of linear type. Using Fourier techniques as in [13] we derive the estimate

$$|x|^{-\delta}u(x)\in L^{\infty}_{\mathrm{loc}}(B_1); \tag{3.24}$$

then easily $\lim_{x\to 0}|x|^{-\delta}u(x)=\gamma\geq 0$. If $\gamma=0$ we get $u\equiv 0$ from Aronszajn's unique continuation theorem [2]; hence a contradiction. □

Remark. There do exist radial solutions satisfying (3.14) for any $\gamma>0$, using fixed point method for the function $r^{-\delta}u$ as in [17]. Moreover there do exist radial solutions satisfying (3.15), as in [4, Thm. 6.8].

THEOREM 3.6. *Assume that* $0<c<\delta^2$ *and* u *is a positive solution of* (3.1) *in* $B_1/\{0\}$. *Then*

(i) *either there is a* $\gamma>0$ *such that*

$$\lim_{x\to 0}|x|^{\sqrt{c}}u(x)=\gamma; \tag{3.25}$$

(ii) *or there is a* $\rho>0$ *such that*

$$\lim_{x\to 0}|x|^{-\sqrt{c}}u(x)=\rho. \tag{3.26}$$

Proof. From [5, Cor. 6.5], we get the estimate

$$|x|^{\sqrt{c}}u(x)\in L^{\infty}_{\mathrm{loc}}(B_1); \tag{3.27}$$

then $v(t)$ decreases exponentially in $C^o(S^1)$ near $+\infty$; we conclude in two steps as in [13]. □

Remark. Obviously there do exist radial solutions satisfying (3.25) or (3.26), since $(0,0)$ is a source of the linearized equation of (0.5).

REFERENCES

[1] D. ARONSON, M. G. CRANDALL, AND L. A. PELETIER, *Stabilization of solutions of a degenerate nonlinear diffusion problem*, Nonlinear Anal., 6 (1982), pp. 1001–1022.

[2] R. ARONSZAJN, *A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of second order*, J. Math. Pures Appl., 36 (1957), pp. 235–249.

[3] P. AVILES, *Local behaviour of solutions of some elliptic equations*, Comm. Math. Phys., 108 (1987), pp. 177–192.

[4] M. F. BIDAUT-VERON, *Local and global behavior of solutions of quasilinear equations of Emden–Fowler type*, Arch. Rational Mech. Anal., 107 (1989), pp. 293–324.

[5] M. F. BIDAUT-VERON AND L. VERON, *Nonlinear elliptic equations on compact Riemannian manifolds and asymptotics of Emden equations*, Invent. Math., 106 (1991), pp. 489–539.

[6] M. BOUHAR AND L. VERON, *Integral representations of solutions of semilinear elliptic systems in cylinders and applications*, Nonlinear Anal., to appear.

[7] R. BELLMAN, *Stability Theory of Differential Equations*, McGraw–Hill, New York, 1953.

[8] H. BREZIS AND P. L. LIONS, *A note on isolated singularities for linear elliptic equations*, J. Math. Anal. Appl., 7A (1981), pp. 263–266.

[9] P. BRUNOVSKI AND B. FIEDLER, *Heteroclinic connections of stationary solutions of scalar reaction diffusion equations*, Banach Center, P. W. N. Warsaw, 19 (1987), pp. 39–47.

[10] L. A. CAFFARELLI, B. GIDAS, AND J. SPRUCK, *Asymptotic symmetry and local behavior of semilinear elliptic equations with critical Sobolev growth*, Comm. Pure Appl. Math., 42 (1989), pp. 271–297.

[11] T. CAZENAVE AND P. L. LIONS, *Solutions globales d'équations de la chaleur semi-linéaires*, Comm. Partial Differential Equations, 9 (1984), pp. 955–978.

[12] N. CHAFEE AND E. F. INFANTE, *A bifurcation problem for a nonlinear partial differential equation of parabolic type*, Appl. Anal., 4 (1974), pp. 17–37.

[13] X. Y. CHEN, H. MATANO, AND L. VERON, *Anisotropic singularities of solutions of nonlinear elliptic equations in* $\mathbb{R}^2$, J. Functional Anal., 83 (1989), pp. 50–97.

[14] V. R. EMDEN, *Gaskugeln*, Teubner, Leipzig, Germany, 1897.

[15] R. H. FOWLER, *Further studies on Emden's and similar differential equations*, Q. J. Math., 2 (1931), pp. 259–288.

[16] B. GIDAS AND J. SPRUCK, *Global and local behavior of positive solutions of nonlinear elliptic equations*, Comm. Pure Appl. Math., 34 (1981), pp. 525–598.

[17] M. GUEDDA AND L. VERON, *Local and global properties of solutions of quasilinear equations*, J. Differential Equations, 76 (1988), pp. 159–189.

[18] B. GUERCH AND L. VERON, *Local properties of stationary solutions of some nonlinear singular Schrödinger equations*, Rev. Mat. Iberoamericana, 7 (1991), pp. 65–114.

[19] D. B. HENRY, *Some infinite-dimensional Morse–Smale systems defined by parabolic partial differential equations*, J. Differential Equations, 59 (1985), pp. 165–205.

[20] P. L. LIONS, *Isolated singularities in semilinear problems*, J. Differential Equations, 38 (1980), pp. 441–450.

[21] H. MATANO, *Convergence of solutions of one-dimensional semilinear parabolic equations*, J. Math. Kyoto Univ., 18 (1978), pp. 221–227.

[22] ———, *Existence of nontrivial unstable sets for equilibrium of strongly order-preserving systems*, J. Fac. Sci. Univ. Tokyo, 30 (1983), pp. 645–673.

[23] ———, *Nonlinear partial differential equations and infinite dimension dynamical systems*, Sûgaku, 42 (1990), pp. 289–303. (In Japanese.)

[24] W. M. NI AND J. SERRIN, *Existence and nonexistence theorems for ground states of quasilinear partial differential equations: the anomalous case*, Acad. Naz. dei Lincei, 77 (1986), pp. 231–257.

[25] J. SERRIN, *Local behavior of solutions of quasilinear equations*, Acta Math., 111 (1964), pp. 247–302.

[26] ———, *Isolated singularities of solutions of quasilinear equations*, Acta Math., 113 (1965), pp. 219–240.

[27] L. SIMON, *Asymptotics for a class of nonlinear evolution equations with applications to geometric problems*, Ann. Math., 118 (1983), pp. 525–571.

[28] ———, *Isolated singularities of extrema of geometric variational problems*, in *Harmonic Mappings and Minimal Immersions*, E. Giusti, ed., Lecture Notes in Math. 1161, Springer-Verlag, New York, 1985, pp. 206–277.

[29] L. VERON, *Comportement asymptotique des solutions d'équations elliptiques semi-linéaires dans* $\mathbb{R}^N$, Ann. Mat. Pura. Appl., 127 (1981), pp. 25–50.

[30] ———, *Singular solutions of some nonlinear elliptic equations*, Nonlinear Anal., 5 (1981), pp. 225–242.

[31] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, UK, 1973.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 876–893, May 1994

004

ON GLOBAL WEAK SOLUTIONS OF THE NONSTATIONARY TWO-PHASE STOKES FLOW*

YOSHIKAZU GIGA† AND SHUJI TAKAHASHI‡

Abstract. A global-in-time weak solution of the nonstationary two-phase Stokes flow is constructed for arbitrary given initial phase configuration (under periodic boundary condition) when two viscosities are close. The solution presented here tracks the evolution of the interface after it develops singularities. The theory of viscosity solutions is adapted to solve the interface equation. Surface tension effects are ignored here.

Key words. global solutions, two-phase Stokes system, interface equation, generalized evolution, upper semicontinuous convexification

AMS subject classifications. 35Q30, 35R35, 58C06, 76T05

1. Introduction.

This paper studies the dynamics of the interface (free boundary) of two immiscible, incompressible viscous fluids with the same constant density, say, one. We are interested in slow motions so that each fluid velocity satisfies the Stokes equations with different viscosities. The interface is assumed to move with the fluid velocities. No surface tension on the interface is considered in this paper.

Let ν_+ and ν_-, simply denoted by $\nu_\pm$, be the viscosities of each fluid. Let $\Omega_\pm(t)$ be the disjoint open sets in a bounded rectangle $R(\subset \mathbf{R}^n (n \geqslant 2))$ occupied with the fluids of viscosities $\nu_\pm$ at time t, respectively. The complement of the union of $\Omega_+(t)$ and $\Omega_-(t)$ is called the interface and denoted by $\Gamma(t)$. To write down the equation we assume that the interface $\Gamma(t)$ is a smooth hypersurface so that $\Gamma(t)$ is the boundary between $\Omega_+(t)$ and $\Omega_-(t)$. Let $u_\pm = u_\pm(t,x)$ and $\pi_\pm = \pi_\pm(t,x)$ denote the velocities and pressures of fluids with viscosities $\nu_\pm$, respectively. The motion of the fluids determines the dynamics of the interface. Let $V = V(t,x)$ denote the velocity of $\Gamma(t)$ at $x \in \Gamma(t)$ in the direction of the unit normal vector $\mathbf{n}$ from $\Omega_+(t)$ to $\Omega_-(t)$. We consider an interface equation for $\Gamma(t)$:

$$V = u_+ \cdot \mathbf{n} \quad \text{on} \quad \Gamma(t) \quad \text{with initial data} \quad \Omega_\pm(0) = \Omega_{\pm 0} \tag{1.1}$$

coupled with the incompressible Stokes system:

$$\partial_t u_\pm - \nu_\pm \Delta u_\pm + \nabla \pi_\pm = \nabla \cdot f_\pm \quad \text{in} \quad \Omega_\pm(t), 0 < t < T, \tag{1.2}$$

$$\nabla \cdot u_\pm = 0 \quad \text{in} \quad \Omega_\pm(t), 0 < t < T \tag{1.3}$$

$$u_+ = u_- \quad \text{on} \quad \Gamma(t), \tag{1.4}$$

$$T_+(u_+, \pi_+) \cdot \mathbf{n} = T_-(u_-, \pi_-) \cdot \mathbf{n} \quad \text{on} \quad \Gamma(t), \tag{1.5}$$

$$u_\pm(0,x) = 0 \quad \text{in} \quad \Omega_\pm(0), \tag{1.6}$$

where $T_\pm(u_\pm, \pi_\pm) := \nu_\pm D(u_\pm) - \pi_\pm I$ denotes the stress tensors with

$$D(u) = (D_{k\ell}(u)) := \frac{\partial u^k}{\partial x_\ell} + \frac{\partial u^\ell}{\partial x_k}.$$

* Received by the editors June 1, 1992; accepted for publication (in revised form) March 18, 1993.
† Department of Mathematics, Hokkaido University, Sapporo 060, Japan. The work of this author was partially supported by the Inamori Foundation.
‡ Department of Mathematical Sciences, Faculty of Science and Engineering, Tokyo Denki University, Hatoyama, Saitama 350-03, Japan.

Here $0 < \nu_- < \nu_+ < \infty$, $0 < T \leqslant \infty$, and $\nabla \cdot f_\pm = \sum_{j=1}^n \partial f_{\pm ij}/\partial x_j$ for $f_\pm = (f_{\pm ij}(t,x))$ $(i,j = 1,\dots,n)$. The initial velocities are assumed to be zero for simplicity.

Our goal is to construct global weak solutions of the two-phase Stokes system (1.1)–(1.6) for arbitary given initial regions $\Omega_{\pm 0}$ and external forces $f_\pm$ under the assumption that ν_+ and ν_- are close. We impose here periodic boundary conditions to avoid technical difficulties. Although it is possible to construct local solutions (cf. [DeSol]), there is an intrinsic difficulty to construct global solutions since the interface $\Gamma(t)$ may develop singularities in a finite time.

We first introduce a weak formulation of the transport equation (1.1). Since the boundary of our $\Omega_\pm(t)$ may not be regular, we consider a generalized evolution of (1.1) through a level set of an auxiliary function. This idea goes back to [ESou]. Recently, the level set approach is extended to other equations including the mean curvature flow equations (cf. [ES] and [CGG1]). However, our velocity field u is merely continuous, so one cannot apply these known theories directly to our setting. We are forced to extend the usual definition of generalized evolutions to (1.1) (cf. [ESou]). It turns out that our generalized evolution uniquely exists for any initial data $\Omega_{\pm 0}$ and any continuous velocity u.

Next we introduce a step function ν to give a weak formulation of (1.2)–(1.6). The region occupied with high (respectively, low) viscous fluid corresponds to the place where ν takes the value ν_+(respectively, ν_-). The interface corresponds to the jump discontinuity of ν. The velocity u is defined by $u = u_+$ on Ω_+ and $u = u_-$ on Ω_-, and also the pressure π is defined in the same manner. The system (1.2) and (1.5) is formally equivalent to

$$u_t - \nabla \cdot (\nu D(u)) + \nabla \pi = \nabla \cdot f \quad \text{in} \quad (0,T) \times \mathbf{T}, \tag{1.7}$$

where $\mathbf{T}$ is the torus obtained by identifying each end of R. Equation (1.7) should be understood in the sense of distributions. Condition (1.5) is implicit in (1.7) since ν has the jump discontinuity on the interface. Condition (1.4) is automatic if u is assumed to be continuous. Thus we obtain a weak formulation of (1.1)–(1.6) by using generalized evolutions of (1.1).

To construct a solution we seek a fixed point of the mapping defined as follows. For a continuous function v we solve (1.1) with $u_+ = v$ and find generalized evolutions $\Omega_+^v \cup \Omega_-^v$. Let $\nu = \nu_v$ be a step function with $\nu = \nu_\pm$ on $\Omega_\pm^v$ and $\nu = (\nu_+ + \nu_-)/2$ outside $\Omega_\pm^v$. We next solve (1.7) with $\nabla \cdot u = 0$ and $u(0,x) = 0$, and obtain a mapping $S : v \longmapsto u$. Unfortunately S may not be continuous, so Leray–Schauder's fixed point theory does not apply. We extend mapping S to an upper semicontinuous convex, set valued mapping so that we can apply Kakutani's fixed point theory. To apply Kakutani's theory we need a compactness which follows from a priori L^p estimates (for large p) for the Stokes system (1.7) and $\nabla \cdot u = 0$ with discontinuous viscosity. A perturbation argument similar to [Cam] and [GY] is applied here. To get the L^p estimates for large p we need to assume that $(\nu_+ - \nu_-)/\nu_+$ is sufficiently small.

Our formulation allows the possibility that the interface $\Gamma(t)$ becomes thick. If the interface has the zero Lebesgue measure in space-time, our weak solution satisfies (1.7). It is an open problem whether there is an example such that the interface actually becomes thick in finite time for the system (1.1)–(1.6). See §7 for more discussion.

In [GGI] and [GY] global solutions for the interface equations coupled with other equations are studied in different contexts.

There are related free boundary problems for one-phase incompressible viscous fluid motion. Solonnikov extensively studied the evolution of the free boundary when

the initial surface is a connected boundary of a bounded domain. He constructed a unique local smooth solution for $\sigma = 0$ in [Sol 5] and $\sigma > 0$ in [Sol 6], where σ is the surface tension. If the data is close to some equilibrium state, he showed that his solution can be extended globally in time; see [Sol 1], [Sol 2], and [Sol 3] for $\sigma > 0$ and [Sol 4] for $\sigma = 0$.

The same problem is studied when the domain is occupied with fluid, like an ocean with finite depth whose top is the free boundary. Local existence is established by Beale [Be2] and Allain [Al] for $\sigma > 0$ and by Beale [Be1] for $\sigma = 0$. Global-in-time existence of smooth solution is established by Beale [Be2] for $\sigma > 0$. Sylvester [Sy] studied global existence for $\sigma = 0$. Note that the case $\sigma = 0$ is more difficult for establishing global existence because $\sigma > 0$ gives some regularizing effect.

For the two-phase Navier–Stokes system, using a priori estimates in [De], Denisova and Solonnikov [DeSol] constructed a local solution with or without the surface tension. Tanaka [Tana] proved a global existence for $\sigma > 0$ when the initial surface is close to some equilibrium state.

The problem (1.1)–(1.6) is regarded as the two-phase Stokes system with no surface tension. The only difference between our problem and the two-phase Navier–Stokes system mentioned above is that our equations for the fluid motion are not the Navier–Stokes equations but the Stokes equations. So far even to our problem no global smooth solutions are constructed for nontrivial initial data. We note that our method actually extends to the two-phase Navier–Stokes system with inhomogeneous Dirichlet condition. This will be discussed in a forthcoming paper [T] of the second author.

It is an open problem whether our weak solution actually agrees with (unique) classical solution as far as the latter exists.

Finally we point out that Kohn and Lipton [KL] discussed the homogenization problem for the two-phase Navier–Stokes flow with no surface tension in a formal level.

We note that two-phase problems for compressible viscous fluid are extensively studied by Tani. We refer to [Tani 1], [Tani 2], and [Tani 3].

2. Interface equations. We consider the motion of interfaces with a given velocity under periodic boundary conditions. For $\alpha_i > 0$ $(i = 1, \dots, n)$ let R be a bounded rectangle in $\mathbf{R}^n$ of the form

$$R = \{(x_1, \cdots x_n) \in \mathbf{R}^n;\ 0 \leqslant x_i \leqslant \alpha_i,\ 1 \leqslant i \leqslant n\}.$$

We identify faces $x_i = 0$ and $x_i = \alpha_i$ $(1 \leqslant i \leqslant n)$ of R to get an n-dimensional flat torus $\mathbf{T}$. A motion of interfaces in R under periodic boundary conditions is interpreted as that in $\mathbf{T}$. We consider $\mathbf{T}$ rather than $\mathbf{R}^n$ for technical convenience because $\mathbf{T}$ is compact and has no boundary. The periodic boundary condition is important because it is often used in numerical experiments.

Let Ω_+ and Ω_- be disjoint open sets in $M = [0, \infty) \times \mathbf{T}$. Let Γ denote the complement of the union of Ω_+ and Ω_- in M. Physically, $\Gamma(t)$ is called an interface at time t bounding two phases $\Omega_\pm(t)$ of fluids. Here $W(t)$ denotes the cross-section of $W \subset M$ at time t, i.e.,

$$W(t) = \{x \in \mathbf{T}; (t, x) \in W\}.$$

Suppose that $\Gamma(t)$ is a smooth hypersurface and let $\mathbf{n}$ denote the unit normal vector field pointing from $\Omega_+(t)$ to $\Omega_-(t)$. Let $V = V(t, x)$ denote the velocity of $\Gamma(t)$ at

$x \in \Gamma(t)$ in the direction $\mathbf{n}$. Suppose that $u : \overline{Q} \longrightarrow \mathbf{R}^n$ is a continuous vector field, i.e., $u \in C(\overline{Q})$ where $Q = (0,T) \times \mathbf{T}$ $(0 < T \leqslant \infty)$ and that $\overline{Q}$ denotes the closure of Q in M. Here and hereafter we do not distinguish the space of real, vector or tensor valued functions. The equation for $\Gamma(t)$ we consider here is

$$V = u \cdot \mathbf{n} \quad \text{on} \quad \Gamma(t), \tag{2.1}$$

where $\cdot$ denotes the standard inner product in $\mathbf{R}^n$.

If $u(t,x)$ is Lipschitz continuous in x (uniformly in t), one can construct a unique short time classical solution for a given smooth initial data $\Gamma(0)$ by a method of characteristics. In the periodic case a unique global-in-time weak solution is constructed in [GGI] by a level set approach developed by Chen, Giga, and Goto [CGG1] and Evans and Spruck [ES]; see also [ESou]. However, if u is merely continuous, classical solutions may not exist even for a short time and they are not uniquely determined by the initial data even if they exist. The level set approach in [GGI] does not apply to this case so we are forced to extend the approach. By the way in [CGG2] we actually need to assume a uniform bound on the gradient of T in [CGG2, eq. (1.6)] and of ω in [CGG2, eq. (2.13)] although it is not written there.

Largest and smallest solutions. Let $u \in C(\overline{Q})$ and $a \in C(\mathbf{T})$. We say $\psi : Q \longrightarrow \mathbf{R}$ is a *subsolution* of

$$\psi_t + (u \cdot \nabla)\psi = 0 \quad \text{in } Q, \tag{2.2}$$

$$\psi(0,x) = a(x) \tag{2.3}$$

if ψ is a viscosity subsolution of (2.2) on Q and $\psi_*(0,x) = a(x)$, where h_* denotes the lower semicontinuous envelope of $h : I \longrightarrow \mathbf{R}$, i.e.,

$$h_*(y) = \lim_{\epsilon \downarrow 0} \inf\{h(z); |z-y| < \epsilon,\ z \in I\}, \qquad y \in \overline{I}.$$

If $-\psi$ is a subsolution of (2.2)–(2.3) with $-\psi(0,x) = -a(x)$, we say ψ is a *supersolution* of (2.2)–(2.3). If ψ is both super- and subsolution of (2.2)–(2.3), we simply say ψ is a *solution* of (2.2)–(2.3). For a general theory of viscosity solutions, see [CIL].

As is well known, there exists a comparison theorem on solutions provided that $|\nabla u|$ is uniformly bounded. However, for general $u \in C(\overline{Q})$ there is no uniqueness of solutions of (2.2)–(2.3). We thus consider largest and smallest solutions. Let λ (respectively, σ) be a solution of (2.2)–(2.3). We say λ (respectively, σ) is a *largest* (respectively, *smallest*) *solution* if $\lambda \geqslant \psi$ (respectively, $\sigma \leqslant \psi$) for all other solutions ψ of (2.2)–(2.3).

PROPOSITION 2.1. (i) *Suppose that ψ is a viscosity sub-(super)solution of* (2.2) *on Q, where $u \in C(\overline{Q})$. Then ψ is also a viscosity sub-(super)solution of*

$$\psi_t - L|\nabla\psi| = 0 \tag{2.4}$$

$$(\textit{respectively, } \psi_t + L|\nabla\psi| = 0) \tag{2.5}$$

on Q with $L \geqq \sup_Q |u|$.

(ii) *Suppose that ψ is a viscosity super-(sub)solution of* (2.4) (*respectively,* (2.5)). *Then ψ is also a viscosity super-(sub)solution of* (2.2) *on Q.*

Proof. We only present the proof of (i) when ψ is a viscosity subsolution of (2.2) because the remaining three cases can be proved similarly. Suppose that $\zeta \in C^2(Q)$

and $(t_0, x_0) \in Q$ satisfy

$$\max_{Q}(\psi - \zeta) = (\psi - \zeta)(t_0, x_0).$$

Since ψ is a viscosity subsolution of (2.2),

$$\zeta_t + (u \cdot \nabla)\zeta \leqslant 0 \quad \text{at } (t_0, x_0).$$

The Schwarz inequality now yields

$$\zeta_t - L|\nabla\zeta| \leqslant \zeta_t + (u \cdot \nabla)\zeta \leqslant 0 \quad \text{at } (t_0, x_0),$$

so ψ is a viscosity subsolution of (2.4) on Q. □

LEMMA 2.2. *Suppose that $u \in C(\overline{Q})$ and $a \in C(\mathbf{T})$. There are unique largest and smallest solutions λ and σ of (2.2)–(2.3) which are bounded on every compact set in $\overline{Q}$. Moreover, λ and σ are expressed as*

$$\lambda(t, x) = \sup\{\psi(t, x); \psi \text{ is a subsolution of (2.2)–(2.3)}\}, \tag{2.6}$$

$$\sigma(t, x) = \inf\{\psi(t, x); \psi \text{ is a supersolution of (2.2)–(2.3)}\}. \tag{2.7}$$

Proof. Let Λ denote the right hand side of (2.6). As it is well known, there exists a unique viscosity solution ψ^+ (respectively, ψ^-) of (2.4) (respectively, (2.5)) with (2.3). By Proposition 2.1 ψ^+ and ψ^- are, respectively, super- and subsolutions of (2.2)–(2.3). Also any subsolution ψ of (2.2)–(2.3) is a subsolution of (2.4)–(2.3), so a comparison theorem for (2.4) yields $\psi \leqslant \psi^+$. By Perron's method (cf. [Ish]) we see Λ is a solution of (2.2)–(2.3) with

$$\psi^- \leqslant \Lambda \leqslant \psi^+.$$

Since $\psi^{\pm}$ is continuous on $\overline{Q}$, Λ is bounded on every compact set in $\overline{Q}$. The solution Λ is a unique largest solution λ because otherwise there would exist a solution φ of (2.2)–(2.3), which is not smaller than Λ, so this contradicts the definition of Λ. We thus proved all statements on $\lambda = \Lambda$. The proof for σ is completely parallel, so is omitted. □

LEMMA 2.3 (uniqueness of level sets). *Let λ and σ be, respectively, the largest and smallest solutions of (2.2)–(2.3). Let*

$$\Omega_+ = \{(t, x) \in [0, T) \times \mathbf{T};\ \sigma_*(t, x) > 0\}, \tag{2.8}$$

$$\Omega_- = \{(t, x) \in [0, T) \times \mathbf{T};\ \lambda^*(t, x) < 0\}, \tag{2.9}$$

where $\lambda^* = -(-\lambda)_*$. The set Ω_+ (respectively, Ω_-) is completely determined by the initial data $\Omega_+(0)$ (respectively, $\Omega_-(0)$) and u, and is independent of choice of a defining $\Omega_\pm(0)$, i.e.,

$$\Omega_\pm(0) = \{x \in \mathbf{T};\ a(x) \gtrless 0\}.$$

Proof. Suppose that $a_i \in C(\mathbf{T})$ $(i = 1, 2)$ satisfies

$$\Omega_+(0) = \{x \in \mathbf{T};\ a_i(x) > 0\}.$$

Let σ_i denote the smallest solution of (2.2)–(2.3) with $a = a_i$. We first take $\theta \in C(\mathbf{R})$ (strictly) increasing with $\theta(0) = 0$ and $a_1 \leqslant \theta(a_2)$. Such a function θ, of course, exists (cf. [CGG1, Lemma 7.2]). Since (2.2) is geometric, $\varphi := \theta(\sigma_2)$ is a solution of (2.2)–(2.3) with $a = \theta(a_2)$ (cf. [CGG1, Thm. 5.2] or [CGG2, Thm. 2.3]). Moreover φ is the smallest solution of (2.2)–(2.3) with $a = \theta(a_2)$ since θ and θ^{-1} preserve the order in $\mathbf{R}$.

We next observe that $\sigma_1 \leqslant \varphi$. Indeed, $\psi = \min(\sigma_1, \varphi)$ is a supersolution of (2.2)–(2.3) with $a = a_1$ (cf. [CGG1, Prop. 2.2]). If $\sigma_1 \leqslant \varphi$ were not true, there would exist $(t, x) \in Q$ such that $\psi(t, x) < \sigma_1(t, x)$. This contradicts the representation (2.7) of the smallest solution σ_1.

The inequality $\sigma_1 \leqslant \varphi$ yields

$$\{(t, x); \sigma_{1*}(t, x) > 0\} \subset \{(t, x); \sigma_{2*}(t, x) > 0\}.$$

If we choose θ so that $a_2 \leqslant \theta(a_1)$, the other side inclusion also holds, so Ω_+ is completely determined by $\Omega_+(0)$.

The proof for Ω_- is parallel, so is omitted. □

Remark. Evans and Souganidis [ESou, Thm. 7.1] proved the uniqueness of level sets in $\mathbf{R}^n$ when (2.2) is of the form

$$u_t + H(x, \nabla u) = 0, \tag{2.10}$$

where $H : \mathbf{R}^n \times \mathbf{R}^n \longrightarrow \mathbf{R}$ is uniformly Lipschitz and positively homogeneous of degree one in the second variable. In this case there is no need to consider largest and smallest solutions because solutions of (2.10) with (2.3) are unique by comparison. The proof given there is different from that in [CGG1] and [CGG2] and does not seem to apply to second-order equations. Of course, the proof in [CGG1] and [CGG2] does apply to second-order equations.

Generalized evolution. Let Ω_+ (respectively, Ω_-) be an open set in M. We say Ω_+ (respectively, Ω_-) is a $+$ (respectively, $-$) *generalized evolution with velocity* $u \in C(\overline{Q})$ *and initial data* $\Omega_+(0)$ (respectively, $\Omega_-(0)$) on interval $[0, T)$ if there is the smallest (respectively, largest) solution σ (respectively, λ) of (2.2)–(2.3) satisfying (2.8) (respectively, (2.9)) with some $a \in C(\mathbf{T})$ defining $\Omega_\pm(0)$.

Note that each level set of solutions of (2.2)–(2.3) independently moves by (2.1) at least formally. The sign $\pm$ reflects on the orientation of the interface.

For a given open set Ω_{+0} in $\mathbf{T}$ there is $a \in C(\mathbf{T})$ satisfying $\Omega_{+0} = \{x; a(x) > 0\}$, so Lemmas 2.2 and 2.3 yield the following.

THEOREM 2.4. *For a given open set* Ω_{+0} (*respectively,* Ω_{-0}) *in* $\mathbf{T}$ *there is a unique* $+$ (*respectively,* $-$) *generalized evolution* Ω_+ (*respectively,* Ω_-) *with velocity* $u \in C(\overline{Q})$ *and initial data* $\Omega_\pm(0) = \Omega_{\pm 0}$ *on* $[0, T)$. *If* Ω_{+0} *and* Ω_{-0} *are disjoint, so are* Ω_+ *and* Ω_-.

THEOREM 2.5 (stability). *Suppose that* $T < \infty$ *and* $u_j \longrightarrow u$ *in* $C(\overline{Q})$ *as* $j \to \infty$. *Let* Ω_{+j} *be the* $+$ *generalized evolution with velocity* $u_j \in C(\overline{Q})$ *and initial data* $\Omega_{+j}(0) = \Omega_{+0}$ *on* $[0, T)$, *where* $j = 1, 2, \ldots$ *and* $Q = (0, T) \times \mathbf{T}$. *Let* Ω_+ *be the* $+$ *generalized evolution on* $[0, T)$ *with velocity* u *and* $\Omega_+(0) = \Omega_{+0}$. *Let* K *be a compact set in* Ω_+. *Then* K *is also contained in* Ω_{+j} *for sufficiently large* j. *The same holds for the* $-$ *evolution.*

Proof. Let σ_j be the smallest solution of

$$\psi_t + (u_j \cdot \nabla)\psi = 0, \qquad \psi(0, x) = a(x) \in C(\mathbf{T})$$

with $\Omega_{+0} = \{x; a(x) > 0\}$. By the stability result of Barles and Perthame [BP, Appendix] the function

$$\varphi(t,x) := \lim\nolimits_* \sigma_j(t,x) := \lim_{\substack{j\to\infty \\ \varepsilon\downarrow 0}} \inf\{\sigma_j(s,y); |t-s| < \varepsilon, |y-x| < \varepsilon\}$$

is a viscosity supersolution of (2.2) on Q since $u_j \longrightarrow u$ in $C(\overline{Q})$. Let L be a constant such that $\sup_Q |u_j| \leqslant L$ for all j. We take a continuous solution ψ^+ (respectively, ψ^-) of (2.4) (respectively, (2.5)) with (2.3). As in the proof of Lemma 2.2, we have $\psi^- \leqslant \sigma_j \leqslant \psi^+$. This implies that $\psi^- \leqslant \varphi \leqslant \psi^+$ on $[0,T) \times \mathbf{T}$, so we have $\varphi_*(0,x) = a(x)$. Therefore φ is a supersolution of (2.2)–(2.3). Let σ be the smallest solution of (2.2)–(2.3) so that $\varphi \geqslant \sigma$ by (2.7). For any compact set $K \subset \Omega_+$ there is $\delta > 0$ such that $\inf_K \sigma_* \geqslant \delta$ since σ_* is lower semicontinuous. Since $\varphi \geqslant \sigma$ and K is compact we see $\inf_K \sigma_{j*} \geqslant \delta/2$ for sufficiently large j. This implies $K \subset \Omega_{+j}$ for large j. The proof for $-$ evolution is parallel, so it is omitted. □

3. Global existence of weak solutions. We introduce a weak formulation of the problem (1.2)–(1.5) on $\mathbf{T}$. Let $\Omega_\pm$ be two disjoint open sets in $[0,T) \times \mathbf{T}$. Let ν be a step function such that $\nu = \nu_\pm$ in $\Omega_\pm$ and $\nu = (\nu_+ + \nu_-)/2$ outside $\Omega_+ \cup \Omega_-$, where $0 < \nu_- < \nu_+$. We take the mean value just to fix the idea. We may assign any value between ν_- and ν_+ provided that ν is measurable. Let f be a tensor field on $Q = (0,T) \times \mathbf{T}$ such that $f = f_\pm$ on $\Omega_\pm$. We say u is a *weak solution* of (1.2)–(1.5) for $\Omega_\pm$ in Q if $u \in C(\overline{Q})$ with $\nabla u \in L^q(Q)$ (for some $1 < q < \infty$) and it solves

$$u_t - \nabla\cdot(\nu D(u)) + \nabla\pi = \nabla\cdot f + \nabla\cdot g, \qquad \nabla\cdot u = 0 \quad \text{in} \quad Q = (0,T)\times\mathbf{T}, \tag{3.1}$$

in the sense of distribution with some π and some tensor field g whose support spt g is contained in $\Gamma = \overline{Q}\backslash(\Omega_+ \cup \Omega_-)$. By $L^p(Q)$ we mean the space of all periodic (in space) functions f on $(0,T)\times\mathbf{R}^n$ with period $\alpha = (\alpha_1,\dots,\alpha_n)$ such that $f|_{(0,T)\times R} \in L^p((0,T)\times R)$.

If the Lebesgue measure of the interface Γ is zero, then (3.1) yields (1.2)–(1.3) by interpreting $u = u_\pm$ in $\Omega_\pm$. If $\{\Gamma(t)\}_{t\geq 0}$ is a smooth family of smooth hypersurfaces, the boundary condition (1.5) is contained in (3.1). The condition (1.4) is automatic since $u \in C(\overline{Q})$.

We now state our main result in this paper.

THEOREM 3.1. *Let $p > 2(n+1)$. Assume that $\Omega_{\pm 0}$ are two disjoint open sets in $\mathbf{T}$ and that $f \in L^p(Q)$ is a tensor field. Then there exists a positive constant $\delta = \delta(n,p)$ such that if*

$$\frac{\nu_+ - \nu_-}{\nu_+} < \delta, \tag{3.2}$$

then there exist $u \in C(\overline{Q})$ with $\nabla u \in L^p(Q)$ and $\Omega_\pm \subset \overline{Q}$ such that u is a weak solution of (1.2)–(1.5) for $\Omega_\pm$ with (1.6) and that $\Omega_\pm$ are generalized evolutions with the velocity u and initial data $\Omega_{\pm 0}$. Moreover, g in (3.1) can be taken as an element of $L^p((0,T_0)\times\mathbf{T})$ for all finite $T_0 \leq T$. Here T is allowed to be infinite.

4. Upper semicontinuous convexification. This section establishes a crucial abstract theory for (set-valued) mappings so that we apply Kakutani's fixed point theory. For this purpose we extend a mapping to an upper semicontinuous convex set-valued mapping.

For a given subset A of a vector space X let $\mathrm{co}A$ denote the convex hull of A, i.e.,

$$\mathrm{co}A = \{tx + (1-t)y\ ; x,\ y \in A,\ 0 \leqslant t \leqslant 1\}.$$

Let X and Y be a normed space and a Banach space equipped with norms $||\cdot||_X$ and $||\cdot||_Y$, respectively. For a set-valued mapping $S : X \longrightarrow 2^Y$ we define $S_\varepsilon : X \longrightarrow 2^Y$ by

$$S_\varepsilon(u) = \bigcup\{S(w); ||u-w||_X < \varepsilon\} \subset Y$$

for $u \in X$ and $\varepsilon > 0$. Here 2^Y denotes the family of all subsets of Y. We introduce another set-valued mapping $\mathcal{S} : X \longrightarrow 2^Y$ defined by

$$\mathcal{S}(u) = \bigcap_{\varepsilon>0} \overline{\mathrm{co}S_\varepsilon(u)}, \qquad u \in X,$$

where $\overline{B}$ denotes the closure of $B \subset Y$. In this paper we call $\mathcal{S}$ *upper semicontinuous convexification* of S since it has the following properties.

LEMMA 4.1. (i) *For each $u \in X$ the set $\mathcal{S}(u)$ is closed and convex in Y.*

(ii) *The mapping $\mathcal{S}$ is upper semicontinuous. In other words, if $u_j \longrightarrow u$ in X, $v_j \in \mathcal{S}(u_j)$ and $v_j \longrightarrow v$ in Y, then $v \in \mathcal{S}(u)$.*

(iii) *If $S(u)$ is nonempty for all $u \in X$, so is $\mathcal{S}$.*

Proof. (i) Clearly, $\mathcal{S}(u)$ is closed. Since the closure of a convex set is convex and the intersection of a family of convex sets is also convex, we see $\mathcal{S}(u)$ is convex as well.

(ii) Suppose that $v \notin \mathcal{S}(u)$. Then there would exist $\delta > 0$ such that

$$v \notin A_\delta(u) := \overline{\mathrm{co}S_\delta(u)}.$$

Since $A_\delta(u)$ is closed, there would exist k such that $j \geqslant k$ implies that $v_j \notin A_\delta(u)$. Since $u_j \longrightarrow u$ we may assume that $||u_j - u||_X < \delta/2$ for $j \geqslant k$ by taking k larger. By the definition of S_ε we see

$$A_\delta(u) \supset A_{\delta/2}(u_j), \qquad j \geqslant k.$$

This inclusion now would imply $v_j \notin A_{\delta/2}(u_j)$, i.e., $v_j \notin \mathcal{S}(u_j)$ for $j \geqslant k$, which leads a contradiction.

(iii) Since $S_\varepsilon(u)$ contains $S(u)$, so does $\mathcal{S}(u)$. □

We have introduced the upper semicontinuous convexification so that we apply Kakutani's fixed point theory. We state an easy consequence of this fixed point theory for later use.

PROPOSITION 4.2. *Let K be a convex compact subset of a Banach space X and let $S : X \longrightarrow 2^K \subset 2^X$ be a nonempty set-valued mapping. Let $\mathcal{S}$ be the upper semicontinuous convexification of S. Then $\mathcal{S}$ has a fixed point $\overline{u} \in K \cap \mathcal{S}(\overline{u})$.*

Proof. Since K is convex and closed, values of $\mathcal{S}$ are contained in K. By Lemma 4.1 we see $\mathcal{S}$ is an upper semicontinuous set-valued mapping $X \longrightarrow 2^K$ with nonempty closed convex values. The existence of a fixed point of $\mathcal{S}$ now follows from Kakutani's fixed point theorem [AF]. □

5. Stokes equations with discontinuous coefficients. Let us recall anisotropic Sobolev spaces of fractional orders (cf. [Ya, Ex. 1.1 and §3] and [Tri, §2.13]) though our notation differs from them. For $1 < p < \infty$ and $0 < s < \infty$, let $H_p^{s,2s} = H_p^{s,2s}(\mathbf{R}\times\mathbf{R}^n)$ denote

$$H_p^{s,2s} = \{f \in L^p(\mathbf{R}\times\mathbf{R}^n);\ ||f||_{H_p^{s,2s}} = ||\mathcal{F}^{-1}\langle\tau,\xi\rangle^{2s}\mathcal{F}f||_{L^p(\mathbf{R}\times\mathbf{R}^n)} < \infty\}$$

with $\langle \tau, \xi \rangle = [\{1 + |\xi|^2 + (1 + 2|\xi|^2 + |\xi|^4 + 4\tau^2)^{1/2}\}/2]^{1/2}$. Here $(\mathcal{F}f)(\tau, \xi)$ denotes the Fourier transform of $f(t, x)$ on $\mathbf{R} \times \mathbf{R}^n \ni (t, x)$. The multiplier $\langle \tau, \xi \rangle$ is actually the unique positive root t of

$$t^{-2} + t^{-2 \cdot 2}\tau^2 + t^{-2}|\xi|^2 = 1.$$

For a domain D in $\mathbf{R}^{n+1}$ let $H_p^{s,2s}(D)$ denote the space of all $f \in L^p(D)$ which can be extended to an element $\bar{f}$ of $H_p^{s,2s}$. The space $H_p^{s,2s}(D)$ is equipped with the norm

$$||f||_{H_p^{s,2s}(D)} = \inf\{||\bar{f}||_{H_p^{s,2s}};\ \bar{f}|_D = f\}.$$

Let $H_p^{s,2s}(Q)$ denote the space of all periodic (in space) functions f defined on $(0, T) \times \mathbf{R}^n$ with the period $\alpha = (\alpha_1, \cdots, \alpha_n)$ such that

$$f|_{(0,T)\times R} \in H_p^{s,2s}((0, T) \times R).$$

The space $H_p^{s,2s}(Q)$ is equipped with the norm

$$||f||_{H_p^{s,2s}(Q)} = ||f||_{H_p^{s,2s}((0,T)\times R)}.$$

We shall write $H_p^{1/2,1}(Q)$ simply by $\mathcal{H}_p(Q)$. We begin with a priori estimates for the heat equation.

LEMMA 5.1. *Let $0 < T \leqslant \infty$ and $Q = (0, T) \times \mathbf{T}$. Let $2 < p < \infty$ and let $F \in L^p(Q)$ be a vector field. Then there exists a unique solution $u \in \mathcal{H}_p(Q)$ of*

$$u_t - \Delta u = \nabla \cdot F \quad \text{in } Q,$$
$$u|_{t=0} = 0.$$

Moreover there is a constant $C_1 = C_1(n, p)$ such that

$$||u||_{\mathcal{H}_p(Q)} \leqslant C_1 ||F||_{L^p(Q)}.$$

The restriction $2 < p$ guarantees that $u \in \mathcal{H}_p(Q)$ has a trace at $t = 0$. However if we interpret $u|_{t=0} = 0$ in a suitable way, the restriction $p > 2$ is weakened as $p > 1$.

Proof. The uniqueness is standard. For example, multiplying u with $u_t - \Delta u = 0$ and integrating in space by parts yields a differential inequality which implies $u \equiv 0$.

We extend u and F periodically outside R so that u solves

$$u_t - \Delta u = \nabla \cdot F \quad \text{in } (0, T) \times \mathbf{R}^n.$$

The solution u is expressed as

$$u(t, x) = \int_0^t \int_{\mathbf{R}^n} (\nabla g)(t - s, x - y) \cdot F(s, y)\, dyds,$$

where $g(t, x) = (4\pi t)^{-n/2} \exp(-|x|^2/4t)$ is the heat kernel. Since F is periodic with period $\alpha = (\alpha_1, \cdots, \alpha_n)$, we observe that

$$u(t, x) = \sum_{\sigma \in \mathbf{Z}^n} v(t, x - \sigma\alpha), \qquad v(t, x) = \int_0^t \int_R (\nabla g)(t - s, x - y) \cdot F(s, y)\, dyds,$$

where $\sigma\alpha = (\sigma_1\alpha_1, \cdots, \sigma_n\alpha_n)$. Note that

$$||u||_{\mathcal{H}_p(Q)} \leq \sum_{\sigma\in\mathbf{Z}^n} ||v(\cdot, \cdot - \sigma\alpha)||_{H_p^{1/2,1}((0,T)\times R)} \leq ||v||_{H_p^{1/2,1}(D)}$$

with $D = (0,T) \times \mathbf{R}^n$. By Mikhlin's lemma (cf. [MS]) we see

$$||v||_{H_p^{1/2,1}(D)} \leq C_1||F||_{L^p(Q)}$$

with $C_1 = C_1(n,p)$. These two inequalities yield Lemma 5.1. □

We apply Lemma 5.1 and a perturbation argument (cf. [Cam] and [GY]) to the Stokes system with discontinuous coefficients and obtain the following results.

PROPOSITION 5.2. *Assume that* $0 < T \leqslant \infty$ *and* $2 < p < \infty$. *Assume that* $\nu \in L^\infty(Q)$ *satisfies*

$$0 < \nu_- \leq \nu \leq \nu_+ \tag{5.1}$$

with some constants $\nu_\pm$. *Let* $f \in L^p(Q)$ *be a tensor field. Then there exists a positive constant* $\delta = \delta(n,p)$ *such that*

$$\frac{\nu_+ - \nu_-}{\nu_+} < \delta \tag{5.2}$$

implies that the Stokes system

$$\begin{aligned} &u_t - \nabla\cdot(\nu D(u)) + \nabla\pi = \nabla\cdot f, \qquad \nabla\cdot u = 0 \quad \text{in} \quad Q, \\ &u|_{t=0} = 0, \end{aligned} \tag{5.3}$$

has a unique solution $u \in \mathcal{H}_p(Q)$ (*with some function* π) *satisfying*

$$||u||_{\mathcal{H}_p(Q)} \leqslant \frac{C_2}{\nu_+}||f||_{L^p(Q)} \tag{5.4}$$

with $C_2 = C_2(n,p)$.

Proof. Let P be the projection of $L^p(\mathbf{T})$ to $L^p_\sigma(\mathbf{T})$ associated with the Helmholtz decomposition

$$L^p(\mathbf{T}) = L^p_\sigma(\mathbf{T}) \oplus \{\nabla\pi \in L^p(\mathbf{T});\ \pi \in L^p(\mathbf{T})\},$$

$$L^p_\sigma(\mathbf{T}) = \{u \in L^p(\mathbf{T});\ \nabla\cdot u = 0 \text{ in } \mathbf{T}\}.$$

Since $\mathbf{T}$ has no boundary, P commutes with partial derivatives on $\mathbf{T}$. Applying P to the first equation of (5.3) yields

$$u_t - \nabla\cdot(P\nu D(u)) = \nabla\cdot(Pf). \tag{5.5}$$

Here Pf is a tensor field defined by

$$(Pf)_{ij} = (Pf_j)_i, \qquad 1 \leq i,j \leq n$$

for a tensor field f and f_j represents a vector field defined by $f_j = (f_{ij})_{1\leq i\leq n}$. From (5.5) it follows

$$u_t - \nu_+ \Delta u = \nabla \cdot P(f + (\nu - \nu_+)D(u)) \tag{5.6}$$

since $\nabla \cdot u = 0$.

We shall solve (5.6) with $u|_{t=0} = 0$ by a successive approximation. Let u_{j+1} be a solution of

$$\begin{aligned} \partial_t u_{j+1} - \nu_+ \Delta u_{j+1} &= \nabla \cdot P(f + (\nu - \nu_+)D(u_j)), \\ u_{j+1}|_{t=0} &= 0, \end{aligned} \tag{5.7}$$

for $j \geq 1$ and let $u_1 \equiv 0$. Since P is bounded from $L^p(\mathbf{T})$ to $L^p_\sigma(\mathbf{T})$ and $||D(u)||_{L^p(Q)} \leq 2C_0||u||_{\mathcal{H}_p(Q)}$ (cf. Appendix, Lemma A.1(vi)), it follows from (5.1) that

$$||P(f + (\nu - \nu_+)D(u_j))||_{L^p(Q)} \leq C(||f||_{L^p(Q)} + 2C_0(\nu_+ - \nu_-)||u_j||_{\mathcal{H}_p(Q)}).$$

The bound C of P here is actually independent of $\mathbf{T}$. Indeed, note that $Pu = u - \nabla q$ with $\Delta q = \nabla \cdot u$ in $\mathbf{T}$. Extend q and u periodically outside R so that $\Delta q = \nabla \cdot u$ is regarded as an equation on $\mathbf{R}^n$. As in the proof of Lemma 5.1, applying Mikhlin's lemma to the integral representation of ∇q we obtain

$$||\nabla q||_{L^p(R)} \leq C'||u||_{L^p(R)}$$

with $C' = C'(n,p)$.

Applying Lemma 5.1 with a change of a variable $s = t/\nu_+$ to (5.7), we now obtain

$$\begin{aligned} ||u_{j+1}||_{\mathcal{H}_p(Q)} &\leqslant \frac{C_1 C}{\nu_+}||f||_{L^p(Q)} + C''\frac{\nu_+ - \nu_-}{\nu_+}||u_j||_{\mathcal{H}_p(Q)}, \\ C'' &= 2C_0C_1C. \end{aligned} \tag{5.8}$$

We thus observe that $u_j \in \mathcal{H}_p(Q)$ for all $j \geq 1$.

Choose δ such that $C''\delta < \frac{1}{2}$. Since (5.7) is linear in u_{j+1} and u_j, the difference $w_{j+1} = u_{j+1} - u_j$ solves

$$\partial_t w_{j+1} - \nu_+ \Delta w_{j+1} = \nabla \cdot P((\nu - \nu_+)D(w_j)) \quad \text{in } Q,$$

$$w_{j+1}|_{t=0} = 0.$$

As in deriving (5.8) applying Lemma 5.1 we observe, by (5.2), that

$$\begin{aligned} ||w_{j+1}||_{\mathcal{H}_p(Q)} &\leqslant C''\frac{\nu_+ - \nu_-}{\nu_+}||w_j||_{\mathcal{H}_p(Q)} \\ &\leq \frac{1}{2}||w_j||_{\mathcal{H}_p(Q)} \end{aligned}$$

for $j \geq 2$. This implies that $\{u_j\}$ is a Cauchy sequence in $\mathcal{H}_p(Q)$.

The limit u of $\{u_j\}$ solves (5.6) with $u|_{t=0} = 0$. The estimate (5.8) yields

$$||u||_{\mathcal{H}_p(Q)} \leqslant \frac{C_1 C}{\nu_+}||f||_{L^p(Q)} + \frac{1}{2}||u||_{\mathcal{H}_p(Q)}.$$

We now obtain (5.4) with $C_2 = 2C_1C$. Since P commutes with partial derivatives and u solves (5.6), we see $\nabla \cdot u = 0$. We have thus constructed a solution $u \in \mathcal{H}_p(Q)$ of (5.3) with (5.4) under (5.2). The uniqueness of solutions follows from (5.4). □

6. Proof of Theorem 3.1. Assume that $0 < T < \infty$. For $u \in C(\overline{Q})$ let $\Omega_\pm \subset \overline{Q}$ be generalized evolutions with velocity u and initial data $\Omega_{\pm 0}$. Let $\nu = \nu_u$ be a step function such that $\nu = \nu_\pm$ in $\Omega_\pm$ and $\nu = (\nu_+ + \nu_-)/2$ outside $\Omega_+ \cup \Omega_-$ with $0 < \nu_- < \nu_+$. Assume that $f \in L^p(Q)$. If positive constant δ is chosen as in Proposition 5.2, then there is a unique solution $\tilde{u}$ of (5.3) for $\nu = \nu_u$ such that

$$\tilde{u} \in K = \left\{ w \in \mathcal{H}_p(Q); \|w\|_{\mathcal{H}_p(Q)} \leqslant \frac{C_2}{\nu_+} \|f\|_{L^p(Q)} \right\}.$$

We define a mapping $S : C(\overline{Q}) \longrightarrow 2^K$ by $S(u) := \{\tilde{u}\}$. If $p > 2(n+1)$, the inclusion

$$\mathcal{H}_p(Q) \subset C^\mu(\overline{Q})$$

for $\mu = 1/2(n+1) - 1/p$ is continuous (see Appendix) and Ascoli–Arzela's theorem implies that K is compact in Banach space $C(\overline{Q})$ since $T < \infty$. Unfortunately Leray–Schauder's fixed point theory does not apply to S since S may not be continuous. We consider the upper semicontinuous convexification $\mathcal{S}$ of S in §4.

LEMMA 6.1. *Let $\mathcal{S}$ be the upper semicontinuous convexification of S. If $v \in \mathcal{S}(u)$, then v is a weak solution of* (1.2)–(1.5) *with* (1.6) *for generalized evolutions $\Omega_\pm$ with velocity u and initial data $\Omega_{\pm 0}$. Moreover, g in* (3.1) *belongs to $L^p(Q)$.*

Proof. By the definition of $\mathcal{S}$, if $v \in \mathcal{S}(u)$, then for each $k = 1, 2, \ldots$, there is a sequence $\{v_m^k\}_{m=1}^\infty$ converging to v in $C(\overline{Q})$ such that $v_m^k \in \mathrm{co} S_{1/k}(u)$. In other words v_m^k is expressed as

$$v_m^k = \sum_{j=m}^{\ell} \lambda_j^{mk}\, \tilde{u}_j^k, \quad \{\tilde{u}_j^k\} = S(u_j^k), \quad \ell = \ell(m,k),$$

with some λ_j^{mk} and $u_j^k \in C(\overline{Q})$ such that

$$\sum_{j=m}^{\ell} \lambda_j^{mk} = 1, \quad \lambda_j^{mk} \geqslant 0, \quad \|u_j^k - u\|_{C(\overline{Q})} < 1/k.$$

By a diagonal argument there are a sequence u_j converging to u in $C(\overline{Q})$ and $\lambda_m^m, \ldots, \lambda_{\ell_m}^m$ with

$$\sum_{j=m}^{\ell_m} \lambda_j^m = 1, \qquad \lambda_j^m \geqslant 0$$

such that

$$v_m = \sum_{j=m}^{\ell_m} \lambda_j^m\, \tilde{u}_j, \qquad \{\tilde{u}_j\} = S(u_j)$$

converges to v in $C(\overline{Q})$ as $m \to \infty$. In the definition of S, $\tilde{u}_j$ solves

$$\begin{aligned} &\partial_t \tilde{u}_j - \nabla \cdot (\nu_{u_j} D(\tilde{u}_j)) + \nabla \tilde{\pi}_j = \nabla \cdot f \quad \text{in } Q, \\ &\nabla \cdot \tilde{u}_j = 0 \quad \text{in } Q, \\ &\tilde{u}_j|_{t=0} = 0, \end{aligned}$$

with some $\tilde{\pi}_j$. Multiplying λ_j^m and adding from m to ℓ_m we see

$$
\begin{aligned}
&\partial_t v_m - \nabla \cdot (\nu_u D(v_m)) + \nabla \pi_m = \nabla \cdot f + \nabla \cdot g_m \quad \text{in } Q, \\
&\nabla \cdot v_m = 0 \quad \text{in } Q, \\
&v_m|_{t=0} = 0,
\end{aligned} \tag{6.1}
$$

with

$$
\pi_m = \sum_{j=m}^{\ell_m} \lambda_j^m \tilde{\pi}_j,
$$

$$
g_m = \sum_{j=m}^{\ell_m} \lambda_j^m (\nu_{u_j} - \nu_u) D(\tilde{u}_j).
$$

Since K is convex and bounded, the sequence $\{v_m\}$ is bounded in $\mathcal{H}_p(Q)$ so that $\{D(v_m)\}$ is bounded in $L^p(Q)$ (cf. Appendix). We thus observe that $D(v_m) \rightharpoonup D(v)$ weakly in $L^p(Q)$ since $v_m \to v$ in $C(\overline{Q})$. Since $\tilde{u}_j \in K$, the sequence $\{g_m\}$ is bounded in $L^p(Q)$. Taking a subsequence if necessary, $g_m \rightharpoonup g$ weakly in $L^p(Q)$ for some $g \in L^p(Q)$. Letting $m \to \infty$ in (6.1) yields

$$
\begin{aligned}
&\partial_t v - \nabla \cdot (\nu_u D(v)) + \nabla \pi = \nabla \cdot f + \nabla \cdot g \quad \text{in } Q, \\
&\nabla \cdot v = 0 \quad \text{in } Q, \\
&v|_{t=0} = 0
\end{aligned}
$$

for some π.

It remains to prove that $\operatorname{spt} g \subset \overline{Q} \setminus (\Omega_+ \cup \Omega_-)$. Let C be a compact set in $\Omega_+ \cup \Omega_-$. Since $u_j \to u$ in $C(\overline{Q})$, we see, by Theorem 2.5, $\nu_{u_j} = \nu_u$ on C for sufficiently large j. This implies that $g_j \equiv 0$ on C for sufficiently large j. Since $g_j \rightharpoonup g$ weakly in $L^p(Q)$ and C can be taken as an arbitrary ball in $\Omega_+ \cup \Omega_-$, we conclude that $g \equiv 0$ on $\Omega_+ \cup \Omega_-$. □

If $T < \infty$ and $p > 2(n+1)$, K is compact and convex in $X = C(\overline{Q})$. By Proposition 4.2 $\mathcal{S}$ has a fixed point $u \in K \cap \mathcal{S}(u)$. By Lemma 6.1 this u is a desired weak solution in Theorem 3.1.

To complete the proof of Theorem 3.1, it remains to construct a global solution in $(0, \infty)$. For $0 < T < \infty$ we write Q by Q_T, K by K_T and $\mathcal{S}$ by $\mathcal{S}_T$ to emphasize the dependence of T. For $T_1 < T_2 < \cdots < T_i \longrightarrow \infty$ let u_{T_i} be a fixed point in $K_{T_i} \cap \mathcal{S}_{T_i}(u_{T_i})$. Since δ in Proposition 5.2 is independent of time, for each $T < \infty$ the restrictions $\{\hat{u}_i\}$ of $\{u_{T_i}\}$ on $t \le T$ are bounded in $K_T \subset \mathcal{H}_p(Q_T)$ for sufficiently large i. Since the inclusion $\mathcal{H}_p(Q_T) \subset C(\overline{Q_T})$ is compact for $p > 2(n+1)$ and $T < \infty$, a diagonal argument yields a subsequence $\{\hat{u}_{i'}\}$ and $w \in C((0,\infty) \times \mathbf{T})$ satisfying

$$
\hat{u}_{i'} \longrightarrow w \quad \text{in} \quad C(\overline{Q_T}). \tag{6.2}
$$

Since $\hat{u}_{i'} \in \mathcal{S}_T(\hat{u}_{i'}) \subset C(\overline{Q_T})$ and since the graph of $\mathcal{S}_T : C(\overline{Q_T}) \longrightarrow 2^{C(\overline{Q_T})}$ is closed in $C(\overline{Q_T}) \times C(\overline{Q_T})$, (6.2) implies $w|_{Q_T} \in \mathcal{S}_T(w|_{Q_T}) \subset C(\overline{Q_T})$. Since T is arbitrary, this yields a desired global solution in $(0, \infty)$. □

7. Discussion: fattening of interface. Suppose that Ω_{+0} and Ω_{-0} are disjoint open sets in $\mathbf{T}$. As proved in Theorem 2.4 there are unique $\pm$ (mutually disjoint) evolutions $\Omega_\pm$ with velocity $u \in C(\overline{Q})$ and initial data $\Omega_{\pm 0}$. The complement Γ of the union of Ω_+ and Ω_- is called the generalized interface evolution. We do not know whether $\Gamma(t)$ becomes to have a positive Lebesgue measure at some time t for smooth initial data $\Gamma(0)$. Even if we assume div $u = 0$, we do not know because u is merely continuous. If ∇u is bounded in Q, $\Gamma(t)$ has the zero measure (for all t) if so does $\Gamma(0)$.

LEMMA 7.1. *Suppose that* $|\nabla u|$ *is bounded on* Q *and that* div $u = 0$. *Then the Lebesgue measure* $\mathcal{L}^n(\Omega_\pm(t))$ *is independent of time. In particular,* $\mathcal{L}^n(\Gamma(t)) = 0$ *so that* $\mathcal{L}^{n+1}(\Gamma) = 0$ *provided that* $\mathcal{L}^n(\Gamma(0)) = 0$.

Proof. By Lipschitz continuity of u in x there is a unique viscosity solution $\psi \in C([0,T) \times \mathbf{T})$ of

$$\begin{aligned} &\psi_t + (u \cdot \nabla)\psi = 0 \quad \text{in } Q \\ &\psi(0,x) = a(x) \end{aligned} \tag{7.1}$$

for given $a \in C(\mathbf{T})$. If $\Omega_{+0} = \{x \in \mathbf{T}; a(x) > 0\}$,

$$\Omega_+ = \{(t,x) \in [0,T) \times \mathbf{T}; \psi(t,x) > 0\}.$$

Such a solution is a uniform limit of solution ψ^ε of approximate equation

$$\psi_t + (u \cdot \nabla)\psi = \varepsilon \Delta \psi \quad \text{in } Q$$

as $\varepsilon \downarrow 0$. For $T > t > 0$ using div $u = 0$, we have

$$\frac{d}{dt} \int_{\mathbf{T}} \psi^\varepsilon dx = -\int_{\mathbf{T}} (\operatorname{div}(u\psi^\varepsilon) - \varepsilon \Delta \psi^\varepsilon) dx = 0.$$

Sending ε to zero now yields

$$\int_{\mathbf{T}} \psi(t,x) dx = \int_{\mathbf{T}} a(x) dx \quad \text{for all } t > 0. \tag{7.2}$$

We set θ_m by

$$\theta_m(\xi) = \begin{cases} 1, & \xi \geqslant 1/m, \\ m\xi, & 0 \leqslant \xi \leqslant 1/m, \\ 0, & \xi \leqslant 0, \end{cases}$$

so that θ_m approximates the Heaviside function. Since (7.1) is geometric (cf. [CGG1]), $\psi_m = \theta_m(\psi)$ solves (7.1) with $\psi_m(0,x) = \theta_m(a(x))$ and

$$\Omega_+ = \{(t,x) \in [0,T) \times \mathbf{T}; \psi_m(t,x) > 0\}.$$

By (7.2) we observe that

$$\int_{\mathbf{T}} \psi_m(t,x) dx = \int_{\mathbf{T}} \theta_m(a(x)) dx.$$

Letting $m \to \infty$ yields

$$\mathcal{L}^n(\Omega_+(t)) = \mathcal{L}^n(\Omega_{+0}) \quad \text{for} \quad 0 < t < T.$$

The proof for Ω_- is the same. In particular, we observe $\mathcal{L}^n(\Gamma(t)) = \mathcal{L}^n(\Gamma(0))$. By Fubini's theorem we now have $\mathcal{L}^{n+1}(\Gamma) = 0$ if $\mathcal{L}^n(\Gamma(0)) = 0$. □

Uniqueness problem. We do not know the uniqueness of our weak solutions even if a (unique) classical solution exists. We hope our solution agrees with the classical one as far as the latter exists.

Our solution satisfies (3.1) (not, in general, (1.7)); the term $\nabla \cdot g$ comes from convexification. However, since support of $g \in L^p(Q)$ is contained in the interface Γ, our solution actually solves (1.7) in the usual sense provided that Γ has the zero Lebesgue measure in Q. We hope uniquess of solutions in this case. However, we do not know any example such that the interface Γ of our solution has a positive Lebesgue measure for smooth initial surface $\Gamma(0)$.

If the interface has a positive measure, we hope no uniqueness. Actually, if Γ has an interior point, we easily observe no uniqueness of solutions by modifying u on the interface. We need further constitutive information on the interface to get uniqueness.

Appendix. We list a couple of properties of anisotropic Sobolev spaces for the reader's convenience since such spaces may be less familiar than isotropic ones.

LEMMA A.1. (i) *For $0 \le s \le 1$ the space $H_p^{s,2s}(\mathbf{R}^{n+1})$ is isomorphic to the complex interpolation space $[L^p(\mathbf{R}^{n+1}), H_p^{1,2}(\mathbf{R}^{n+1})]_s$ as Banach spaces.*

(ii) *The norm $||f||_{H_p^{1,2}}$ is equivalent to the norm*

$$||f||_{L^p} + ||\nabla^2 f||_{L^p} + ||\partial_t f||_{L^p}.$$

(iii) *Let D be a domain in $\mathbf{R}^{n+1}$ of the form $(t_0, t_1) \times \Omega$ with smoothly bounded domain Ω in $\mathbf{R}^n$. There is a continuous linear operator e from $H_p^{1,2}(D)$ to $H_p^{1,2}(\mathbf{R}^{n+1})$ such that $ef = f$ on D.*

(iv) *$H_p^{s,2s}(D) = [L^p(D), H_p^{1,2}(D)]_s$ for $0 \le s \le 1$.*

(v) *For $p > 2(n+1)$ the space $H_p^{1/2,1}(Q)$ is continuously embedded in $C^\mu(\overline{Q})$ with $\mu = 1/2(n+1) - 1/p$.*

(vi) *There is a constant $C_0 = C_0(n,p)$ such that*

$$\left\| \frac{\partial u}{\partial x_j} \right\|_{L^p(Q)} \le C_0 ||u||_{\mathcal{H}_p(Q)} \quad \text{for all} \quad u \in \mathcal{H}_p(Q), j = 1, \dots, n.$$

Proof. (i) For $f \in H_p^{1,2}(\mathbf{R}^{n+1})$ we set

$$Af = \mathcal{F}^{-1}\langle \tau, \xi \rangle^2 \mathcal{F} f.$$

The operator A is closed in $L^p(\mathbf{R}^{n+1})$ with the domain $\mathcal{D}(A) = H_p^{1,2}(\mathbf{R}^{n+1})$. By Mikhlin's lemma the operator norm in L^p of the pure imaginary power A^{iy} is bounded by a constant multiple of $e^{\gamma|y|}$ for some $\gamma > 0$. A standard argument (see, e.g., [GS, §6]) yields

$$H_p^{s,2s}(\mathbf{R}^{n+1}) = \mathcal{D}(A^s) = [L^p(\mathbf{R}^{n+1}), H_p^{1,2}(\mathbf{R}^{n+1})]_s.$$

(ii) We observe through Mikhlin's lemma that

$$\begin{aligned} ||Af||_{L^p} &\le C||(\partial_t - \Delta + 1)f||_{L^p} \\ &\le C(||\partial_t f||_{L^p} + ||\nabla^2 f||_{L^p} + ||f||_{L^p}) \\ &\le C||Af||_{L^p}. \end{aligned}$$

(iii) We may assume $t_0 = 0$. It is well known that there is a continuous extension e_1: $H_p^2(\Omega) \to H_p^2(\mathbf{R}^n)$. For $f \in H_p^{1,2}(D)$ we set

$$\tilde{f}(t,x) = \begin{cases} (e_1 f)(-t,x) & \text{for } (t,x) \in (-t_1,0) \times \mathbf{R}^n, \\ (e_1 f)(2t_1 - t,x) & \text{for } (t,x) \in (t_1,2t_1) \times \mathbf{R}^n, \end{cases}$$

so that $\tilde{f}$ is defined on $(-t_1,2t_1) \times \mathbf{R}^n$. We then take $\varphi \in C_0^\infty((-t_1,2t_1) \times \mathbf{R}^n)$ so that $\varphi \equiv 1$ on D. By the characterization of $H_p^{1,2}$ norm in (ii) we observe that the operator $ef := \varphi \tilde{f}$ is continuous from $H_p^{1,2}(D)$ to $H_p^{1,2}(\mathbf{R}^{n+1})$. Clearly $ef = f$ on D.

(iv) Interpolating $e : H_p^{1,2}(D) \to H_p^{1,2}(\mathbf{R}^{n+1})$ and $e : L^p(D) \to L^p(\mathbf{R}^{n+1})$, we observe that e is a bounded linear operator

$$e : [L^p(D), H_p^{1,2}(D)]_s \to [L^p(\mathbf{R}^{n+1}), H_p^{1,2}(\mathbf{R}^{n+1})]_s = H_p^{s,2s}(\mathbf{R}^{n+1}).$$

Since the restriction $r : H_p^{s,2s}(\mathbf{R}^{n+1}) \to H_p^{s,2s}(D)$ is continuous, there is a continuous inclusion from $[L^p(D), H_p^{1,2}(D)]_s$ to $H_p^{s,2s}(D)$.

Interpolating $r : H_p^{1,2}(\mathbf{R}^{n+1}) \to H_p^{1,2}(D)$ and $r : L^p(\mathbf{R}^{n+1}) \to L^p(D)$, we observe that $H_p^{s,2s}(D)$ is continuously included in $[L^p(D), H_p^{1,2}(D)]_s$ since r is surjective and the topology of $H_p^{s,2s}(D)$ is strongest such that r is continuous. This proves the identity of (iv).

(v) We take $D = (0,T) \times \Omega$ such that Ω contains the closed rectangle R. For $f \in H_p^{1/2,1}(Q)$ the mapping

$$j : f \mapsto f|_D$$

is continuous from $H_p^{1/2,1}(Q)$ to $H_p^{1/2,1}(D)$ since D is bounded.

Note that $H_p^{1,2}(D) \subset H_p^1(D)$ by (ii), where $H_p^1(D)$ denotes an isotropic L^p Sobolev space of order one. By (iv) we observe that $H_p^{1/2,1}(D) \subset H_p^{1/2}(D)$ since

$$H_p^{1/2}(D) = [L^p(D), H_p^1(D)]_{1/2}$$

(cf. [Tri, §4.3.1]). The Sobolev inequality implies

$$H_p^{1/2}(D) \subset C^\mu(\overline{D})$$

with $\mu = 1/2(n+1) - 1/p$ provided that $p > 2(n+1)$ (see [Tri p. 327, §4.6.1]). Thus $H_p^{1/2,1}(D)$ is continuously embedded in $C^\mu(\overline{D})$. The mapping j now gives a continuous mapping

$$H_p^{1/2,1}(Q) \to C^\mu(\overline{D})$$

such that

$$jf = f \quad \text{on } (0,T) \times R.$$

This implies that the inclusion

$$H_p^{1/2,1}(Q) \subset C^\mu(\overline{Q})$$

is continuous.

(vi) For $u \in \mathcal{H}_p(Q)$ let $v \in H_p^{1/2,1}$ be an extension of u such that

$$||v||_{H_p^{1/2,1}} \leq 2||u||_{\mathcal{H}_p(Q)}.$$

By Mikhlin's lemma we have

$$||\frac{\partial u}{\partial x_j}||_{L^p(Q)} \leq C||v||_{H_p^{1/2,1}}, \qquad j = 1, \ldots, n$$

with $C = C(n,p)$. These two inequalities yield (vi). □

Acknowledgments. We are grateful to Professor Hitoshi Ishii and Professor Hisashi Okamoto for criticism of solutions of the transport equations.

REFERENCES

[Al] G. ALLAIN, *Small-time existence for the Navier–Stokes equations with a free surface*, Appl. Math. Optim., 16 (1987), pp. 37–50.

[AF] J. P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, Basel, and Berlin, 1990.

[BP] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modèl Math. Anal. Numer., 21 (1987), pp. 557–579.

[Be1] J. T. BEALE, *The initial value problem for the Navier–Stokes equations with a free surface*, Comm. Pure Appl. Math., 34 (1981), pp. 359–392.

[Be2] ———, *Large-time regularity of viscous surface waves*, Arch. Rational Mech. Anal., 84 (1984), pp. 307–352.

[Cam] S. CAMPANATO, *L^p regularity for weak solutions of parabolic systems*, Ann. Scuola Norm. Sup. Pisa, 7 (1980), pp. 65–85.

[CGG1] Y.-G. CHEN, Y. GIGA, AND S. GOTO, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geometry, 33 (1991), pp. 749–786.

[CGG2] ———, *Analysis Toward Snow Crystal Growth*, Proc. of International Symposium on Functional Analysis and Related Topics, S. Koshi, ed., World Scientific, Singapore, New Jersey, London, and Hong Kong, 1991, pp. 43–57.

[CIL] 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., 27 (1992), pp. 1–67.

[De] I. V. DENISOVA, *A priori estimates of the solution of a linear time-dependent problem connected with the motion of a drop in a fluid medium*, Trudy Mat. Inst. Steklov., 188 (1990), pp. 3–21. (English translation: Proc. Steklov Inst. Math. 3 (1991), pp. 1–24.)

[DeSol] I. V. DENISOVA AND V. A. SOLONNIKOV, *Solvability of the linearized problem on the motion of a drop in a fluid flow*, Zap. Nauchn. Semin. Lenigrad. Otdel. Mat. Inst. Steklov. 171 (1989), pp. 53–65. (English translation in J. Soviet Math. 56 (1991), pp. 2309–2316.)

[ES] L. C. EVANS AND J. SPRUCK, *Motion of level sets by mean curvature* I, J. Differential Geometry, 33 (1991), pp. 635–681.

[ESou] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaccs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[GGI] Y. GIGA, S. GOTO, AND H. ISHII, *Global existence of weak solutions for interface equations coupled with diffusion equations*, SIAM J. Math. Anal., 23 (1992), pp. 821–835.

[GS] Y. GIGA AND H. SOHR, *On the Stokes operator in exterior domains*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 36 (1989), pp. 103–130.

[GY] Y. GIGA AND Z. YOSHIDA, *A dynamic free-boundary problem in plasma physics*, SIAM J. Math. Anal., 21 (1990), pp. 1118–1138.

[Ish] H. ISHII, *Perron's method for Hamilton–Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.

[KL] R. V. KOHN AND R. LIPTON, *The effective viscosity of a mixture of two Stokes fluids*, in Advances in Multiphase Flow and Related Problems, G. Papanicolaou, ed., SIAM, Philadelphia, PA, pp. 123–135, 1986.

[MS] V. G. MAZ'YA AND T. O. SHAPONIKOVA, *Theory of Multipliers in Spaces of Differentiable Functions*, Pitman, Boston, MA, 1985.

[Sol 1] V. A. SOLONNIKOV, *Unsteady motions of a finite isolated mass of a self-gravitating fluid*, Algebra i Analiz, 1 (1989), pp. 207–246. (In Russian.) (English translation in Leningrad Math. 1 (1990), pp. 227–276.)

[Sol 2] ———, *On the evolution of an isolated volume of viscous incompressible capillary fluid for large values of time*, Vestnik Leningrad. Univ. Math., 20 (1987), pp. 52–58. (English translation from Vestnik Leningrad. Univ. 20 (1987), pp. 49–55.)

[Sol 3] ———, *Unsteady motion of a finite mass of fluid, bounded by a free surface*, J. Soviet Math., 40 (1988), pp. 672–686. (English translation from Zap. Nauchn. Sem. LOMI 152 (1986), pp. 137–157.)

[Sol 4] ———, *On the transient motion of an isolated volume of viscous incompressible fluid*, Math. USSR-Izv. 31 (1988), pp. 381–405. (English translation from Izv. Akad. Nauk SSSR Ser. Mat. 51 (1987), pp. 1065–1087.)

[Sol 5] ———, *Solvability of a problem on the motion of a viscous incompressible fluid bounded by a free surface*, Math. USSR-Izv, 11 (1977), pp. 1323–1358. (English translation from Izv. Akad. Nauk SSSR Ser. Mat., 41 (1977), pp. 1388–1424.)

[Sol 6] ———, *Solvability of the problem of evolution of an isolated volume of viscous incompressible capillary fluid*, J. Soviet Math. 32 (1986), pp. 223–228. (English translation from Zap. Nauchn. Sem. LOMI, 140 (1984), pp. 179–186.)

[Sy] G. SYLVESTER, *Large-time existence of small viscous surface waves without surface tension*, Comm. Partial Differential Equations, 15 (1990), pp. 823–903.

[T] S. TAKAHASHI, *On global weak solutions of the nonstationary two-phase Navier-Stokes flow*, Adv. Math. Sci. Appl., to appear.

[Tana] N. TANAKA, *Global existence of two phase non-homogeneous viscous incompressible fluid flow*, Comm. Partial Differential Equations, 18 (1993), pp. 41–81.

[Tani 1] A. TANI, *On the evolution equations of the two-phase compressible viscous capillary fluids*, in preparation. (Announcement in: *Free boundary problems for general fluids.* Kokyuroku. RIMS. Kyoto Univ. 698, 146–170 (1988).)

[Tani 2] ———, *Multi-phase free boundary problem for the equation of motion of general fluids*, Comm. Math. Univ. Carolinae, 26 (1985), pp. 201–208.

[Tani 3] ———, *Two-phase free boundary problem for compressible viscous fluid motion*, J. Math. Kyoto Univ., 21 (1981), pp. 839–859.

[Tri] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

[Ya] M. YAMAZAKI, *A quasi-homogeneous version of paradifferential operators*, I. *Boundedness on spaces of Besov type*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 33 (1986), pp. 131–174.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 894–914, May 1994

005

A REMARK ON QUASI-STATIONARY APPROXIMATE INERTIAL MANIFOLDS FOR THE NAVIER–STOKES EQUATIONS*

D. A. JONES† AND E. S. TITI‡

This paper is dedicated to the memory of Professor Charles J. Amick.

Abstract. The two dimensional Navier–Stokes equations with time-dependent external body forces is considered. Under appropriate assumptions on the temporal properties of the forcing term the authors are able to construct a time-dependent deterministic approximate inertial manifold. It is shown that all solutions converge exponentially fast to a thin neighborhood of this manifold. If the forcing term is too oscillatory in time, it is shown by example that the techniques used in the construction of certain approximate inertial manifolds for the autonomous case, in general, do not extend to the time-dependent case. Also it is shown that if the forcing term is time-independent and spatially smooth (Gevrey class), then the global attractor lies exponentially close to the linear manifold spanned by the first m eigenfunctions of the Stokes operator, provided m is large enough.

Key words. Navier–Stokes equations, approximate inertial manifolds, nonlinear Galerkin methods

AMS subject classifications. 35A40, 35Q10, 65N30, 76D05

1. Introduction.

In recent years there have been extensive, rigorous studies of the long-time behavior of the Navier–Stokes equations (NSE) in an attempt to understand the phenomenon of turbulence. All these studies have indicated that the long-time behavior of the dynamical system generated by the solutions has a finite number of "degrees of freedom."

For instance, the existence of a finite number of *determining Fourier modes* (cf. [25] and [23]), the existence of a finite number of *determining nodes* (cf. [33], [35], [46], and [47]), and the existence of a finite-dimensional *global (universal) attractor* (see, e.g., [2], [7], [11], [8], [36], [40], and [67] and the references therein). In the latter case, most of the studies were made under the assumption that the forcing term $f(t) \equiv f$ is time independent (for related topics with time-periodic forces see, e.g., [37] and [61]). In this case, when the force is time-independent, the solution $u(t)$ is given by a nonlinear semigroup $u(t) = S(t)u_0$, and the global attractor is characterized as the maximal bounded invariant set under $S(t)$ (i.e., $S(t)\mathcal{A} = \mathcal{A}$ for all $t \in \mathbb{R}$; see, e.g., [40] and [66], and the references therein).

The existence of a finite number of determining Fourier modes suggests that for $m \gg 1$ the long-time dynamics of the high modes, say q, is, roughly speaking, determined by the dynamics of the lower ones, say p. Hence, it is natural to search for a "global function" that gives q in terms of p, asymptotically in time. Most recently, the theory of *inertial manifolds* provided sufficient conditions for the existence of such a function $q = \Phi(p)$ (cf. [27] and [28], and [3], [9], [10], [24], [29], and [55]). More precisely, an inertial manifold (IM) for a dissipative evolution partial differential equation is a smooth finite-dimensional manifold in the phase space, which is positively invariant under the solution operator, and which is uniformly attracting every bounded subset of phase space in an exponential rate. The problem of existence of an IM for the NSE is still open. However, such a manifold exists for numerous interesting

* Received by the editors May 4, 1992; accepted for publication (in revised form) April 6, 1993.
† Department of Mathematics, University of California, Irvine, California 92717.
‡ Center of Applied Mathematics, Cornell University, Ithaca, New York 14853. The work of this author was supported by Air Force Office of Scientific Research, National Science Foundation grant DMS-8915672.

partial differential equations (see, e.g., [9] and the references therein). It is clear that if the IM exists, then it must contain the global attractor. Moreover, the reduction of the partial differential equation to the IM yields an ordinary differential system, which is called the inertial form.

In particular, we consider an abstract evolutionary equation on a Hilbert space H (see §2) of the form

$$\frac{du}{dt} + Au + R(u) = f.$$

We denote by P_m the orthogonal projection onto the span of the first m eigenvectors of the linear dissipative operator A, $P_m H = \text{span}\{\varphi_1, \dots, \varphi_m\}$, $Q_m = I - P_m$, and $p = P_m u, q = Q_m u$. Then the evolution equation is equivalent to the system

$$\frac{dp}{dt} + Ap + P_m R(p+q) = P_m f,$$
$$\frac{dq}{dt} + Aq + Q_m R(p+q) = Q_m f.$$

If the IM is given by $\mathcal{M} = \text{Graph}(\Phi)$ and $u(t) = p(t) + \Phi(p(t))$, the inertial form is given by

$$\frac{dp}{dt} + Ap + P_m R(p + \Phi(p)) = P_m f, \qquad p \in P_m H.$$

Although the existence of an IM for the NSE is still unknown, the NSE does have an inertial form for periodic boundary conditions with special periods [52]. That is, the dynamics on the global attractor is given by the dynamics of an ordinary differential system.

In general, one does not have an explicit form for the IM when it exists, except in certain cases (see, e.g., [3]). One must therefore approximate it. Even if the IM does not exist, the theory suggests looking for a global function, Φ_{app}, whose graph, $\mathcal{M}_{app} := \text{Graph}(\Phi_{app})$, in phase space approximates the global attractor. Such manifolds are called approximate inertial manifolds (see [6], [12], [15], [20]–[22], [29], [32], [44], [45], [48], [57], [59], [68], [69], and [71]). One then studies, in either case, the approximate inertial form (AIF)

$$\frac{dp}{dt} + Ap + P_m R(p + \Phi_{app}(p)) = P_m f, \qquad p \in H_m, \tag{1.1}$$

or a variant of (1.1) (see [12], [13], [45], and [58]). If one puts $\Phi_{app}(p) = 0$ in (1.1), then the AIF is just the usual Galerkin scheme. For this reason the study of AIMs and the associated approximate inertial forms are also called nonlinear Galerkin methods. Thus, if one can model, asymptotically in time, the small scales, q, as a function, $\Phi_{app}(p)$, of the large scales, p, in a nontrivial way, (1.1) may better reflect the dynamics of the original PDE compared with the standard Galerkin scheme. Indeed, the nonlinear Galerkin methods have yielded new numerical schemes that may be appropriate for approximating solutions for long intervals of time. Computational results using (1.1) are encouraging. They have shown improved stability, accuracy, and a significant gain in computing time (see [5], [19], [43], [56], [44], and [45]). Also see [18], [38], and [62] for other computational and stability aspects of these schemes.

However, it is important to verify that these schemes would reflect and predict the correct qualitative dynamical features of the equation and its solutions, such as

bifurcations, dissipation, stability, hyperbolicity, etc. One of the essential features of the PDEs under study is that they are dissipative. We remark, however, that even to preserve this basic property in the AIF, one needs to be careful. See, for example, [17], [18], [35], and [45].

Our goal in this paper is to construct an AIM for time-dependent forces. The theory of attractors and IMs was originally developed for autonomous systems-time-independent forces. Indeed, the idea of AIMs is to approximate the global attractor. When the force is time-dependent, and no assumptions are made about the asymptotic behavior of $f(t)$, the system may not possess a global attractor. We will show, however, that we can extend the properties of several AIMs to time-dependent forces under appropriate assumptions on the force, $f(t)$. Our method is similar to the one used in [65], which studies the AIM in [69] and [71] for certain reaction-diffusion equations with time-dependent forces.

We organize the paper as follows. In §2 we present the abstract framework and recall certain known estimates for the NSE that we will use later. In §3 we study a particular AIM and its inertial form; namely, the flat linear space, $\Phi_{app}(p) = 0$, and its corresponding inertial form that is the usual Galerkin scheme. However, we show that if the forcing term is in the Gevrey class, then after a sufficiently large time, the orbits of the solutions to the NSE remain at a distance of the order $\lambda_{m+1}^{-1/2} e^{-\sigma \lambda_{m+1}^{1/2}}$ from the flat manifold, where λ_{m+1} is the $(m+1)$th eigenvalue of the linear Stokes operator. The consequences of this is that the AIMs given in [21], [29], [57], [68], and [69] lead to algebraic improvements in the rate of convergence of the distance solutions are attracted to their respective manifolds over the flat manifold. Moreover, based on the work of [13] one can show exponential convergence of the Galerkin and nonlinear Galerkin systems in this case. We would like to add that these results are not restricted to the NSE and can be extended easily to many other equations such as the Kuramoto–Sivashinsky equation, the complex Ginzburg–Landau equation, and the convection in porous media (see, e.g., [14], [54], [60], and [72]). In §4 we consider time-dependent forces. We show by example that if the forcing term is too oscillatory in the time variable, then the assumptions used in the AIMs in [21], [29], [57] [68], and [69], for the autonomous case do not carry over to the time-dependent case. That is, dq/dt is relatively small with respect to the other terms in the equation for q above for $m >> 1$. In §5 we assume that the forcing term in Hölder continuous in the time variable. We are able to construct a time-dependent AIM in this case, and estimate the error resulting in the associated nonlinear Galerkin method.

We remark that our manifold is obtained from an algebraic closure resulting from a decomposition of the solutions with respect to the spatial scales and by negelecting the time derivative of the small scales. Perhaps a more justified method, given an ensemble of solutions, would be to apply the Karhunen–Loève procedure [1], [63] to the time derivatives (see [64]). In this way one could decompose the system into two parts: one finite-dimensional where the time derivatives are important, and the other in which the time derivatives could be neglected. This latter system would then give a way to close the system algebraically. We investigate this in [49].

2. Functional setting and preliminary results. The two-dimensional Navier–Stokes equations for a viscous incompressible fluid filling a region Ω are of the form

$$\text{(2.1)}\qquad \begin{cases} \dfrac{\partial u}{\partial t} - \nu\Delta u + (u\cdot\nabla)u + \nabla p = f, \\ \nabla\cdot u = 0, \\ u(x,0) = u_0(x). \end{cases}$$

$f = f(x,t)$, the external body force, and $\nu > 0$, the kinematic viscosity, are given; $u = u(x,t)$, the velocity vector, and $p = p(x,t)$, the pressure, are the unknowns.

We supplement (2.1) with two types of boundary conditions: the first is nonslip or homogeneous Dirichlet boundary condition,

$$\text{(2.2)}\qquad u|_{\partial\Omega} = 0;$$

the other is periodic boundary conditions. $\Omega = (0,L)\times(0,L)$ and

$$\text{(2.3)}\qquad \begin{aligned} u(x_1,x_2,t) &= u(x_1,x_2+L,t), \\ u(x_1,x_2,t) &= u(x_1+L,x_2,t) \end{aligned}$$

for all $(x_1,x_2)\in\mathbb{R}^2$. We assume in the latter case that the integrals of u and f on Ω vanish at all time.

In the case (2.2) we denote

$$\text{(2.4)}\qquad \mathcal{V} = \left\{v\in (C_0^\infty(\Omega))^2,\ \operatorname{div} v = 0\right\},$$

and in the case (2.3) we denote

$$\mathcal{V} = \{\ u : \mathbb{R}^2 \mapsto \mathbb{R}^2,\ \text{vector-valued trigonometric polynomials with period } L, \nabla\cdot u = 0,\ \text{and } \textstyle\int_\Omega u dx = 0\}.$$

We suppose that from now on in the case (2.2), Ω has a sufficiently smooth boundary. In both cases we set

$$\begin{aligned} H &= \text{the closure of } \mathcal{V} \text{ in } (L^2(\Omega))^2, \\ V &= \text{the closure of } \mathcal{V} \text{ in } (H^1(\Omega))^2, \end{aligned}$$

where $H^l(\Omega)$ $(l = 1,2,\ldots)$ denote the usual L^2-Sobolev spaces. H is a Hilbert space with the inner product and norm

$$(u,v) = \int_\Omega u(x)\cdot v(x)dx, \qquad |u| = \left(\int_\Omega |u(x)|^2 dx\right)^{1/2},$$

respectively, and $u(x)\cdot v(x)$ is the usual Euclidean scalar product. Thanks to the Poincaré inequality, V is also a Hilbert space with inner product and norm

$$((u,v)) = \sum_{i,j}^{2}\int_\Omega \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}dx, \qquad \|v\|^2 = \sum_{i,j}^{2}\int_\Omega \left|\frac{\partial v_i}{\partial x_j}\right|^2 dx,$$

respectively. Let P denote the orthogonal projection in $L^2(\Omega) \times L^2(\Omega)$ onto H. We denote by A the Stokes operator

$$Au = -P\Delta u$$

(notice that in the periodic case $Au = -\Delta u$) and the bilinear operator

$$B(u,v) = P((u \cdot \nabla)v)$$

for all u, v in $\mathcal{D}(A) = V \cap (H^2(\Omega) \times H^2(\Omega))$. We recall that the operator A is a positive self-adjoint operator with compact inverse. Thus there exists a complete orthonormal set φ_j of eigenfunctions of A such that $A\varphi_j = \lambda_j \varphi_j$ and $0 < \lambda_1 \leq \lambda_2 \leq \cdots$.

Now the NSE, (2.1), is equivalent to the differential equation in H,

$$\frac{du}{dt} + \nu Au + B(u,u) = f, \tag{2.5}$$

where from now on $f = Pf$, and it is assumed that f satisfies $f \in L^\infty((0,\infty); H)$. That is, $\sup_{t \geq 0} |f(t)| < f_\infty$. (For details see, e.g., [7], [53], or [66].) For questions related to existence, uniqueness, and regularity of solutions the reader is referred to [7], [30], [42], [41], [50], [53], [66], and the references therein.

It is also well known that there exists constants M_0, M_1, which depend only on ν, f_∞, λ_1, such that for every solution $u(t)$ of (2.5) there is a time T_* depending on u_0, ν, f_∞, λ_1, such that

$$|u(t)| \leq M_0 \quad \text{and} \quad \|u(t)\| \leq M_1 \tag{2.6}$$

for $t \geq T_*$ (see [7], [30], and [66]). Some of our results will require $|Au(t)| \leq M_2$ for sufficiently large t. Sufficient conditions for this bound would be, for example, $\sup_{t \geq 0} \|f(t)\| < \infty$, or $f(t)$ is analytic on a strip containing $\mathbb{R}^+$ and $\sup_{t \geq 0} |f(t)| < \infty$ (see [7], [30], and [66]) or $\sup_{t \geq 0} |f'(t)| < \infty$, (see [39], [53], and [66]). We remark also that the existence of the constant M_1 such that (2.6) holds is not known in the three-dimensional case. Therefore, in order to extend the results of this work to the three-dimensional case we need to consider the approximation of invariant sets which are bounded in V.

We recall the following inequalities that are satisfied by $B(u,v)$ (cf. [7], [53], and [66]).

$$|(B(u,v),w)| \leq c_1 |u|^{1/2} \|u\|^{1/2} \|v\|^{1/2} |Av|^{1/2} |w| \qquad \begin{matrix} \forall u \in V, v \in \mathcal{D}(A), \\ w \in H, \end{matrix} \tag{2.7}$$

$$|(B(u,v),w)| \leq c_2 |u|^{1/2} \|u\|^{1/2} \|v\| |w|^{1/2} \|w\|^{1/2} \quad \forall u, v, w \in V. \tag{2.8}$$

We recall the following estimates from [70]:

$$|(B(u,v),w)| \leq c_3 \|u\| \|v\| |w| \left[1 + \log\left(\frac{\|w\|}{|w|\lambda_1^{1/2}}\right)\right]^{1/2} \quad \forall u, v, w \in V, \tag{2.9}$$

$$|(B(u,v),w)| \leq c_4 |u| \|v\| \|w\| \left[1 + \log\left(\frac{\|u\|}{|u|\lambda_1^{1/2}}\right)\right]^{1/2} \quad \forall u, v, w \in V, \tag{2.10}$$

$$|(B(u,v),Aw)| \le c_5\|u\|\|v\||Aw|\left[1+\log\left(\frac{|Av|}{\|v\|\lambda_1^{1/2}}\right)\right]^{1/2} \qquad \begin{matrix}\forall u\in V,\\ v,w\in\mathcal{D}(A).\end{matrix} \tag{2.11}$$

We also have the estimate

$$|B(u,v)| \le c_6\|u\|_{L^\infty(\Omega)}\|v\| \quad \forall u\in\mathcal{D}(A), v\in V.$$

Then we may either use from [4] that

$$\|u\|_{L^\infty(\Omega)} \le c_7\|u\|\left[1+\log\left(\frac{|Au|}{\lambda_1\|u\|}\right)\right]^{1/2} \quad \forall u\in\mathcal{D}(A) \tag{2.12}$$

or Agmon's inequality,

$$\|u\|_{L^\infty(\Omega)} \le c_8|u|^{1/2}|Au|^{1/2} \quad \forall u\in\mathcal{D}(A), \tag{2.13}$$

to obtain

$$|B(u,v)| \le c_9\|u\|\|v\|\left[1+\log\left(\frac{\|Au\|}{\|u\|\lambda_1^{1/2}}\right)\right]^{1/2} \quad \forall u\in\mathcal{D}(A), v\in V, \tag{2.14}$$

$$|B(u,v)| \le c_{10}|u|^{1/2}|Au|^{1/2}\|v\| \quad \forall u\in\mathcal{D}(A), v\in V, \tag{2.15}$$

respectively.

In addition, the operator B enjoys the following fundamental property:

$$(B(u,v),w) = -(B(u,w),v). \tag{2.16}$$

(See, e.g., [7], [53], and [66].)

3. The Galerkin approximation revisited. Denote by P_m the orthogonal projection of H onto $H_m = \operatorname{span}\{\varphi_1,\dots,\varphi_m\}$, $Q_m = I - P_m$, and $p = P_m u, q = Q_m u$. Then (2.5) is equivalent to the system

$$\frac{dp}{dt} + \nu Ap + P_m B(p+q,p+q) = P_m f, \tag{3.1}$$

$$\frac{dq}{dt} + \nu Aq + Q_m B(p+q,p+q) = Q_m f. \tag{3.2}$$

Before turning to the time-dependent forces we examine the case where f is time-independent and smooth. We consider only the case (2.3), the periodic boundary conditions, in this section. More specifically, we suppose that f is in the Gevrey class. This means that for some $\sigma > 0$, $f \in \mathcal{D}(e^{\sigma A^{1/2}})$. That is,

$$|e^{\sigma A^{1/2}} f|^2 = \sum_{j=1}^{\infty} e^{2\sigma\lambda_j^{1/2}}|f_j|^2 < \infty,$$

where $f = \sum_{j=1}^{\infty} f_j\varphi_j$. We wish to reconsider the classical Galerkin method under this assumption on f.

It is shown in [20] and [21] that on the attractor

$$|q(t)| \le K_0 L_m^{1/2}\lambda_{m+1}^{-1} \quad \forall t\in\mathbb{R},$$

where $L_m = (1 + \log(\lambda_m/\lambda_1))$ and f is assumed to be in $L^2(\Omega)$ only. We remark that this estimate has been shown by example to be sharp, asymptotically in m, as $m \to \infty$, up to the logarithmic term [69], [71]. This suggests replacing the mapping Φ_{app} in (1.1) by zero. The AIM is just the linear space $P_m H$, and the approximate inertial form is the standard Galerkin approximation,

$$\frac{du_m}{dt} + \nu A u_m + P_m B(u_m, u_m) = P_m f, \qquad p \in H_m.$$

We begin by recalling a result from [34].

THEOREM 3.1. *Suppose $u_0 \in \mathcal{D}(A^{1/2})$ and f is given in $\mathcal{D}(e^{\sigma A^{1/2}})$ for some $\sigma > 0$. Then there exists constants σ_1, T_1 that depend only on u_0 through $|A^{1/2}u_0|$, such that the solution, $u(t)$, of (2.5) satisfies*

$$\left|e^{\sigma_1 A^{1/2}} A^{1/2} u(t)\right| \le R_1 \quad \forall t \ge T_1, \tag{3.3}$$

where R_1 depends only on M_1.

We remark that similar results have been obtained in [41] and [50]. The bounds (2.7), (2.8), (2.14), and (2.15) involving the bilinear term $B(u,v)$ given above have analogies in the Gevrey class (see the appendix). For example, (2.15) becomes

$$\left|e^{\sigma A^{1/2}} B(u,v)\right| \le c_{10} \left|e^{\sigma A^{1/2}} u\right|^{1/2} \left|e^{\sigma A^{1/2}} Au\right|^{1/2} \left|e^{\sigma A^{1/2}} A^{1/2} v\right| \tag{3.4}$$

for all $u \in \mathcal{D}(Ae^{\sigma A^{1/2}}), v \in \mathcal{D}(A^{1/2}e^{\sigma A^{1/2}})$. The next result shows that the flat manifold H_m attracts the solutions of (2.5) to an exponentially thin neighborhood. In what follows, $K_i, i = 1, 2, 3, \ldots$, will denote appropriately chosen constants that depend only on ν, R_1, f, and λ_1.

PROPOSITION 3.2. *Let f be given in the Gevrey class for some $\sigma > 0$ and let $m > 0$. Then for t sufficiently large, any orbit of (2.5) (periodic boundary conditions) satisfies*

$$\operatorname{dist}_V(u(t), H_m) \le \sqrt{2K_1} \left(|e^{\sigma A^{1/2}} Q_m f| + R_1^2 L_m^{1/2} + \frac{R_1^3}{\nu} \right) \frac{e^{-\sigma_2 \lambda_{m+1}^{1/2}}}{\nu \lambda_{m+1}^{1/2}},$$

where $\sigma_2 = \sigma_1/2$ and σ_1 is given by Theorem 3.1.

Proof. We have that

$$\frac{dq}{dt} + \nu A q + Q_m B(u,u) = Q_m f,$$

where $q(t) = Q_m u(t)$ and $u(t)$ solves (2.5). Taking the inner product with $e^{2\sigma A^{1/2}} Aq$ (to make this completely rigorous we should consider a Galerkin approximation based on the eigenfuctions of A and then pass to the limit, since we do not know a priori that $|e^{2\sigma A^{1/2}} Aq|$ is bounded), we find

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|e^{\sigma A^{1/2}} q\|^2 + \nu |Ae^{\sigma A^{1/2}} q|^2 \le\ & |e^{\sigma A^{1/2}} Q_m f| |Ae^{\sigma A^{1/2}} q| \\ & + |(e^{\sigma A^{1/2}} B(p,u), Ae^{\sigma A^{1/2}} q)| \\ & + |(e^{\sigma A^{1/2}} B(q,u), Ae^{\sigma A^{1/2}} q)|. \end{aligned}$$

Using the Cauchy–Schwarz inequality, (2.14), and (2.15) appropriately modified as (3.4), we majorize the right-hand side of the last inequality by

$$|e^{\sigma A^{1/2}}Q_m f||Ae^{\sigma A^{1/2}}q| + c_9\|e^{\sigma A^{1/2}}p\|\,\|e^{\sigma A^{1/2}}u\|L^{1/2}|Ae^{\sigma A^{1/2}}q| + c_{10}|e^{\sigma A^{1/2}}q|^{1/2}|Ae^{\sigma A^{1/2}}q|^{3/2}\|e^{\sigma A^{1/2}}u\|.$$

After applying Young's inequality and (3.3) we bound this quality by

$$\frac{\nu}{2}\left|Ae^{\sigma A^{1/2}}q\right|^2 + \frac{6}{\nu}\left|e^{\sigma A^{1/2}}Q_m f\right|^2 + \frac{6}{\nu}c_9^2 R_1^4 L_m + \frac{27^3 c_{10}^4 R_1^6}{32\nu^3\lambda_{m+1}}.$$

Thus,

$$\frac{d}{dt}\|e^{\sigma A^{1/2}}q\|^2 + \nu\lambda_{m+1}\|e^{\sigma A^{1/2}}q\|^2 \le K_1\left(\frac{1}{\nu}|e^{\sigma A^{1/2}}Q_m f|^2 + \frac{R_1^4 L_m}{\nu} + \frac{R_1^6}{\nu^3}\right)$$

and we conclude that

$$\|e^{\sigma A^{1/2}}q(t)\|^2 \le \|e^{\sigma A^{1/2}}q(T_1)\|^2 e^{-\nu\lambda_{m+1}(t-T_1)} + \frac{K_1}{\nu\lambda_{m+1}}\left(\frac{1}{\nu}|e^{\sigma A^{1/2}}Q_m f|^2 + \frac{R_1^4 L_m}{\nu} + \frac{R_1^6}{\nu^3}\right).$$

Then after some further time, depending on R_1, ν, λ_{m+1}, the term involving t becomes negligible, and we obtain

$$\|e^{\sigma A^{1/2}}q(t)\|^2 \le \frac{2K_1}{\nu\lambda_{m+1}}\left(\frac{1}{\nu}|e^{\sigma A^{1/2}}Q_m f|^2 + \frac{R_1^4 L_m}{\nu} + \frac{R_1^6}{\nu^3}\right). \tag{3.5}$$

We have

$$\|e^{\sigma A^{1/2}}q(t)\|^2 = \sum_{j=m+1}^{\infty} e^{\sigma\lambda_j^{1/2}}\lambda_j^{1/2}|q_j(t)|^2 \ge e^{\sigma\lambda_{m+1}^{1/2}}\|q(t)\|^2,$$

where $q(t) = Q_m u(t) = \sum_{j=m+1}^{\infty} q_j(t)\varphi_j$. Combining this last inequality with (3.5) and

$$\mathrm{dist}_V(u(t), H_m) \le \|(p(t)+q(t)) - p(t)\| = \|q(t)\|,$$

the result follows. □

A consequence of this last result is that under the assumptions that f is in the Gevrey class, the AIMs given in [21], [29], [57], [68], and [69] will lead to algebraic improvements in the upper bounds of the rates of convergence over the flat manifold $\mathcal{M} = P_m H$. That is, since under these conditions the Fourier coefficients of the solutions decay exponentially, there may be little gain in approximating $q(t) = \Phi_{app}(p(t))$ with any other choice than $\Phi_{app} \equiv 0$ (see, e.g., [38]). We illustrate this for the AIM given in [21].

It is known [34] that if $f \in \mathcal{D}(e^{\sigma A^{1/2}})$, for some $\sigma > 0$, then the solution of the complexified equation (in time) of (2.5), $u(t)$, is analytic with values in $\mathcal{D}(A^{1/2}e^{\sigma A^{1/2}})$. Furthermore, the region of analyticity $\{\mathrm{Re}\, z \ge K_2, \quad \mathrm{Im}\, z \le \sqrt{2}/2K_2\}$ is independent of the individual solution. It follows from (3.5) and the Cauchy integral formula that

$$\left|e^{\sigma A^{1/2}}\frac{dq}{dt}\right| \le \frac{K_3}{\lambda_{m+1}}. \tag{3.6}$$

Consider the AIM given in [21],

$$\Phi_1(p) = Q_m(\nu A)^{-1}(f - B(p,p)). \tag{3.7}$$

We set $\mathcal{M}_1 = \text{Graph}(\Phi_1)$.

THEOREM 3.3. *Let f be given in $\mathcal{D}(e^{\sigma A^{1/2}})$ for some $\sigma > 0$. Then for t sufficiently large, any orbit of* (2.5) *(periodic boundary conditions) satisfies*

$$\text{dist}_V(u(t), \mathcal{M}_1) \leq K_4 \lambda_{m+1}^{-1} e^{-\sigma_2 \lambda_{m+1}^{1/2}}.$$

Proof. Set $u_1(t) = p(t) + q_1(t)$, where $p(t)$ solves (3.1) and $q_1(t) = \Phi_1(p(t))$, where Φ_1 is given by (3.7). Then $u_1(t) \in \mathcal{M}_1$ for $t \geq 0$. Subtracting (3.2) from (3.7) and setting $\Theta_1 = q - q_1$ we have

$$\nu A \Theta_1 = Q_m B(p,q) + Q_m B(q,p) + Q_m B(q,q) + \frac{dq}{dt}.$$

Multiply this last equation by $e^{\sigma_2 A^{1/2}}$ and take the inner product with $e^{\sigma_2 A^{1/2}}\Theta_1$ using equations (2.8) and (2.14) to obtain

$$\begin{aligned}
\|e^{\sigma_2 A^{1/2}}\Theta_1\| \leq\ & \frac{c_9 L_m^{1/2}}{\nu} \|e^{\sigma_2 A^{1/2}} p\| \, \|e^{\sigma_2 A^{1/2}} q\| \, |e^{\sigma_2 A^{1/2}}\Theta_1| \\
& + \frac{c_2}{\nu} |e^{\sigma_2 A^{1/2}} q|^{1/2} \|e^{\sigma_2 A^{1/2}} q\|^{1/2} \|e^{\sigma_2 A^{1/2}} p\| \, |e^{\sigma_2 A^{1/2}}\Theta_1|^{1/2} \|e^{\sigma_2 A^{1/2}}\Theta_1\|^{1/2} \\
& + \frac{c_2}{\nu} |e^{\sigma_2 A^{1/2}} q|^{1/2} \|e^{\sigma_2 A^{1/2}} q\|^{3/2} |e^{\sigma_2 A^{1/2}}\Theta_1|^{1/2} \|e^{\sigma_2 A^{1/2}}\Theta_1\|^{1/2} \\
& + \frac{1}{\nu} \left| e^{\sigma_2 A^{1/2}} \frac{dq}{dt} \right| |e^{\sigma_2 A^{1/2}}\Theta_1|.
\end{aligned}$$

It follows from Proposition 3.2 that for $t >> 1$ and some constant R_2, $\|e^{\sigma_2 A^{1/2}} q(t)\| \leq R_2 \nu^{-1} \lambda_{m+1}^{-1}$. Using (3.3) and (3.6).

$$\begin{aligned}
\|e^{\sigma_2 A^{1/2}}\Theta_1\| \leq\ & \frac{c_9 L_m^{1/2}}{\nu^2} R_1 R_2 \lambda_{m+1}^{-1} + \frac{c_2}{\nu^2} R_1 R_2 \lambda_{m+1}^{-1} \\
& + \frac{c_2}{\nu^3} R_2^2 \lambda_{m+1}^{-3/2} + \frac{K_3}{\nu} \lambda_{m+1}^{-3/2}.
\end{aligned}$$

Hence for an appropriately chosen constant, say K_4, and $t >> 1$ we have

$$\text{dist}_V(u(t), \mathcal{M}_1) \leq \|\Theta_1\| \leq K_4 \lambda_{m+1}^{-1} e^{-\sigma_2 \lambda_{m+1}^{1/2}}. \qquad \square$$

This last estimate can be viewed as giving the maximum distance in the direction perpendicular to the plane $P_m H$ between the AIM and the attractor. If σ_2 is not too small, then there is only an algebraic improvement in the upper bound for the AIM over the upper bound for the Galerkin approximation in the distance solutions are attracted to their respective manifolds. We remark that the best-known lower estimate for σ_2 is $c/(G + G \log G)$, where c is an absolute constant and G, the Grashof number, is given by $G = |f|/\nu^2 \lambda_1$ [16]. We again mention that following the work of

[13] one can show that the rate of convergence of the nonlinear Galerkin methods is exponential for a given f in the Gevrey class.

4. An example. The AIMs presented in [21], [29], [57], [68], and [69] are based on the observation that if f is time-independent and $f \in H$, then every solution of (3.2) and (3.1) satisfies

$$|q(t)| = O(1/\lambda_{m+1}), \qquad t >> 1.$$

Moreover, it is known that in this case the solution of the complexified equation in time of (2.5) is analytic on a strip that is independent of the particular solution and includes the positive real axis [30]. It follows from the Cauchy formula that

$$\left|\frac{dq}{dt}\right| = O(1/\lambda_{m+1}), \qquad t >> 1$$

and is small compared with $Aq, B(p,q), B(q,p)$. We emphasize that our bounds on $|q(t)|, |dq/dt|$ are asymptotic in m. Furthermore, the upper bounds for the constants in our estimates depend on the Grashof or Reynolds number and in general are quite large. The constant in the bound on $|dq/dt|$ is large due to the width of the band of analyticity, which is inversely proportional to $G^2 \log G$ for the periodic boundary case (see [21]). However, for certain points on the attractor the constant can be of order one. For instance, near steady states it was shown in [51] that the constant can be made small.

It is important to extend AIMs to time-dependent forces. As mentioned earlier, in meteorology, for example, the forces are time-dependent. Furthermore, some computations with AIMs that were originally constructed for autonomous systems have been implemented with time-dependent forces. The results of these computations are encouraging (see, for example, [43]). We would like, therefore, to justify the use of AIMs with time-dependent forces. However, we show that for time-dependent forces the term dq/dt may be the dominate term in (3.2). We make our example more dramatic by choosing a volume force, $f(t)$, so that the solution, $u(t)$, of (2.5) is in the Gevrey class. We make our choice for $f(t)$ as follows. Set

$$u(t) = \sum_{j=1}^{\infty} \alpha_j \cos(\pi/4 + 2^j t)\varphi_j, \tag{4.1}$$

where the $\{\alpha_j\}$ are chosen such that

$$\sum_{j=1}^{\infty} \alpha_j^2 2^{2j} < \infty.$$

Since there exist constants such that $c_0 j \le \lambda_j \le c_1 j$ (see, for example, [7]), it follows that

$$\left|e^{\sigma_1 A^{1/2}} u(t)\right|^2 \le \sum_{j=1}^{\infty} e^{2\sigma_1 \lambda_j^{1/2}} \alpha_j^2 < \infty$$

(for all $\sigma_1 > 0$). Thus $u(t)$ is in a Gevrey class. Set

$$f(t) := \frac{du}{dt} + \nu Au + B(u,u).$$

Then $u(t)$ given by (4.1) solves (2.5), and at least $f(t) \in L^\infty(0,\infty;H)$. Notice, however, that for $t = 2\pi l, l \in \mathbb{N}$, we have

$$\left|\frac{dq}{dt}\right|^2 = \frac{1}{2}\sum_{j=m+1}^{\infty} 2^{2j}\alpha_j^2,$$

while

$$|Aq|^2 = \frac{1}{2}\sum_{j=m+1}^{\infty} \lambda_j^2\alpha_j^2.$$

Since $\lambda_j \sim j$ as $j \to \infty$ (see, e.g., [7]), $|Aq| << |dq/dt|$ and also $|B(u,u)| << |dq/dt|$. Therefore, one cannot in general neglect $|dq/dt|$ in (3.2) when the force is time-dependent.

5. An approximate inertial manifold. Denote by

$$\mathcal{B} = \{p \in P_mH : \ \|p\| \le 2M_1\}$$

and

$$\mathcal{B}^\perp = \{q \in Q_mV : \ \|q\| \le 2M_1\},$$

where M_1 satisfies (2.6). The assumptions on $f(t)$ in this section are as follows:

$$|(f(t_1) - f(t_2))| \le L_1|t_1 - t_2|^\theta. \tag{5.1}$$

Furthermore, $f \in L^\infty((0,\infty);H)$. That is, $\sup_{t\ge 0}|f(t)| \le f_\infty < \infty$. Notice that we do not require any further assumption about the asymptotic behavior of $f(t)$. Thus it may be that for our forces there is no global attractor for the system. However, if one is willing to make further assumptions about the force, for example, that it enters a compact set in H in finite time (see [61] and the references therein), or that f periodic in time (see, for example, [37]), it is possible to obtain a universal (global) attractor for the system.

If we require sufficient conditions on $f(t)$ (see §2) so that $|Au|$ is uniformly bounded, then the solution $p(t)$ of

$$\frac{dp}{dt} + \nu Ap + P_mB(u,u) = P_mf(t)$$

$$p(0) = P_mu_0$$

is uniformly Lipschitz in time. That is,

$$|p(t_1) - p(t_2)| \le L_2|t_1 - t_2|, \tag{5.2}$$

where L_2 depends only on ν, f_∞, λ_1 and not on t. Before constructing our time-dependent AIM we first estimate the distance the solutions are attracted to the flat manifold $P_mH = H_m$. If we take the inner product of (3.2) with Aq and use the estimates (2.14) and (2.15) (cf. the proof of Proposition 3.2), we have that for $t \ge T_*$,

$$\|q(t)\| \le \frac{K_5}{\lambda_{m+1}^{1/2}},$$

where T_* is as in (2.6). Therefore, without requiring further smoothness of $f(t)$ in time, and for any orbit $u(t)$ of (2.1), we have the following.

THEOREM 5.1. *Let* $\sup_{t\geq 0}|f(t)| \leq f_\infty < \infty$. *Then for* $t \geq T_*$,

$$\operatorname{dist}_V(u(t), P_m H) = \|q(t)\| \leq \frac{K_5}{\lambda_{m+1}^{1/2}}. \tag{5.3}$$

We remark that it has been shown by example in [69] and [71] that this last estimate is sharp asymptotically, as $m \to \infty$, up to a logarithmic term for a chosen $f \in L^2(\Omega)$. The goal then is to produce a nonlinear Galerkin scheme that will improve this error. In the arguments that follow we temporarily suppose that f is time-independent. Whenever f is time-dependent we will explicitly write $f(t)$. We recall (cf. [26], [31], [69], and [71]) for m large enough there exists a mapping $\Phi^s : \mathcal{B} \mapsto Q_m V$ that satisfies

$$A\Phi^s(p) + Q_m B(p + \Phi^s(p), p + \Phi^s(p)) = Q_m f \tag{5.4}$$

for all $p \in \mathcal{B}$. It is important to note that m depends only on M_1, ν, f_∞, and λ_1. Furthermore, the graph of Φ^s, denoted by $\mathcal{M}^s$, is a C-analytic manifold [26], [31]. In addition, it contains all of the stationary solutions of (2.5). Notice also that Φ^s depends on $Q_m f$. Therefore, we will write $\Phi^s(p, Q_m f)$. Hence, for $f(t)$ uniformly bounded in time (i.e., $|f(t)| \leq f_\infty < \infty$) the dimension, m, of the system can be obtained independently of t. That is, for some fixed m, $\Phi^s(p, Q_m f(t))$ exists for all time for all $p \in B$, where here the size of $\mathcal{B}$ depends on M_1, which depends on f_∞ and is uniform in time. We recall the following theorem of [69], which gives a lower bound for m.

THEOREM 5.2. *Let* m *be large enough such that*

$$\lambda_{m+1} \geq \max\left\{4r_2^2, \frac{r_1^2}{4M_1^2}\right\}. \tag{5.5}$$

Then there exists a unique mapping $\Phi^s : \mathcal{B} \mapsto Q_m V$ *that satisfies* (5.4). *Moreover,*

$$\|\Phi^s(p, Q_m f)\| \leq \lambda_{m+1}^{-1/2} r_1, \tag{5.6}$$

where

$$r_1 = \nu^{-1} c_9 8 M_1^2 L_m^{1/2} + \nu^{-1} c_2 8 M_1^2 + \nu^{-1} \lambda_{m+1}^{-1/2} |f|,$$

$$r_2 = \nu^{-1} c_9 2 M_1 L_m^{1/2} + \nu^{-1} c_2 6 M_1,$$

$$L_m = \left(1 + \log\left(\frac{\lambda_m}{\lambda_1}\right)\right).$$

In what follows we will need Φ^s bounded in $\mathcal{D}(A)$. Suppose m is chosen so large that (5.5) is satisfied and

$$\lambda_{m+1} \geq \max\left\{\left(\frac{4c_{10}(M_1 + r_1)}{\nu}\right)^2, \left(\frac{16 c_2 M_1}{\nu}\right)^4\right\}. \tag{5.7}$$

We have the following.

COROLLARY 5.3. *Let Φ^s be as in* (5.4). *Then*

$$|A\Phi^s(p,Q_mf)| \le \alpha_1 + 2\nu^{-1}|Q_mf| \quad \forall p \in \mathcal{B}, \tag{5.8}$$

where $\alpha_1 = 8c_9\nu^{-1}M_1^2\lambda_1^{-1}L_m^{1/2} + 4c_9\nu^{-1}M_1\lambda_1^{-1/2}r_1L_m^{1/2}\lambda_{m+1}^{-1/2}$.

Proof. From (5.4) we have

$$\nu|A\Phi^s| \le |B(p,p)| + |B(p,\Phi^s)| + |B(\Phi^s,p)| + |B(\Phi^s,\Phi^s)| + |Q_mf|.$$

Using the estimates (2.14) and (2.15) we get

$$\begin{aligned}\nu|A\Phi^s| \le\ & c_9|p|^2L_m^{1/2} + c_9|p|L_m^{1/2}\|\Phi^s\| + c_{10}|\Phi^s|^{1/2}|A\Phi^s|^{1/2}\|p\| \\ & + c_{10}|\Phi^s|^{1/2}|A\Phi^s|^{1/2}\|\Phi^s\| + |Q_mf|,\end{aligned}$$

and from (5.6),

$$\begin{aligned}\nu|A\Phi^s| \le\ & 4c_9M_1^2\lambda_1^{-1}L_m^{1/2} + 2c_9M_1\lambda_1^{-1/2}L_m^{1/2}r_1\lambda_{m+1}^{-1/2} + 2c_{10}M_1|A\Phi^s|\lambda_{m+1}^{-1/2} \\ & + c_{10}r_1\lambda_{m+1}^{-1}|A\Phi^s| + |Q_mf|.\end{aligned}$$

From (5.7) we conclude that

$$|A\Phi^s| \le 8c_9\nu^{-1}M_1^2L_m^{1/2} + 4c_9\nu^{-1}M_1r_1L_m^{1/2}\lambda_{m+1}^{-1/2} + 2\nu^{-1}|Q_mf|. \qquad \square$$

We need one more property of Φ^s.

LEMMA 5.4. *For all $p_i \in \mathcal{B}, f_i \in H$ with $|f_i| \le f_\infty$, $i = 1,2$ we have*

$$\begin{aligned}&\|\Phi^s(p_1,Q_mf_1) - \Phi^s(p_2,Q_mf_2)\| \\ &\le\ \alpha_2|p_1-p_2| + \nu^{-1}\lambda_{m+1}^{-1/2}|Q_mf_1 - Q_mf_2|,\end{aligned} \tag{5.9}$$

where $\alpha_2 = \nu^{-1}(c_38M_1L_m^{1/2} + c_48M_1L_m^{1/2})$.

Proof. Set $\Delta p = p_1 - p_2$, $q_i = \Phi^s(p_i,Q_mf_i)$, $\Delta q = q_1 - q_2$ and $\Delta f = f_1 - f_2$. Again Φ^s solves

$$A\Phi^s(p_i,Q_mf_i) + Q_mB(p_i + \Phi^s(p_i,Q_mf_i), p_i + \Phi^s(p_i,Q_mf_i)) = Q_mf_i$$

for $i = 1,2$. Subtracting the two equations we obtain

$$\begin{aligned}A\Delta q = Q_m\Delta f - Q_m(&B(p_1+q_1,\Delta p) \\ &+ B(p_1+q_1,\Delta q) + B(\Delta p, p_2+q_2) + B(\Delta q, p_2+q_2)).\end{aligned} \tag{5.10}$$

We take the scalar product in H of (5.10) with Δq, and we use (2.16). Then we use (2.9), (2.8), (2.10), (2.9), respectively, to obtain

$$\begin{aligned}\nu\|\Delta q\|^2 \le\ & c_3\|p_1+q_1\|\|\Delta q\||\Delta p|\left[1 + \log\left(\frac{\|\Delta p\|}{|\Delta p|\lambda_1^{1/2}}\right)\right]^{1/2} \\ & + c_2|p_1+q_1|^{1/2}\|p_1+q_1\|^{1/2}\|\Delta q\||\Delta q|^{1/2}\|\Delta q\|^{1/2} \\ & + c_4|\Delta p|\,\|p_2+q_2\|\,\|\Delta q\|\left[1 + \log\left(\frac{\|\Delta p\|}{|\Delta p|\lambda_1^{1/2}}\right)\right]^{1/2} \\ & + c_2|\Delta q|^{1/2}\|\Delta q\|^{1/2}\|p_2+q_2\|\,|\Delta q|^{1/2}\|\Delta q\|^{1/2} + |Q_m\Delta f||\Delta q|\end{aligned}$$

or

$$\nu\|\Delta q\|^2 \le c_3 4M_1 L_m^{1/2}\|\Delta q\|\|\Delta p| + c_2 4M_1\lambda_{m+1}^{-1/4}\|\Delta q\|^2 + c_4 4M_1 L_m^{1/2}|\Delta p|\|\Delta q\| + c_2 4M_1\lambda_{m+1}^{-1/2}\|\Delta q\|^2 + \lambda_{m+1}^{-1/2}|Q_m\Delta f|\|\Delta q\|.$$

We conclude from (5.7) that

$$\|\Delta q\| \le \nu^{-1}(c_3 8M_1 L_m^{1/2} + c_4 8M_1 L_m^{1/2})|\Delta p| + 2\nu^{-1}\lambda_{m+1}^{-1/2}|Q_m\Delta f|. \qquad \square$$

Our aim here is to show that solutions of equations (3.1) and (3.2) are attracted to a thin neighborhood of the (now time-dependent manifold) given by $\mathcal{M}(t) = \text{Graph}(\Phi^s(p, Q_m f(t)))$. A crucial element in estimating the size of this neighborhood in the time-independent case for the AIMs given in [21], [29], [57], [68], and [69] is obtaining a bound on $|dq/dt|$. As seen in the last section, if $f(t)$ is time-dependent, it may not be that dq/dt is small compared to the other terms of the equation. We avoid estimating dq/dt directly (as in [65] and [29]) by considering the system

$$\begin{aligned} &\frac{dw}{dt} + \nu Aw + Q_m B(p+w, p+w) = Q_m f, \\ &w(0) = q_0, \end{aligned} \tag{5.11}$$

where p is fixed in $\mathcal{B}$ and $q_0 \in \mathcal{B}^\perp$ is given. We will show that solutions of (5.11) decay exponentially to the unique stationary solution $\Phi^s(p, Q_m f)$. We will also see that solutions of the NSE remain close to solutions of (5.11) for short time. This will enable us to estimate the distance that solutions of the NSE are attracted to our time-dependent manifold.

The existence of solutions to (5.11) follows just as for (2.5) (see [7] and [66]). Furthermore, under the assumptions on p and q_0 solutions of (5.11) remain bounded for all $t \ge 0$. That is, $\|p + w(t)\| \le M_3$ for all $t \ge 0$ and $p \in \mathcal{B}$. The next lemma shows the exponential decay of solutions of (5.11) to the unique stationary solution $\Phi^s(p, Q_m f)$. Again we suppose that m is chosen so large that (5.5) and (5.7) are satisfied. Furthermore, we require

$$\lambda_{m+1} \ge \max\left\{\left(\frac{2c_1 M_3\lambda_1^{-1/4} + 4c_{10}M_1 L_m^{1/2} + 2c_1 r_1^{1/2}(\alpha_1 + 2\nu^{-1}f_\infty)^{1/2}}{\nu}\right)^4, \left(\frac{8c_1 M_1\lambda_1^{-1/4} + 4c_{10}M_3}{\nu}\right)^4\right\}. \tag{5.12}$$

LEMMA 5.5. *Let m be large enough such that* (5.5), (5.7), (5.12) *are satisfied. Let $p \in \mathcal{B}, |f| \le f_\infty$, and $q_0 \in \mathcal{B}^\perp$ be given. Then, if $w(t)$ is the solution of* (5.11),

$$\|w(t) - \Phi^s(p, Q_m f)\|^2 \le e^{-\nu\lambda_{m+1}t}\|q_0 - \Phi^s(p, Q_m f)\|^2 \tag{5.13}$$

for all $t \ge 0$.

Proof. Set $\Delta = w(t) - \Phi^s(p, Q_m f)$. Δ solves

$$\frac{d}{dt}\Delta + \nu A\Delta + Q_m B(p+w, \Delta) + Q_m B(\Delta, p + \Phi^s) = 0. \tag{5.14}$$

Taking the inner product of (5.14) with $A\Delta$ and using estimates (2.7) and (2.11) we find that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\Delta\|^2+\nu|A\Delta|^2 &\le |(Q_mB(p+w,\Delta),A\Delta)|\\ &\quad+|(Q_mB(\Delta,p),A\Delta)|+|(Q_mB(\Delta,\Phi^s),A\Delta)|\\ &\le c_1|p+w|^{1/2}\|p+w\|^{1/2}\|\Delta\|^{1/2}|A\Delta|^{1/2}|A\Delta|\\ &\quad+c_5\|\Delta\|2M_1L_m^{1/2}|A\Delta|\\ &\quad+c_1|\Delta|^{1/2}\|\Delta\|^{1/2}\|\Phi^s\|^{1/2}|A\Phi^s|^{1/2}|A\Delta|.\end{aligned}$$

Using Theorem 5.2 and Corollary 5.3,

$$\le\left(c_1M_3\lambda_1^{-1/4}\lambda_{m+1}^{-1/4}+c_52M_1L_m^{1/2}\lambda_{m+1}^{-1/2}+c_1r_1^{1/2}(\alpha_1+2\nu^{-1}f_\infty)^{1/2}\lambda_{m+1}^{-1}\right)|A\Delta|^2,$$

and from (5.12),

$$\frac{d}{dt}\|\Delta\|^2+\nu\lambda_{m+1}\|\Delta\|^2\le 0.$$

The result follows after an application of Gronwall's inequality. □

We again return to the case of time-dependent forces. We suppose that $f(t)$ is time-dependent satisfying (5.1) and $\sup_{t\ge0}|f(t)|\le f_\infty$. We suppose that $t\ge T_*$, where T_* is as in (2.6), so that $u(t)=p(t)+q(t)\in\mathcal{B}\cup\mathcal{B}^\perp$, and we let $\bar t\in[T_*,\infty)$. Set $\overline{p}=p(\bar t),\overline{q}=q(\bar t),Q_m\overline{f}=Q_mf(\bar t)$. Consider the initial value problems

$$\begin{aligned}&\frac{dq}{dt}+\nu Aq+Q_mB(u,u)=Q_mf(t)\\ &\qquad\qquad q(\bar t)=\overline{q},\end{aligned}\tag{5.15}$$

$$\begin{aligned}&\frac{dw}{dt}+\nu Aw+Q_mB(\overline{p}+w,\overline{p}+w)=Q_m\overline{f},\\ &\qquad\qquad w(\bar t)=\overline{q}.\end{aligned}\tag{5.16}$$

Since $\overline{q}\in\mathcal{B}^\perp$, we have $\|\overline{p}+w(t)\|\le M_3$ for $t\ge\bar t$. The next lemma shows that $w(t)$ and $q(t)$ are close for short time.

LEMMA 5.6. *Under the above assumptions the solutions $q(t),w(t)$ of* (5.15) *and* (5.16), *respectively,*

$$\|q(t)-w(t)\|\le\lambda_{m+1}^{-1/2}2\nu^{-1}\left(\alpha_3L_2\lambda_m^{1/2}\tau^\theta+L_1\tau\right),\qquad\bar t\le t\le\bar t+\tau,\tag{5.17}$$

where $\alpha_3=2M_1c_5L_m^{1/2}+M_3c_9L_m^{1/2}$.

Proof. Set $\Delta=q(t)-w(t)$. Then Δ solves

$$\begin{aligned}&\frac{d\Delta}{dt}+\nu A\Delta+Q_mB(u,p-\overline{p}+\Delta)\\ &\quad+Q_mB(p-\overline{p}+\Delta,\overline{p}+w)=Q_m(f(t)-\overline{f}),\\ &\qquad\qquad\Delta(\bar t)=0.\end{aligned}\tag{5.18}$$

Taking the inner product of (5.18) with $A\Delta$, we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\Delta\|^2+\nu|A\Delta|^2 \le\ & |(B(u,p-\overline{p}),A\Delta)|+|(B(u,\Delta),A\Delta)|\\ &+|(B(p-\overline{p},\overline{p}+w),A\Delta)|\\ &+|(B(\Delta,\overline{p}+w),A\Delta)|+|Q_m(f(t)-\overline{f})||A\Delta|.\end{aligned}$$

After an application of (2.11), (2.7), (2.14), and (2.15), respectively, we find

$$\begin{aligned}\le\ & c_5\|u\|\|p-\overline{p}\||A\Delta|L_m^{1/2}+c_1|u|^{1/2}\|u\|^{1/2}\|\Delta\|^{1/2}|A\Delta|^{1/2}|A\Delta|\\ &+c_9\|p-\overline{p}\|\|\overline{p}+w\||A\Delta|L_m^{1/2}+c_{10}|\Delta|^{1/2}|A\Delta|^{1/2}\|\overline{p}+w\||A\Delta|\\ &+|Q_m(f(t)-\overline{f})||A\Delta|.\end{aligned}$$

After an application of Young's inequality,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\Delta\|^2+\frac{\nu}{2}|A\Delta|^2 \le\ & (\nu)^{-1}\left(2M_1c_5L_m^{1/2}+M_3c_9L_m^{1/2}\right)^2\|p-\overline{p}\|^2\\ &+\nu^{-1}|Q_m(f(t)-\overline{f})|^2\\ &+\left(2c_1M_1\lambda_{m+1}^{-1/4}+c_{10}M_3\lambda_{m+1}^{-1/2}\right)|A\Delta|^2.\end{aligned}$$

It follows from (5.12), (5.2), and (5.1) that

$$\frac{d}{dt}\|\Delta\|^2+\frac{\nu\lambda_{m+1}}{2}\|\Delta\ |^2\le 2\nu^{-1}\left(\alpha_3^2L_2^2\lambda_m\tau^{2\theta}+L_1^2\tau^2\right).$$

The result follows after an application of Gronwall's inequality and the fact that $\Delta(\overline{t})=0$. □

THEOREM 5.7. *Let $f(t)$ satisfy* (5.1) *and be such that* $\sup_{t\ge 0}|f(t)|\le f_\infty$ *and* (5.2) *hold. Furthermore, let m be chosen so large that* (5.5), (5.7), *and* (5.12) *are satisfied. Then for any solution $u(t)=p(t)+q(t)$ of* (2.5) *and $t\ge T_*$. We have*

$$\|q(t)-\Phi^s(p(t),Q_mf(t))\|\le\frac{\alpha_5}{\lambda_{m+1}^{\theta}}+d_0e^{-\lambda_{m+1}(t-T_*)},\tag{5.19}$$

where $\alpha_5=\alpha_4(1+(1+e)^{-1})$, $\alpha_4=\lambda_{m+1}^{-1/2}2\nu^{-1}(\alpha_3L_2\lambda_m^{1/2}+L_1)+\alpha_2L_2+\nu^{-1}L_1\lambda_{m+1}^{-1/2}$, *and* $d_0=\|q(T_*)-\Phi^s(p(T_*),Q_mf(T_*))\|$.

Proof. Set $\tau=(\nu\lambda_{m+1})^{-1}$. Define the sequences $\{t_n\},\{p_n\},\{Q_mf_n\},\{d_n\}$ by $t_n=T_*+n\tau$, $p_n=p(t_n)=P_mu(t_n),Q_mf_n=Q_mf(t_n)$ and $d_n=\|q(t_n)-\Phi^s(p_n,Q_mf_n)\|$, respectively. Let w_n solve

$$\frac{dw_n}{dt}+\nu Aw_n+Q_mB(p_n+w_n,p_n+w_n)=Q_mf_n$$

with initial condition

$$w_n(t_n)=q(t_n)$$

on the interval $t_n\le t\le t_{n+1}$. From Lemma 5.5 we see that

$$\|w_n(t)-\Phi^s(p_n,Q_mf_n)\|\le d_ne^{-\nu\lambda_{m+1}(t-t_n)}.\tag{5.20}$$

We then have

$$\begin{aligned}d_{n+1}\le\ & \|q(t_{n+1})-w(t_{n+1})\|+\|w(t_{n+1})-\Phi^s(p_n,Q_mf_n)\|\\ &+\|\Phi^s(p_{n+1},Q_mf_{n+1})-\Phi^s(p_n,Q_mf_n)\|.\end{aligned}$$

Upon using (5.17), (5.20), (5.9), (5.1), and (5.2) we find

$$d_{n+1} \leq \alpha_4 \tau^\theta + d_n e^{-\nu\lambda_{m+1}\tau}. \tag{5.21}$$

Thanks to our choice of τ, $e^{-\nu\lambda_{m+1}\tau} = e^{-1} < 1$, and we may iterate (5.21) to obtain

$$d_n \leq \frac{\alpha_4 \tau^\theta}{1 - e^{-\nu\lambda_{m+1}\tau}} + d_0 e^{-n}, \qquad n \geq 1.$$

Now using (5.17) and (5.21), and the estimate on d_n we can obtain a continuous estimate on the interval $[t_n, t_{n+1}]$:

$$\begin{aligned} \|q(t) - \Phi^s(p_n, Q_m f_n)\| &\leq \|q(t) - w_n(t)\| + \|w_n(t) - \Phi^s(p_n, Q_m f_n)\| \\ &\leq 2\nu\lambda_{m+1}^{-1/2}(\alpha_3 L_2 \lambda_m^{1/2} + L_1)(\tau)^\theta \\ &\quad + \left(\frac{\alpha_4 \tau^\theta}{1 - e^{-1}} + e^{-n}\right) e^{-\nu\lambda_{m+1}(t-t_n)}. \end{aligned}$$

For $t \geq T_*$ choose n so that $T_* + n\tau \leq t \leq T_* + (n+1)\tau$. Then

$$\begin{aligned} \|q(t) - \Phi^s(p(t), Q_m f(t))\| &\leq \|q(t) - \Phi^s(p_n, Q_m f_n)\| \\ &\quad + \|\Phi^s(p_n, Q_m f_n) - \Phi^s(p(t), Q_m f(t))\|. \end{aligned}$$

After using (5.9), (5.1), and (5.2) the result follows. □

COROLLARY 5.8. *Under the hypotheses of Theorem* (5.7) *the solution* $u(t)$ *of* (2.5) *approaches a neighborhood of the manifold*

$$\mathcal{M}(t) = \text{Graph}\Phi^s(p, Q_m f(t)), \qquad p \in \mathcal{B}.$$

In particular,

$$\begin{aligned} \text{dist}_V(u(t), \mathcal{M}(t)) &\leq \|q(t) - \Phi^s(p(t), Q_m f(t))\| \\ &\leq \text{dist}_V(u(T_*) - \mathcal{M}(T_*)) e^{-\lambda_{m+1}(t-T_*)} + \frac{\alpha_5}{\lambda_{m+1}^\theta} \end{aligned}$$

for $t \geq T_*$.

Remark 5.9. If m_c is the dimension of the manifold given by Theorem 5.7, then one notices that the proof of Lemma 5.17, and consequently Theorem 5.7, only requires $|Q_{m_c}(f(t_1) - f(t_2))| \leq L_1 |t_1 - t_2|^\theta$. No such condition is required on $|P_{m_c} f(t)|$.

Remark 5.10. If f is time-independent, then θ can be chosen to be arbitrarily large. However, we see from (5.17) that in this case the estimate given in Theorem 5.7 for the distance solutions are attracted to our time-dependent manifold is $\alpha_5 \lambda_{m+1}^{-1}$. The estimate obtained in [69] for this case was obtained using a different approach based on the analyticity in time of the solutions. The estimate obtained there is $C\lambda_{m+1}^{-3/2}$. If one assumes $f(t)$ is analytic in time, and uniformly bounded, then following the arguments in [69], one would obtain the same estimate as in [69] for the distance solutions are attracted to the time-dependent manifold.

A similar analysis is possible for other time-dependent manifolds. For example, we may extend the AIM given in [21] to time-dependent forces. Set

$$\Phi_1(p, Q_m f(t)) = (\nu A)^{-1}(Q_m f(t) - Q_m B(p, p))$$

(cf. §3) and

$$\mathcal{M}_1(t) = \text{Graph}(\Phi_1).$$

We have the following.

THEOREM 5.11. *Under the hypotheses of Theorem* 5.7 *and for* t, m *sufficiently large, any orbit of* (2.1) *satisfies*

$$\text{dist}_V(u(t), \mathcal{M}_1(t)) \leq \|q(t) - \Phi_1(p(t), Q_m f(t))\| \leq \frac{4\alpha_5}{\lambda_{m+1}^{\theta}}.$$

Proof. The proof uses the same arguments as above, but one uses

$$\frac{dw}{dt} + \nu Aw + Q_m B(p,p) = Q_m f$$

instead of (5.11). □

Remark 5.12. If f is time-independent, then the bound given in Corollary 5.11 agrees with the estimates given in [21].

Notice that if $\theta < 1/2$, then the estimate we obtain for the upper bound in the distance solutions are attracted to our time-dependent manifolds is worse than the estimate for the flat manifold in (5.3) (standard Galerkin scheme). We can handle the case $\theta \leq 1/2$ in the following proposition. We also do not require $|Au(t)|$ to be bounded as $t \to \infty$.

PROPOSITION 5.13. *Assume* $\sup_{t \geq 0} |f(t)| \leq f_\infty < \infty$. *Then for* $t \geq T_*$,

$$\text{dist}_V(u(t) - \mathcal{M}(t)) \leq \|q(t) - \Phi^s(p(t), Q_m f(t))\| \leq \frac{K_6}{\lambda_{m+1}^{1/2}}, \tag{5.22}$$

$$\text{dist}_V(u(t) - \mathcal{M}_1(t)) \leq \|q(t) - \Phi_1(p(t), Q_m f(t))\| \leq \frac{K_7}{\lambda_{m+1}^{1/2}}. \tag{5.23}$$

Proof. From Theorems 5.3 and 5.2, respectively, we have that $\|q(t)\| \leq K_5 \lambda_{m+1}^{-1/2}$ and $\|\Phi^s(p(t), Q_m f(t))\| \leq r_1 \lambda_{m+1}^{-1/2}$ for $t \geq T_*$. Equation (5.22) then follows with $K_6 = r_1 + K_5$. The estimate (5.23) follows in a similar fashion. □

If $f(t) \in L^\infty(0, \infty; H)$ and $f'(t) \in L^\infty(0, \infty; H)$, then one can show that $|du/dt|$ is bounded, and hence $|Au|$ is uniformly bounded as $t \to \infty$ (see [53]). Furthermore, Theorem 5.7 and Corollary 5.11 apply with $\theta = 1$. They yield that the error of the nonlinear Galerkin methods are of the order λ_{m+1}^{-1}, whereas the standard Galerkin scheme gives an error of the order $\lambda_{m+1}^{-1/2}$. Moreover, in view of the example in [69], the bound $|q(t)| \leq K_5 \lambda_{m+1}^{1/2}$ is sharp asymptotically, as $m \to \infty$, up to a logarithmic term for a chosen f which is time-independent. In particular, this f satisfies $f \in L^\infty(0, \infty; H)$ and $f'(t) \in L^\infty(0, \infty; H)$. Therefore, we may expect that in general the nonlinear Galerkin methods will give an improvement over the usual Galerkin scheme in this case as well.

Acknowledgments. Part of the work was done while the authors enjoyed the hospitality of the Center for Nonlinear Studies and the Institute of Geophysics and Planetary Physics at Los Alamos National Laboratory.

REFERENCES

[1] N. Aubry, P. Holmes, J. L. Lumley, and E. Stone, *The dynamics of coherent structures in the wall region of a turbulent boundary layer*, J. Fluid Mech., 192 (1988), pp. 115–173.

[2] A. V. Babin and M. I. Vishik, *Attractors of partial differential equations and estimate of their dimension,* Uspekhi Mat. Nauk, 38 (1983), pp. 133–187 (in Russian); Russian Math. Surveys, 38 (1983), pp. 151–213. (In English.)

[3] A. Bloch and E. S. Titi, *On the dynamics of rotating elastic beams*, Proc. of New Trends in Systems Theory, July 9-11, 1990, Genoa, Italy, G. Conte, A.M. Perdon and B. Wyman, eds., Birkhäuser Verlag, Boston, Basel, Berlin.

[4] H. Brézis and T. Gallouet, *Nonlinear Schrödinger evolution equations*, Nonlinear Anal., TMA, 4 (1980), pp. 677–681.

[5] H. S. Brown, M. S. Jolly, I. G. Kevrekidis, and E. S. Titi, *Use of approximate inertial manifolds in bifurcation calculations,* Proceedings of NATO Advanced Research Workshop on: Continuations and Bifurcations: Numerical Techniques and Applications, D. Roose et. al., eds., Kluwer Academic Publishers, Norwell, MA, 1989, pp. 9–23.

[6] W. Chen, *Approximate inertial manifolds for the 2D Navier–Stokes equations*, J. Math. Anal. Appl., 165 (1992), pp. 399–418.

[7] P. Constantin and C. Foias, *Navier–Stokes Equations,* University of Chicago Press, Chicago, IL, 1988.

[8] P. Constantin, C. Foias, O. P. Manley, and R. Temam, *Determining modes and fractal dimension of turbulent flows,* J. Fluid Mech., 150 (1985), pp. 427–440.

[9] P. Constantin, C. Foias, B. Nicolaenko, and R. Temam, *Integral manifolds and inertial manifolds for dissipative partial differential equations*, Appl. Math. Sciences, Springer-Verlag, New York, 1989.

[10] ———, *Spectral barriers and inertial manifolds for dissiapative partial differential equations*, J. Dynam. Differential Equations, 1 (1989), pp. 45–73.

[11] P. Constantin, C. Foias, and R. Temam, *On the dimension of the attractors in two-dimensional turbulence*, Physica, D30 (1988), pp. 284–296.

[12] C. Devulder and M. Marion, *A class of numerical algorithms for large time integration: the nonlinear Galerkin methods*, SIAM J. Numer. Anal., 29 (1992), pp. 462–483.

[13] C. Devulder, M. Marion, and E. S. Titi, *On the rate of convergence of the nonlinear Galerkin methods*, Math. Comp., 60 (1993), pp. 495–514.

[14] A. Doelman and E. S. Titi, *The exponential decay of modes in the Ginzburg–Landau equation*, Proceedings of the NATO Advanced Research Workshop: Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters, M. Garbey and H. G. Kaper, eds., Kluwer Academic Publishers, Dordrecht, Germany, 1993, pp. 241–252.

[15] E. Fabes, M. Luskin, and G. Sell, *Construction of inertial manifolds by elliptic regularization*, J. Differential Equations, 89 (1991), pp. 355–387.

[16] C. Foias, private communication.

[17] C. Foias, M. S. Jolly, I. G. Keverkidis, and E. S. Titi, *Dissipativity of numerical schemes*, Nonlinearity, 4 (1991), pp. 591–613.

[18] ———, *On some dissipative fully discrete nonlinear Galerkin schemes for the Kuramoto–Sivashinsky equation*, Phys. Lett. A, 1994, to appear.

[19] C. Foias, M. S. Jolly, I. G. Keverkidis, G. R. Sell, and E. S. Titi, *On the computation of inertial manifolds*, Phys. Lett. A, 31 (1988), pp. 433–436.

[20] C. Foias, O.P. Manley, and R. Temam, *Sur l'interaction des petits et grands tourbillans dans les écoulements turbulents*, C.R. Acad. Sci. Paris, Sér. I, 305 (1987), pp. 487–500.

[21] ———, *Modelization of the interaction of small and large eddies in two dimensional turbulent flows,* Math. Mod. Numer. Anal., 22 (1988), pp. 93–114.

[22] ———, *Approximate inertial manifolds and effective viscosity in turbulent flows*, Phys. Fluids A, 3 (1991), pp. 898–911.

[23] C. Foias, O. P. Manley, R. Temam, and Y. Treve, *Asymptotic analysis of the Navier–Stokes equations*, Physica, D9 (1983), pp. 157–188.

[24] C. Foias, B. Nicolaenko, G. R. Sell, and R. Temam, *Inertial manifolds for the Kuramoto–Sivashinsky equation and an estimate of their lowest diminsions*, J. Math. Pures Appl., 67 (1988), pp. 197–226.

[25] C. Foias and G. Prodi, *Sur le comportement global des solutions non stationnaires des équations de Navier–Stokes en dimension two*, Rend. Sem. Mat. Univ. Padova, 39 (1967), pp. 1–34.

[26] C. Foias and J. C. Saut, *Remarques sur les équations de Navier–Stokes stationnaires*, Ann. Scuola Norm. Sup. Pisa, 10 (1983), pp. 169–177.

[27] C. FOIAS, G. SELL, AND R. TEMAM, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309–353.

[28] ———, *Variétés inertielles des équations différentielles dissipatives*, C. R. Acad. Sci. Paris, Sér. I, 301 (1985), pp. 139–142.

[29] C. FOIAS, G. SELL, AND E. S. TITI, *Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations*, J. Dynam. Differential Equations, 1 (1989), pp. 199–243.

[30] C. FOIAS AND R. TEMAM, *Some analytic and geometric properties of the solutions of the Navier–Stokes equations*, J. Math. Pures Appl., 58 (1979), pp. 339–368.

[31] ———, *Remarques sur les équations de Navier–Stokes stationnaires et les phénomènes successifs de bifurcation*, Ann. Scuola Norm. Sup. Pisa, 5 (1978), pp. 29–63.

[32] ———, *The algebraic approximation for attractors: the finite dimensional case*, Physica, D32 (1988), pp. 163–182.

[33] ———, *Determination of the solutions of the Navier–Stokes equations by a set of nodal values*, Math. Comput., 43 (1984), pp. 117–133.

[34] ———, *Gevrey class regularity for the solutions of the Navier–Stokes equations*, J. Funct. Anal., 87 (1989), pp. 359–369.

[35] C. FOIAS AND E. S. TITI, *Determining nodes, finite difference schemes and inertial manifolds*, Nonlinearity, 4 (1991), pp. 135–153.

[36] J-M GHIDAGLIA AND R. TEMAM, *Lower bound on the dimension of the attrator for the Navier–Stokes equations in space dimension* 3, in Mechanics, Analysis and Geomerty: 200 years after Lagrange, M. Francaviglia, ed., Elsevier Science Publishers B.V. New York, (1991), pp. 33–60,

[37] ———, *Dimension of the universal attractor describing the periodically driven sine-Gordon equations*, Transport Theory Statist. Phys., 16 (1987), pp. 253–265.

[38] M. D. GRAHAM, P. H. STEEN, AND E. S. TITI, *Computational efficiency and approximate inertial manifolds for a Bénard convection system*, J. of Nonlin. Sci., 3 (1993), pp. 153–167.

[39] C. GUILLOPÉ, *Comportement à l'infini des solutions des équations de Navier–Stokes et propriété des ensembles fonctionnels invariants (ou attracteurs)*, Ann. Inst. Fourier (Grenoble), 32 (1982), pp. 1–37.

[40] J. HALE, *Asymptotic Behavior of Dissipative Systems*, Mathematical Surveys and Monographs, 25, American Mathematical Society, Providence, RI, 1988.

[41] W. D. HENSHAW, H. O. KREISS, AND L. G. REYNA, *Smallest scale estimates for the Navier–Stokes equations for incompressible fluids*, Arch. Rational Mech. Anal., 112 (1990), pp. 21–44.

[42] J. G. HEYWOOD, *The Navier–Stokes equations: On the existence, regularity and decay of solutions*, Indiana Univ. Math. J., 29 (1980), pp. 639–681.

[43] F. JAUBERTEAU, C. ROSIER, AND R. TEMAM, *A nonlinear Galerkin method for the Navier–Stokes equations*, Proc. Conf. on Spectral and High Order Methods for PDEs, ICOSAHOM 89, Como, Italie, 1989.

[44] M. S. JOLLY, I. G. KEVREKIDIS, AND E. S. TITI, *Approximate inertial manifolds for the Kuramoto–Sivashinsky equation: analysis and computations*, Phys. D, 44 (1990), pp. 38–60.

[45] ———, *Preserving dissipation in approximate inertial forms for the Kuramoto–Sivashinsky equation*, J. Dynam. Differential Equations, 3 (1990), pp. 179–197.

[46] D. A. JONES AND E. S. TITI, *On the number of determining nodes for the 2D Navier–Stokes equations*, J. Math. Anal. Appl., 168 (1992), pp. 72–88.

[47] ———, *Determination of the solutions of the Navier–Stokes equations by finite volume elements*, Phys. D, 60 (1992), pp. 165–174.

[48] ———, *C^1 approximating inertial manifolds for dissipative nonlinear equations*, submitted.

[49] ———, *On the existance of slow manifolds for the Navier–Stokes equations, the statistical case*, manuscript.

[50] H. O. KREISS, *Fourier expansions of the Navier–Stokes equations and their exponential decay rate*, Analyse Mathématique et Applications, Gauthier-Villars, Paris, 1988, pp. 245–262.

[51] I. KUKAVICA, *On the time analyticity radius of the solutions of the 2 dimensional Navier–Stokes Equations*, J. Dynam. Differential Equations, 3 (1991), pp. 611–618.

[52] M. KWAK, *Finite dimensional inertial form for the 2D Navier–Stokes equations*, Indiana University Math. J., 41 (1992), pp. 927–982.

[53] J. L. LIONS, *Quelques Method de Résolution de Problém aux Limites Non Linéaire*, Dunod, Paris, 1969.

[54] X. LIU, *Gevrey class regularity and approximate inertial manifolds for the Kuramoto–Sivashinsky equations*, Phys. D, 50 (1991), pp. 135–151.

[55] J. MALLET-PARET AND G. SELL, *Inertial manifolds for reaction-diffusion equations in higher space dimension*, J. Amer. Math. Soc., 1 (1988), pp. 805–866.

[56] L. MARGOLIN AND D. A. JONES, *An approximate inertial manifold for computing Burgers' equation*, Phys. D, 60 (1992), pp. 175–184.

[57] M. MARION, *Approximate inertial manifolds for reaction-diffusion equations in high space dimensions*, J. Dynam. Differential Equations, 1 (1989), pp. 245–267.

[58] M. MARION AND R. TEMAM, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1990), pp. 1139–1157.

[59] ———, *Nonlinear Galerkin methods; the finite elements case*, Numer. Math., 57 (1990), pp. 205–226.

[60] K. PROMISLOW, *The Development and Numerical Implementation of Approximate Inertial Manifolds for the Ginzburg–Landau Equation*, Ph.D. thesis, Indiana University, Bloomington, IN, 1991.

[61] G. RAUGEL, AND G. SELL, *Navier–Stokes equations in thin 3D domains: global regularity of solutions* I, J. Amer. Math. Soc., 6 (1993), pp. 503–568.

[62] J. SHEN, *Long time stability and convergence for fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–229.

[63] L. SIROVICH, *Turbulence and the dynamics of coherent structures, Parts*, I, II, III, Quart. Appl. Math., 45 (1987), pp. 561–590.

[64] L. SIROVICH, B. W. KNIGHT, J. D. RODRIGUEZ, *Optimal low-dimensional dynamical approximations*, Quart. Appl. Math., 48 (1990), pp. 535–548.

[65] M. W. SMILEY, *Global attractors and approximate inertial manifolds for abstract dissipative equations*, IMA Univ. of Minnesota, preprint 672, 1990.

[66] R. TEMAM, *Navier–Stokes Equations and Nonlinear Functional Analysis*, CBMS Regional Conference Series, No. 41, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

[67] ———, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Applied Mathematical Science Series, 68, Springer-Verlag, New York, 1988.

[68] ———, *Attractors for the Navier–Stokes equations: localization and approximation*, J. Fac. Sci., Univ. Tokyo, IA, 36 (1989), pp. 629–647.

[69] E. S. TITI, *On approximate inertial manifolds to the Navier–Stokes equations*, J. Math. Anal. Appl., 149 (1990), pp. 540–557.

[70] ———, *On a criterion for locating stable stationary solutions to the Navier–Stokes equations*, Nonlinear Anal., TMA, 11 (1987), pp. 1085–1102.

[71] ———, *Une variété approximate de l'attracteur universal des équation de Navier–Stokes, nonlinéaire, de dimensions finie*, C.R. Acad. Sci. Paris, Sér. I, 307 (1988), pp. 383–385.

[72] ———, *Gevrey class regularity and long time approximation for 3D convection in porous media*, manuscript.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 915–940, May 1994

006

DISPERSION IN PIPES WITH SLOWLY VARYING CROSS-SECTIONS*

W. P. KOTORYNSKI†

Abstract. An asymptotic solution is developed for the dispersion of solute in fully developed laminar flow through a pipe with cross-section slowly varying in a longitudinal direction. The solution is found as a power series in a small parameter chosen to characterize the slow variation of the pipe boundary. Because of the presence of two different time scales the series is expressed as separate expansions in regular- and short-time scale terms. A perturbative solution is considered in detail for an example of dispersion in a spiralling circular pipe.

Key words. diffusion-convection, slow variations, Poiseuille flow, two-timing approximations, mass-heat transport, effective diffusion coefficient

AMS subject classification. 76R

1. Introduction.

The results presented here comprise a method for constructing asymptotic solutions describing the dispersion of solute—or the diffusion of any other scalar field, such as heat, which may be interest—within a fluid flowing down a pipe of variable cross-section. For an infinite straight pipe of uniform cross-section with a convecting Poiseuille flow the problem is a classic one. The theory was initiated by G. I. Taylor [24] and there is now an extensive literature comprising subsequent contributions of a large number of authors (cf., [1], [4], [7], [10], [14], [22], and [30]). Previous work, as far as we are aware, has not addressed the situation in which the dispersion takes place in those circumstances—of practical importance—where the underlying flow is not strict Poiseuille flow. First-order effects due to secondary flow in differing cross-sectional planes will be present: enhancement of the mixing of solute or increases in heat or mass transfer will occur due to asymmetries.

Recent work by Mercer and Roberts (see [16], [17], [19], and the references therein), in part, is complementary to this paper and, in particular, relates to analysis of long-time asymptotic behavior of the concentration that is addressed in §4 of this paper. The derivations by these authors are based on invariant manifold theory, which is capable of providing a systematic approach to the calculation of successive approximations correcting the leading-order approximate equations and giving a description of the asymptotic evolution of evolving dynamical systems. At the same time, the geometric picture of an invariant manifold gives a way of deriving correct initial conditions for an approximation, a task that frequently demands careful, separate consideration.

In presenting a method for obtaining asymptotic solutions in the case of dispersion in a straight pipe of uniform cross-section, Fife and Nicholes [9] systematized and extended the theory of Taylor and others and, at the same time, clarified conditions under which it is valid. By noting the existence of two different time scales in the initial-boundary value problem they constructed approximations to this nonstandard singular perturbation problem in the form of series of developed and transient terms in powers of a parameter ϵ. Our intent is to extend the ideas introduced by Fife and Nicholes to an analysis of dispersion in pipes with slowly varying cross-sections

* Received by the editors June 25, 1990; accepted for publication (in revised form), March 26, 1993.

† Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada.

of a type previously considered by the author [12]. In that work, a solution of the Navier–Stokes equations in three dimensions is determined in the form of a power series in a small parameter ϵ chosen to characterize the slowly varying nature of a transverse section of the pipe relative to a longitudinal one. The expansion yields a sequence of boundary value problems which, upon completion of the calculations to low orders when pipe geometries permit, provides approximate perturbative solutions. A fully three-dimensional flow is difficult to solve in closed form. By assuming a slow variation in one direction we can distinguish the center-line direction and hence carry out the analysis as a quasi two-dimensional approximation (see, e.g., [27]).

The dispersal of solute in a moving fluid is a complicated process involving the interacting effects of diffusion—primarily radial—and convection. In this paper scalings are such as to make the Péclet number Pé of $O(1)$ so that lateral diffusion and convection terms are comparable. Since our primary goal is to develop a *method* for analyzing the kinds of problems mentioned above, we confine our illustration of the method—via an example in §5—to those aspects of the theory that convey the essence of the perturbative possibilities, omitting any detailed analysis with regard to variations in other parameters on which the solution will depend.

We mention two results of the theory that are exhibited here, the more important of which is not present in the uniform case. First of all, for those flows where the downstream component of the normal to a cross-section of the pipe does not depend on the transverse variables (or, for example, the large classes consisting of curved or twisted pipes of constant cross-section) the dependence of the dispersion on pipe geometry enters as the ratio of perimeter to area of a local cross-section. Secondly, the effective diffusion coefficient D_e appearing in the partial differential equation for the developed mean concentration is not necessarily larger than the molecular diffusion coefficient D of the original convection-diffusion equation. In the definition of D_e *two* terms of $O(\epsilon)$ appear, one of which is of indeterminate sign until specifics of pipe geometry are given. Thus, in the kinds of dispersion problems being considered at present the underlying transport mechanisms may differ significantly (to a first approximation) from those arising in strict Poiseuille flow.

2. Derivation of basic equations. We wish to construct asymptotic solutions for the concentration $C(\overline{x}, t)$ of dissolved matter in an incompressible viscous fluid evolving according to the equation

$$\frac{\partial C}{\partial t} + \overline{U} \cdot \nabla C = D\Delta C, \qquad (\overline{x}, t) \in R^3 \times (0, \infty). \tag{2.1}$$

$\overline{U}$ is the velocity of the convecting fluid, D is the diffusion coefficient, and Δ is the three-dimensional Laplace operator in (2.1). In addition, it is required that a function $f(\overline{x})$ and a constant α be given so that, initially,

$$C(\overline{x}, 0) = f(\overline{x}), \tag{2.2}$$

and on the boundary of the pipe

$$\frac{dC}{dn} + \alpha C = 0. \tag{2.3}$$

We disregard entrance effects and other discontinuities since local solutions for such discontinuities can usually be joined to the slowly varying solutions that we consider using the method of matched asymptotic expansions that render them uniformly valid.

Alternatively, as demonstrated by Roberts [19], an approach using center manifold theory may be employed, which permits a systematic derivation of asymptotically correct boundary conditions for models of physical problems that are based on the slowly varying approximation. When there is no transfer of concentration $C(\overline{x}, t)$ at the pipe boundary to the surrounding medium the constant α is zero. We anticipate applications of the results here not only to dispersion but to analysis of convective and diffusive heat transport as well—thus, (2.1) may govern evolution of a temperature field in which case we permit α to be positive in (2.3). (Note that α will be positive in the concentration case as well if solute is being catalyzed at the pipe walls.) Many previous studies (cf., [1], [5], [10], and [22]) focus on the evolution of the sectionally averaged concentration $\overline{C}(x, t)$. When $\alpha > 0$, however, conservation of the total heat content or solute in the fluid does not hold due to the losses at boundaries; thus, to be useful in the present problem, averaged quantities require some additional care. The initial-boundary value problem (2.1)–(2.3) differs from the one usually studied in that we allow a more general convective term $\overline{U} \cdot \nabla C$, and a more general domain Ω for the pipe in transversal and axial extent, both features reflecting the three-dimensional character of the problem. Of importance—and noted by Fife and Nicholes—is the observation that two time scales are pertinent in the reduction of (2.1)–(2.3) to a problem in dimensionless form. In the present work we wish to preserve scalings for the convecting velocity $\overline{U}$ that were previously employed by the author in [12], and to perform further scalings consistent with those employed by Fife and Nicholes. In what follows, we confine attention to those circumstances in which the underlying flow is steady since the details are readily at hand. Replication of the time-dependent case is direct.

We turn now to the task of constructing a small parameter ϵ, characterizing the slowly varying nature of the pipe boundary from parameters in the present problem. In the nondimensionalization of the Navier–Stokes equations a reference velocity U_0 is provided by the maximum (center-line) velocity of the dominant, i.e., the $O(1)$ term in the expansion for the axial flow velocity in the pipe. A reference axial length L is obtainable in a natural way from the geometry of the pipe boundary (for example, from periodicity of rotation when a pipe is slowly twisted in the downstream flow direction; cf. [13]). A characteristic time T may then be related to the other two reference quantities by $T = U_0^{-1} L$.

There is, however, in addition to L a second (shorter) length scale $\hat{L}$ available from transverse variations of the pipe—and hence, a (shorter) time scale $\hat{T}$ from $\hat{T} = \hat{L}^2 D^{-1}$. $\hat{T}$ is a characteristic measure of the time taken for solute to diffuse radially to the pipe boundary. The dimensionless parameter ϵ is then taken to be $\epsilon = \hat{L} L^{-1}$. We remark at this point that α is required to be small, of order $O(\epsilon)$ in the sequel, and that in subsequent scalings a second dimensionless parameter appears in the governing equations, $\gamma = D U_0^{-1} L^{-1}$, which we will take to be $O(1)$. In the previous [12] determination of $\overline{U}$, and in this work as well, we recognize that the approximation ("slow variations") is founded on the assumption that the pipe boundary is assumed to be slowly varying in the downstream direction X'. (Variables, while still dimensioned, are denoted by primes.) Thus, it is assumed that in the description of the pipe boundary the quantity $\epsilon X'$ appears. In the following expressions the slow variation is in the direction of the dimensionless laboratory coordinate x. To nondimensionalize the problem we let

$$x = \frac{X'}{L}, \quad y = \frac{Y'}{\hat{L}}, \quad z = \frac{Z'}{\hat{L}}, \quad t = \frac{t'}{T}, \tag{2.4}$$

$$\overline{U}(X', Y', Z') = (U_0 u, U_0 \epsilon v, U_0 \epsilon w). \tag{2.5}$$

With this scaling the boundary value problem for the velocity $\overline{U}$ is recast into one in which ϵ no longer appears in the boundary conditions but in the governing Navier–Stokes equations instead. A solution of the Navier–Stokes equations vanishing on the boundaries of the pipe is then sought in a formal power series in ϵ for the pressure p and velocity components u, v, and w.

At each stage k in the expansion it turns out that the coefficients (i.e., dimensionless velocities u_k, v_k, and w_k) are obtained as solutions of the system of partial differential equations (p.d.e.)

$$\begin{aligned} \nabla^2 u_k &= \frac{\partial p_k}{\partial x} - f_k, \\ \frac{\partial p_k}{\partial y} &= g_k, \\ \frac{\partial p_k}{\partial z} &= h_k, \\ \frac{\partial u_k}{\partial x} + \frac{\partial v_k}{\partial y} + \frac{\partial w_k}{\partial z} &= 0, \end{aligned} \tag{2.6}$$

with $u_k = v_k = w_k = 0$ on the boundary. Here, and in what follows, ∇^2 denotes the two-dimensional Laplace operator with respect to y and z, the transverse variables. The right sides f_k, g_k, and h_k are known at each step in terms of previously determined quantities and—with the help of a certain consistency requirement—the u_k, v_k, and w_k are determined uniquely.

The diffusion-convection equation (2.1) becomes

$$\nabla^2 C = \epsilon(C_t - \gamma C_{xx} + uC_x + vC_y + wC_z) \tag{2.7}$$

with

$$\frac{dC}{dn} + \alpha C = 0 \tag{2.8}$$

on the boundary, and, initially,

$$C(x, y, z, 0) = f(x, y, z) \tag{2.9}$$

after the scalings (2.4)–(2.5). In contrast with the regularity of the expansion in the equations governing the convecting velocities, the above problem is singular. Thus, a solution of (2.7)–(2.9) is sought in the form of a sum of short-term (transient) terms W_k and regular time-scale (developed) terms C_k (cf. Fife and Nicholes [9]):

$$C(\overline{x}, t) = \sum_0^\infty \epsilon^k C_k(\overline{x}, t) + \sum_0^\infty \epsilon^k W_k(\overline{x}, \tau), \qquad (\tau = \epsilon^{-1} t). \tag{2.10}$$

Substitution of (2.10) into (2.7) generates for the C_k and W_k the following sequences of problems:

$$\nabla^2 C_k = \frac{\partial C_{k-1}}{\partial t} - \gamma \frac{\partial^2 C_{k-1}}{\partial x^2} + \sum_{j=0}^{k-1} \left[u_{k-1-j} \frac{\partial C_j}{\partial x} + v_{k-1-j} \frac{\partial C_j}{\partial y} + w_{k-1-j} \frac{\partial C_j}{\partial z} \right] \tag{2.11}$$

and

(2.12)
$$\nabla^2 W_k - \frac{\partial W_k}{\partial \tau} = -\gamma \frac{\partial^2 W_{k-1}}{\partial x^2} + \sum_{j=0}^{k-1} \left[u_{k-1-j} \frac{\partial W_j}{\partial x} + v_{k-1-j} \frac{\partial W_j}{\partial y} + w_{k-1-j} \frac{\partial W_j}{\partial z} \right].$$

(Terms with negative subscripts are interpreted as absent.) The initial and boundary conditions (2.2)–(2.3) translate to the coupled conditions

(2.13)
$$C_k(\overline{x}, 0) + W_k(\overline{x}, 0) = \begin{cases} f(\overline{x}), & (k = 0), \\ 0, & (k \geq 1), \end{cases}$$

for $t = 0$, with (writing $\overline{n} = (n_x, n_y, n_z)$)

(2.14)
$$n_y \frac{\partial C_k}{\partial y} + n_z \frac{\partial C_k}{\partial z} = \begin{cases} 0, & (k = 0), \\ -\left(n_x \dfrac{\partial C_{k-1}}{\partial x} + \alpha C_{k-1} \right), & (k \geq 1), \end{cases}$$

and

(2.15)
$$n_y \frac{\partial W_k}{\partial y} + n_z \frac{\partial W_k}{\partial z} = \begin{cases} 0, & (k = 0), \\ -\left(n_x \dfrac{\partial W_{k-1}}{\partial x} + \alpha W_{k-1} \right), & (k \geq 1), \end{cases}$$

on the boundary. Because they are transient or short-scale terms the $W_k(\overline{x}, \tau)$ are required to tend exponentially to zero as τ tends to infinity.

The equations (2.11)–(2.15) are not the same as those of Fife and Nicholes: there are additional terms throughout the equations and boundary conditions arising from transverse velocity components whose presence will be felt, beginning with terms of order ϵ. We can proceed at the start, however, in a way similar to those authors since x is typically carried here as a parameter in some of the calculations. The differences arise in the special care required in consideration of the boundary condition (2.3), where C and the normal derivative dC/dn will depend on all three space variables.

3. Boundary conditions and method of solution. The method of solution is largely guided by the mathematical formulation of the preceding section. Thus, it needs to be shown that (2.11)–(2.12) with conditions (2.13)–(2.15) determine all terms in the assumed expansions (2.10) uniquely. The equations for the C_k have the form

$$\nabla^2 u = f(x, y, z, t)$$

for each $k = 0, 1, 2, \ldots$, while those for the W_k have the form

$$\nabla^2 u - \frac{\partial u}{\partial \tau} = g(x, y, z, \tau).$$

The existence and uniqueness considerations for these equations are standard (and we employ them in the sequel) when f and g are known. They are not, however, the equations that are actually solved in the construction of C_k, W_k. The latter differ in

certain important respects from those of the uniform case due to asymmetry caused by the transverse flow. We show now how the initial and boundary conditions serve to determine uniquely the leading terms in the expansions—other terms being determined analogously—and uncover thereby the interaction between terms possessing transient or nontransient behavior.

The equations for C_0 and C_1 are

$$\nabla^2 C_0 = 0 \tag{3.1}$$

and

$$\nabla^2 C_1 = \frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + \left(u_0 \frac{\partial C_0}{\partial x} + v_0 \frac{\partial C_0}{\partial y} + w_0 \frac{\partial C_0}{\partial z} \right) \tag{3.2}$$

with boundary conditions

$$n_y \frac{\partial C_0}{\partial y} + n_z \frac{\partial C_0}{\partial z} = 0 \tag{3.3}$$

and

$$n_y \frac{\partial C_1}{\partial y} + n_z \frac{\partial C_1}{\partial z} = - \left(n_x \frac{\partial C_0}{\partial x} + \alpha C_0 \right). \tag{3.4}$$

The corresponding equations and boundary conditions for the transient terms, W_0 and W_1, are

$$\nabla^2 W_0 - \frac{\partial W_0}{\partial \tau} = 0 \tag{3.5}$$

and

$$\nabla^2 W_1 - \frac{\partial W_1}{\partial \tau} = -\gamma \frac{\partial^2 W_0}{\partial x^2} + \left(u_0 \frac{\partial W_0}{\partial x} + v_0 \frac{\partial W_0}{\partial y} + w_0 \frac{\partial W_0}{\partial z} \right), \tag{3.6}$$

subject to (3.3)–(3.4) with C_0 and C_1 replaced by W_0 and W_1, respectively.

3.1. Calculation of C_0 and W_0. Equation (3.1), along with the necessary condition for existence of a solution C_1 to (3.2), produces *two* equations for the determination of C_0. To solve (3.1) together with the boundary condition (3.3) it is sufficient that C_0 be independent of y and z. The *particular* function $C_0(x,t)$ is the solution of an initial-boundary value problem in which the partial differential equation is a consequence of the existence requirement. Thus, the problem for C_0 has a solution if and only if C_0 satisfies the solvability condition

$$\int_{\Omega_x} \left(\frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + u_0 \frac{\partial C_0}{\partial x} \right) dA = - \int_{\partial \Omega_x} \left(n_x \frac{\partial C_0}{\partial x} + \alpha C_0 \right) ds, \tag{3.7}$$

where Ω_x means the cross-section at a downstream x. C_0 does not depend on (y,z) at any Ω_x. Furthermore—adapting notation of Fife and Nicholes—we write $u_0(x,y,z,t) = m(x,t)\phi(x,y,z,t)$, where ϕ is a function whose mean over Ω_x is 1 for all t. Thus, ϕ characterizes the spatial structure of the velocity profile while m characterizes its overall speed. The requirement for a solution now becomes

$$\left(\frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + m \frac{\partial C_0}{\partial x} \right) \int_{\Omega_x} dA = - \frac{\partial C_0}{\partial x} \int_{\partial \Omega_x} n_x ds - \alpha C_0 \int_{\partial \Omega_x} ds. \tag{3.8}$$

Until now—apart from its slowly varying character in the x direction—the pipe cross-section has been arbitrary. In order to proced significantly further analytically in the absence of specific pipe geometry, it becomes necessary to restrict somewhat the kinds of cross-sections to which our results apply. For pipes with boundaries such that the x-component of the normal to the boundary depends only on x, so $n_x = \xi(x)$, say, we obtain for $C_0(x,t)$ the partial differential equation

$$\frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + \left[\xi(x)\frac{P(\Omega_x)}{A(\Omega_x)} + m(x,t)\right]\frac{\partial C_0}{\partial x} + \alpha \frac{P(\Omega_x)}{A(\Omega_x)} C_0 = 0, \tag{3.9}$$

where $P(\Omega_x)$ and $A(\Omega_x)$ denote the perimeter and area of Ω_x, respectively. This specialization includes two important classes of pipes that occur in natural applications: those which are curved and those which have undergone torsion. Both effects may be present at the same time, of course. A more general example of a class of pipe geometries that possess mathematical and practical interest and for which the calculations are again simpler is a generalization of surfaces of revolution obtained as follows. A regular plane closed curve $\mathcal{C}$, which does not meet an axis l in the plane, is displaced in a rigid screw motion about l, that is, so that each point of $\mathcal{C}$ describes a helix (or circle) with l as axis. If the screw motion is a pure rotation about l, then the resulting surface is a surface of revolution $\mathcal{S}$. Choose the coordinate axes so that l is the x-axis and $\mathcal{C}$ lies in the yz-plane. If $(f(s), g(s))$ is a parameterization of $\mathcal{C}$ by arclength s, then

$$\overline{x}(s,u) = (f(s)\cos(u), f(s)\sin(u), g(s) + cu), \qquad c = \text{constant}$$

is a parameterization of the surface $\mathcal{S}$. Therefore, for a pipe whose bounding surface is the one above, $\xi(x)$ is $f(s)f'(s)$ so that

$$\int_{\partial\Omega_x} \xi(x)\, ds = \int_{\mathcal{C}} f(s)f'(s)\, ds = \int_{\mathcal{C}} \left(\frac{1}{2}f^2(s)\right)' ds = 0,$$

since $f^2(s)$ is periodic. Many other kinds of pipes whose cross-sections possess the property required of the normal mentioned above have been considered in the literature (cf., the survey paper [2]), and we shall illustrate the ideas of this paper in some detail with such an example in the next section.

Equation (3.9) is of particular interest in that it exhibits an explicit dependence of C_0 on the pipe cross-section Ω_x, that dependence entering as the *ratio* of perimeter to area of Ω_x. In fact, for pipes with cross-sections that have undergone curvature or torsion effects only, this ratio is constant, implying unexpectedly simple dependence of concentration on the pipe geometry to this (and the next) order. It should be noted that even if n_x is not solely a function of x, a partial differential equation

$$\frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + [N(x) + m(x,t)]\frac{\partial C_0}{\partial x} + \alpha M(x) C_0 = 0, \qquad M(x) = \frac{P(\Omega_x)}{A(\Omega_x)} \tag{3.10}$$

is still obtained for $C_0(x,t)$, but this time without the interesting geometrical feature mentioned above.

The initial and boundary conditions (2.13)–(2.14) are sufficient to enable us to solve (3.10) (or (3.9)) for $C_0(x,t)$: for coefficients which are suitably regular (bounded, for example), this is a uniformly parabolic equation, and existence-uniqueness theorems ensure that the problem possesses a unique solution (cf., [18]). In order to

actually write down such a solution, recourse may be had to an extensive literature devoted to transforming equations such as the one above into the classical heat equation (cf., [3] and [26]). The relevance to our application is that the latter equation has a comparatively simple and elegant integral form of solution in terms of initial data $C_0(x,0)$.

Another approach toward closed form solutions—*finite* integral transform methods—becomes possible in circumstances of steady convecting flows in which case $m = m(x)$ in (3.10), and we shall employ one of these methods in the example of §5. Alternatively, by applying a Laplace transform with respect to t, the resulting ordinary differential equation for $C_0(x,t)$ can be put (after some transformations) in eigenvalue form. The complete determination of all the quantities in following this approach would require that $C_0(x,0)$ be known.

The requirement that $C_0(x,0)$ be known in order to employ techniques such as those just mentioned, in fact, applies in general. This initial function $C_0(x,0)$ is determined at the same stage as, and in conjunction with, the transient function $W_0(x,y,z,0)$. The initial-boundary problem for W_0 is (see (3.5) and (3.3) with C_0 replaced by W_0)

$$\nabla^2 W_0 - \frac{\partial W_0}{\partial \tau} = 0, \qquad n_y \frac{\partial W_0}{\partial y} + n_z \frac{\partial W_0}{\partial z} = 0, \tag{3.11}$$

and the data for $W_0(x,y,z,0)$ that is as yet unspecified. Because of the quasi two-dimensional character of the slowly varying approximation, x is carried as a parameter essentially in these equations, and hence a Fourier method of the solution is possible. The solution $W_0(x,y,z,\tau)$ of (3.11) is expressible as

$$W_0(x,y,z,\tau) = \sum_1^\infty c_j(x)\Phi_j(x,y,z)e^{-\lambda_j \tau}, \tag{3.12}$$

where

$$c_j(x) = a_j \int_{\Omega_x} \Phi(x,y,z) W_0(x,y,z,0)\, dA, \qquad (a_j = \text{constant}).$$

In (3.12) the Φ_j are eigenfunctions of the problem

$$\nabla^2 \Phi + \lambda \Phi = 0\,, \qquad \frac{d\Phi}{dn} = 0 \quad \text{on } \partial\Omega_x, \tag{3.13}$$

and the basis functions for the Fourier expansion of $W_0(x,y,z,0)$ in (3.12). It is known that $\lambda_1 = 0$ (and all other λ_j are > 0), and that $\Phi_1(x,y,z)$ is a constant (with respect to y,z). The requirement then that $W_0 \to 0$ as $\tau \to \infty$ translates into the condition

$$c_1(x) = \frac{1}{A(\Omega_x)} \int_{\Omega_x} W_0(x,y,z,0)\, dA \equiv \overline{W}_0(x,0) = 0\,, \tag{3.14}$$

where the bar denotes the average over a cross-section Ω_x. Upon integrating all quantities in (2.13) over Ω_x , and recalling that C_0 does not depend on y or z, we see that

$$C_0(x,0) = \overline{f}(x), \tag{3.15}$$

and then

$$W_0(x,y,z,0) = f(x,y,z) - \overline{f}(x). \tag{3.16}$$

The initial-boundary value problems for C_0 and W_0 are now completely specified.

3.2. Calculation of C_1 and W_1. Since C_1 and W_1—the $O(\epsilon)$ terms in the expansion of $C(x,t)$—measure the largest effects produced by the secondary flow in the pipe, we restrict our efforts toward the calculation of C_k and W_k for $k \geq 1$ to just these two terms. Higher-order terms would be obtained in a similar way. The p.d.e. for C_1,

$$\nabla^2 C_1 = \frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + u_0 \frac{\partial C_0}{\partial x}, \tag{3.17}$$

with the boundary condition

$$n_y \frac{\partial C_1}{\partial y} + n_z \frac{\partial C_1}{\partial z} = -\left(n_x \frac{\partial C_0}{\partial x} + \alpha C_0 \right), \tag{3.18}$$

has a solution of the form

$$C_1(x,y,z,t) = m(x,t)\psi(x,y,z,t)\frac{\partial C_0}{\partial x} + \Psi(x,y,z,t) + C_1^h(x,t), \tag{3.19}$$

where ψ, Ψ, and C_1^h are solutions, respectively, of the problems

$$\begin{aligned} \nabla^2 \psi &= \phi - 1, \\ \frac{d\psi}{dn} &= 0 \quad \text{on } \partial\Omega_x, \\ \int_{\Omega_x} \psi \, dA &= 0, \end{aligned} \tag{3.20}$$

and

$$\begin{aligned} \nabla^2 \Psi &= -\left(N(x)\frac{\partial C_0}{\partial x} + \alpha M(x) C_0 \right), \\ \frac{d\Psi}{dn} &= -\left(n_x \frac{\partial C_0}{\partial x} + \alpha C_0 \right) \quad \text{on } \partial\Omega_x, \\ \int_{\Omega_x} \Psi \, dA &= 0, \end{aligned} \tag{3.21}$$

and

$$\begin{aligned} \nabla^2 C_1^h &= 0, \\ \frac{dC_1^h}{dn} &= 0. \end{aligned} \tag{3.22}$$

In all three partial differential equations above the right sides are known since $C_0(x,t)$ has been determined at the preceding stage. We know that (3.20) possesses a solution since ϕ (defined just prior to (3.8)) has a mean of one over Ω_x.

The partial differential equations in (3.21)–(3.22) above have come about from the solvability requirement for a solution for C_2 at the next stage. In (3.21), the Neumann problem for Ψ has a solution if and only if

$$\begin{aligned} \int_{\Omega_x} \left(N(x)\frac{\partial C_0}{\partial x} + \alpha M(x) C_0 \right) dA &= \int_{\partial\Omega_x} \left(n_x \frac{\partial C_0}{\partial x} + \alpha C_0 \right) ds \\ &= -\int_{\Omega_x} \left(\frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + u_0 \frac{\partial C_0}{\partial x} \right) dA \end{aligned}$$

(after use of (3.7)), which implies existence of a solution if and only if

$$\int_{\Omega_x} \left(\frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + u_0 \frac{\partial C_0}{\partial x} + N(x) \frac{\partial C_0}{\partial x} + \alpha M C_0 \right) dA = 0.$$

This condition *is* satisfied since the integrand is the left-hand side of the p.d.e. (3.10) for C_0.

From (3.22), C_1^h is independent of y and z. The actual form of $C_1^h(x,t)$ is determined from a second p.d.e. which C_1^h must satisfy, and this comes from the existence requirement for a solution of the problem for C_2 that satisfies the p.d.e.

$$\nabla^2 C_2 = \frac{\partial C_1}{\partial t} - \gamma \frac{\partial^2 C_1}{\partial x^2} + u_0 \frac{\partial C_1}{\partial x} + v_0 \frac{\partial C_1}{\partial y} + w_0 \frac{\partial C_1}{\partial z} + u_1 \frac{\partial C_0}{\partial x}. \tag{3.23}$$

After some work, the following solvability condition is seen to be required:

(3.24)

$$\left(\frac{\partial C_1^h}{\partial t} - \gamma \frac{\partial^2 C_1^h}{\partial x^2} + m(x,t) \frac{\partial C_1^h}{\partial x} \right) \int_{\Omega_x} dA + \frac{\partial C_1^h}{\partial x} \int_{\partial \Omega_x} n_x ds + \alpha C_1^h \int_{\partial \Omega_x} ds = \zeta(x,t),$$

where $\zeta(x,t)$ is known in terms of $C_0(x,t)$:

$$\begin{aligned} \zeta(x,t) = & - \int_{\Omega_x} \left[\frac{\partial}{\partial t} \left(m\psi \frac{\partial C_0}{\partial x} \right) - \gamma \frac{\partial^2}{\partial x^2} \left(m\psi \frac{\partial C_0}{\partial x} \right) + m\phi \frac{\partial}{\partial x} \left(m\psi \frac{\partial C_0}{\partial x} \right) \right] dA \\ & - \int_{\partial \Omega_x} \left[n_x \frac{\partial}{\partial x} \left(m\psi \frac{\partial C_0}{\partial x} + \Psi \right) + \alpha \left(m\psi \frac{\partial C_0}{\partial x} + \Psi \right) \right] ds \\ & - \int_{\Omega_x} \left[v_0 m \psi_y \frac{\partial C_0}{\partial x} + w_0 m \psi_z \frac{\partial C_0}{\partial x} \right] dA \\ & - \int_{\Omega_x} \left[\Psi_t - \gamma \Psi_{xx} + m\phi \Psi_x + v_0 \Psi_y + w_0 \Psi_z \right] dA. \end{aligned} \tag{3.25}$$

A simpler equation for C_1^h results—as for C_0 previously—when it is assumed that n_x depends only on x. The resulting equation

$$\frac{\partial C_1^h}{\partial t} - \gamma \frac{\partial^2 C_1^h}{\partial x^2} + \left[\xi(x) \frac{P(\Omega_x)}{A(\Omega_x)} + m(x,t) \right] \frac{\partial C_1^h}{\partial x} + \alpha \frac{P(\Omega_x)}{A(\Omega_x)} C_1^h = \frac{\zeta(x,t)}{A(\Omega_x)} \tag{3.26}$$

once more reveals dependence on pipe geometry to be on the ratio of the perimeter to the area of Ω_x.

We turn next to the determination of W_1, the order ϵ transient term. The analysis will at the same time yield the initial condition for C_1^h. The p.d.e. for W_1 is

$$\begin{aligned} \nabla^2 W_1 - \frac{\partial W_1}{\partial \tau} = & - \gamma \frac{\partial^2 W_0}{\partial x^2} + m(x, \epsilon\tau) \phi(x,y,z,\epsilon\tau) \frac{\partial W_0}{\partial x} \\ & + v_0(x,y,z,\epsilon\tau) \frac{\partial W_0}{\partial y} + w_0(x,y,z,\epsilon\tau) \frac{\partial W_0}{\partial z} \\ \equiv & \; H(x,y,z,\tau). \end{aligned} \tag{3.27}$$

To determine the condition for transience we first integrate the terms in (3.27) over Ω_x, noting that

$$\int_{\Omega_x} \nabla^2 W_1 \, dA = \int_{\partial \Omega_x} \frac{dW_1}{dn} \, ds = - \int_{\partial \Omega_x} \left(n_x \frac{\partial W_0}{\partial x} + \alpha W_0 \right) ds$$

from the boundary condition for W_1, and then integrate with respect to τ, getting

$$\overline{W}_1(x,\tau) = \overline{W}_1(x,0) - \int_0^\tau \overline{H}(x,\tau')\,d\tau' - \int_0^\tau \int_{\partial\Omega_x} \left(n_x \frac{\partial W_0}{\partial x} + \alpha W_0 \right) ds\,d\tau'. \tag{3.28}$$

Letting $\tau \to \infty$ in the last equation (and remembering that $W_1(x,\tau) \to 0$ is a requirement of transience) gives

$$\overline{W}_1(x,0) = \int_0^\infty \overline{H}(x,\tau)\,d\tau + \int_0^\infty \int_{\partial\Omega_x} \left(n_x \frac{\partial W_0}{\partial x} + \alpha W_0 \right) ds\,d\tau, \tag{3.29}$$

and this is the determining equation for $\overline{W}_1(x,0)$. Equation (3.29) serves to determine C_1^h also. For, the initial condition $C_1(x,y,z,0) + W_1(x,y,z,0) = 0$, and the form of the solution $C_1(x,y,z,t)$ gives upon integration over Ω_x,

$$C_1^h(x,0) = -\overline{W}_1(x,0), \tag{3.30}$$

where the right side is once again known in terms of previously determined quantities. Finally, the initial condition for $W_1(x,y,z,0)$ itself is recovered from its average $\overline{W}_1(x,0)$ by tracing back to (2.13) via (3.29)–(3.30) and (3.19).

4. The large-time concentration and effective diffusion coefficient. The kinds of pipe flows under consideration here lead to $O(\epsilon)$ transverse velocities which, in their turn, distort the effective axial diffusion. The latter is a combination of several effects. In addition to molecular diffusion in the transverse direction and distortion of the scalar field by the mean downstream convecting flow—the primary effects in the uniform pipe situation—the secondary flow terms contribute an additional component in the expression for the effective diffusion coefficient, which reflects a further distortion of the scalar field by the lateral convection. We have

$$\overline{C}(x,t) = \frac{1}{A(\Omega_x)} \int_{\Omega_x} C(x,y,z,t)\,dA$$

and then—upon retaining terms in the expansion for $C(x,y,z,t)$ in their explicit form to $O(\epsilon)$—we have that

(4.1)

$$\overline{C}(x,t) = [C_0(x,t) + \epsilon C_1^h(x,t)] + \epsilon \left[\int_{\epsilon^{-1}t}^\infty \overline{H}(x,\sigma)\,d\sigma + \int_{\epsilon^{-1}t}^\infty \int_{\partial\Omega_x} \left(n_x \frac{\partial W_0}{\partial x} + \alpha W_0 \right) ds\,d\sigma \right] + O(\epsilon^2).$$

When t is appreciably large, say $t \gg \epsilon$, the second term in square brackets, which is comprised solely of transient terms, is small, and we neglect it. The remaining term in (4.1) is denoted by

$$C_d(x,t) \equiv C_0(x,t) + \epsilon C_1^h(x,t) \tag{4.2}$$

(*developed mean value* in the terminology of [9]), and we seek an equation satisfied by C_d. With (3.10), (3.24), and (3.25)—and after some algebra—we find

$$\frac{\partial C_d}{\partial t} - \left[\gamma - \epsilon m^2 \overline{\phi\psi} - \frac{\epsilon}{A(\Omega_x)} \int_{\partial\Omega_x} n_x m \psi\,ds \right] \frac{\partial^2 C_d}{\partial x^2}$$

$$
\begin{aligned}
&+ \left[(N+m) + \frac{\epsilon}{A(\Omega_x)} \int_{\partial\Omega_x} n_x (m\psi)_x \, ds + \epsilon \left(m^2 \overline{\phi\psi_x} + m m_x \overline{\phi\psi}\right)\right. \\
&\qquad \left. + \frac{\epsilon m}{A(\Omega_x)} \int_{\Omega_x} (v_0\psi_y + w_0\psi_z) \, dA + \frac{\epsilon\alpha}{A(\Omega_x)} \int_{\partial\Omega_x} m\psi \, ds \right] \frac{\partial C_d}{\partial x} \\
&+ [\alpha M] \, C_d \\
&= \epsilon G(x,t) + O(\epsilon^2),
\end{aligned} \tag{4.3}
$$

where

$$
\begin{aligned}
G(x,t) \equiv &- \frac{1}{A(\Omega_x)} \int_{\Omega_x} (\Psi_t - \gamma\Psi_{xx} + m\phi\Psi_x + v_0\Psi_y + w_0\Psi_z) \, dA \\
&- \frac{1}{A(\Omega_x)} \int_{\partial\Omega_x} (n_x\Psi_x + \alpha\Psi) \, ds.
\end{aligned} \tag{4.4}
$$

The partial differential equation for C_d appears complicated; we note, however, that except for the adjusted coefficients it has the same form as the one for C_0. In specific practical situations—for example, flows through pipes with sections possessing periodicity of n_x, or flows in which the velocity has the properties that $m(x,t)$ is constant or the transverse component of lowest order is radial—a number of the terms comprising the coefficients will drop out of further consideration.

In order to fully specify a problem for C_d there remains the consideration of an initial condition for C_d. For this, we generalize slightly an auxiliary "P-problem" as set up by Fife and Nicholes to conveniently organize the calculations for both $\overline{W}_1(x,0)$ and $C_1^h(x,0)$—and hence, for $C_d(x,0)$—to three (nearly identical) such problems. To see how these intermediate functions can be introduced in a natural manner, substitute for H into $\overline{W}_1(x,0)$ while noting that the term in $\overline{W}_{0_{xx}}(x,\tau)$ is identically zero, and then, because derivatives of W_0 tend to zero exponentially as τ becomes infinite, within the current $O(\epsilon)$ of accuracy,

$$
\begin{aligned}
\overline{W}_1(x,0) = &\int_{\Omega_x} m(x,0)\phi(x,y,z,0) \int_0^\infty \frac{\partial W_0}{\partial x}(x,y,z,\tau) \, d\tau \, dA \\
&+ \int_{\Omega_x} v_0(x,y,z,0) \int_0^\infty \frac{\partial W_0}{\partial y}(x,y,z,\tau) \, d\tau \, dA \\
&+ \int_{\Omega_x} w_0(x,y,z,0) \int_0^\infty \frac{\partial W_0}{\partial z}(x,y,z,\tau) \, d\tau \, dA \\
&+ \int_{\partial\Omega_x} n_x \int_0^\infty \frac{\partial W_0}{\partial x}(x,y,z,\tau) \, d\tau \, ds.
\end{aligned} \tag{4.5}
$$

At this point the calculations may be organized as follows. Introduce functions P, Q, and R defined by the τ integrals just above:

$$
P \equiv \int_0^\infty \frac{\partial W_0}{\partial x} \, d\tau, \qquad Q \equiv \int_0^\infty \frac{\partial W_0}{\partial y} \, d\tau, \qquad R \equiv \int_0^\infty \frac{\partial W_0}{\partial z} \, d\tau.
$$

But now, it can be seen that P is the solution of the problem

$$
\nabla^2 P = \overline{f}_x - f_x, \quad \left.\frac{dP}{dn}\right|_{\partial\Omega_x} = 0, \quad \int_{\Omega_x} P \, dA = 0, \tag{4.6}
$$

where f is the given initial concentration, while the problems for Q and R are similar to the one for P, differing only in the fact that the right sides of their respective

partial differential equations are $-f_y$ and $-f_z$. Once $\overline{W}_1(x,0)$ (and hence, $C_1^h(x,0)$) is determined, the initial condition for $C_d(x,0)$ is obtained from

(4.7)

$$C_d(x,0) = \overline{f}(x) + \epsilon \left(\int_{\Omega_x} [m(x,0)\phi(x,y,z,0)P(x,y,z) + v_0(x,y,z,0)Q(x,y,z) + w_0(x,y,z,0)R(x,y,z)]\, dA + \int_{\partial\Omega_x} n_x P(x,y,z)\, ds \right).$$

The calculation of integral moments of $C(\overline{x},t)$ has been employed by Aris [1] and others (e.g., [5] and [30]) as a way of constructing the concentration distribution to various degrees of accuracy and as a means of linking the effect of the initial distribution of $C(\overline{x},t)$ on its asymptotic form. In particular, the center of mass of the average concentration

$$x_m(t) = \int_{-\infty}^{\infty} x\overline{C}(x,t)\, dx, \qquad x_m(0) = 0$$

in a coordinate system moving with a mean flow $m(t)$ approaches a limiting offset position

$$b = \lim_{t\to\infty} \left(x_m(t) - \int_0^t m(s)\, ds \right).$$

Fife and Nicholes have shown (see [9] for details) that this offset position can be related to the *initial displacement* of the center of mass of the *developed mean*, i.e., from $C_d(x,0)$. Since $W_0(\overline{x},t)$ tends to zero exponentially, $x_m(t)$ can be replaced by

$$x_d(t) = \int_{-\infty}^{\infty} xC_d(x,t)\, dx = x_d(0) + \int_0^t m(s)\, ds \tag{4.8}$$

(neglecting terms of $O(\epsilon^2)$). We then have, after the relevant substitutions, the expression

(4.9)

$$x_d(0) = \epsilon \left(\int_{-\infty}^{\infty} \int_{\Omega_x} [m(x,0)\phi(x,y,z,0)P(x,y,z) + v_0(x,y,z,0)Q(x,y,z) + w_0(x,y,z,0)R(x,y,z)]\, dA\, dx + \int_{-\infty}^{\infty} \int_{\partial\Omega_x} n_x P(x,y,z)\, ds\, dx \right)$$

for the offset position b here. The term with P was previously related to b by Fife and Nicholes. In the present problem we have two additional integrals that contribute to the displacement of the center of mass: first effects due to the transverse velocity component as expressed by the integrals involving v_0, w_0, and, a boundary integral that relates the change in the offset position to nonuniformities in pipe cross-sections.

In the equation (4.3) for $C_d(x,t)$ the coefficient of the term in the second derivative of C_d is of particular interest since it is the one which involves the diffusion coefficient (when quantities revert to their dimensioned forms). We write

$$D_e = D - \frac{U_0^2}{D} \left[m^2 \overline{\phi\psi} + \frac{1}{A(\Omega_{x'})} \int_{\partial\Omega_{x'}} m\psi\xi\, ds \right] \tag{4.10}$$

for this coefficient in its dimensioned form. The part

$$m^2\overline{\phi\psi} = -m^2\frac{1}{A(\Omega_{x'})}\int_{\Omega_{x'}} |\nabla\psi|^2\, dA \tag{4.11}$$

is ≤ 0, in accord with the situation for the Poiseuille case of flow through a pipe of uniform cross-section. The part

$$\frac{1}{A(\Omega_{x'})}\int_{\partial\Omega_{x'}} n_x m\psi\, ds, \tag{4.12}$$

however, is of undetermined sign in the absence of specific pipe geometry and initial conditions. For certain classes of flows it can be easily seen that this term also is ≤ 0. On the other hand, for the interesting class exhibiting flow reversal (e.g., oscillatory pressure driven flows) this term can be > 0, thereby inhibiting rather than augmenting radial diffusion.

5. An example: spiralling circular pipe. Much attention has been devoted to flow through a coiled pipe because of its practical importance (see, for example, the survey by Van Dyke [27]). Our intent now is to illustrate how the theory of preceding sections can be applied to provide a perturbative solution to the dispersion problem for flows in such pipes. First of all, it is necessary to extract certain results concerning low-order terms in the expansions for the three velocity components from our previous study [12] of flows in pipes with slowly varying cross-sections. We consider a spiralling circular pipe that can be viewed as being generated by translating a circle of unit radius so that its center moves along a helix of radius a while its plane remains normal to the axis of the helix. (This differs from what is usually understood as a helically coiled pipe but only in third-order terms). Introduce *local* coordinates (r, θ) at a station x defined by

$$y - a\cos x = r\cos\theta, \qquad z - a\sin x = r\sin\theta.$$

At the stage $k = 0$ in the problems for the velocity components defined by the system of equations (2.6) we have as the solution for u_0 the Poiseuille-like flow through a straight circular pipe:

$$u_0(x, r, \theta) = -\frac{1}{4}\frac{dp_0}{dx}(1 - r^2). \tag{5.1}$$

With u_0 now given, we solve for v_0, w_0 from the pair of equations

$$\Delta(v_{0z} - w_{0y}) = 0, \qquad v_{0y} + w_{0z} = -u_{0x}, \tag{5.2}$$

subject to $v_0 = w_0 = 0$ on the boundary. The first of these is a *consistency relation* for the system of equations (2.6) at the stage $k = 0$, and it is obtained at the stage $k = 2$. Here, we quote the result that, *to terms of order* ϵ, *the flow velocity is independent of* θ, *and, since* $d^2p_0(x)/dx^2 = 0$, $dp_0(x)/dx$ *is a constant*, which we denote by $-\mu$. Then

(5.3)

$$\begin{aligned} u_0(x, r, \theta) &= \frac{\mu}{4}(1 - r^2), \quad v_0(x, r, \theta) = -\frac{\mu}{4}a(1 - r^2)\sin x, \\ w_0(x, r, \theta) &= \frac{\mu}{4}a(1 - r^2)\cos x \end{aligned}$$

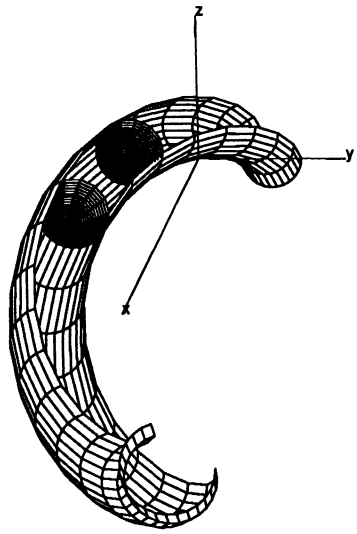


FIG. 1(a). *Unperturbed (zero-order) velocity profile in the spiralling circular pipe.*

where, as in Poiseuille flow μ is the pressure gradient along the pipe. Figures 1(a) and 1(b) display the unperturbed (i.e., u_0) and perturbed (i.e., by v_0, w_0) flow fields at typical axial stations of the pipe. We notice that the perturbed field is slightly inclined to the axis of the basic Poiseuille flow in a straight pipe. The correction to the zero order profile is directed normal to a cross-section and in such a way that the velocity distribution shifts toward the inner wall of the pipe. The resulting shortened fluid path means that the flux ratio, a quantity of interest, defined to be the flow rate through the curved pipe divided by that in a straight pipe under the same pressure gradient, is actually increased by the slight coiling. Further effects of curvature (together with effects of torsion and of the slight deviation from circularity of the pipe) would appear in subsequent orders of the calculations. The circular asymmetry causes enhancement of mixing and relocation of maximum and minimum wall shear stresses from that of straight pipe flow. Since they are the most important terms in applications, we proceed with the analysis for the concentration to terms of order ϵ only. By basing the transverse length scale on a^{-1}, the ratio of the radius of a cross-section of the pipe to the radius of the center-line helix, say, the parameter ϵ could be taken to be a measure of the center-line curvature of the helical pipe. We do not attempt detailed parameter studies with respect to other parameters of order unity because this warrants special consideration beyond the intentions of the example treated here.

For the pipe described in the first paragraph we have, since $n_x = -a \sin(t - s)$, that

$$N(x) = \int_{\partial\Omega_x} n_x \, ds = \int_0^{2\pi} -a \sin(t - s) \, ds = 0.$$

Since $(1/A(\Omega_x)) \int_{\Omega_x} (1 - r^2) \, dA = \frac{1}{2}$, we will also have

$$m(x,t) = \frac{\mu}{8} \equiv U, \qquad M(x) = \frac{P(\Omega_x)}{A(\Omega_x)} = 2. \tag{5.4}$$

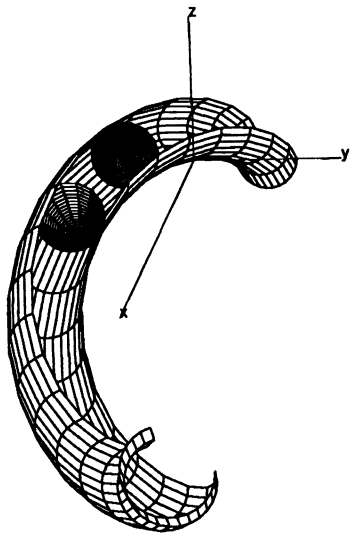


FIG. 1(b). *Perturbed velocity profile incorporating first-order transverse velocities.*

The p.d.e. (3.10) for C_0 then is

$$\frac{\partial C_0}{\partial t} - \gamma \frac{\partial^2 C_0}{\partial x^2} + \frac{\mu}{8} \frac{\partial C_0}{\partial x} + 2\alpha C_0 = 0. \tag{5.5}$$

We can reduce this equation to the equation for pure diffusion by the change of variables

$$C_0 = he^{-2\alpha t}, \qquad x = \xi + \frac{\mu}{8} t,$$

obtaining

$$h_t = \gamma h_{\xi\xi} .$$

A common set of initial conditions for $C_0(x,t)$ is to require

$$C_0(x,0) = \overline{f}(x) = \delta(x - x_0), \tag{5.6}$$

which translates into

$$h(\xi,0) = \delta(\xi - x_0).$$

Such an initial condition may be used to model a concentrated initial solute input distributed uniformly over a cross-section of the pipe at a station x_0. In this case the solution for (5.5)–(5.6) is

$$C_0(x,t) = \frac{e^{-2\alpha t}}{\sqrt{4\pi\gamma t}} e^{-\frac{(x-Ut-x_0)^2}{4\gamma t}}, \qquad \left(U \equiv \frac{\mu}{8}\right). \tag{5.7}$$

The initial delta function profile for C_0 at x_0 is propagated with speed U along lines in xt-space whose slopes are proportional to the axial pressure gradient μ. Figure 2

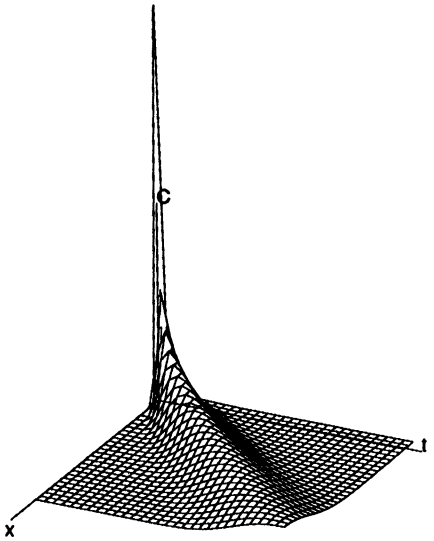


FIG. 2(a). *Evolution of zero-order concentration with initial delta-function data and parameter values* $\gamma = .3$, $\alpha = 0$ (*Neumann boundary condition*).

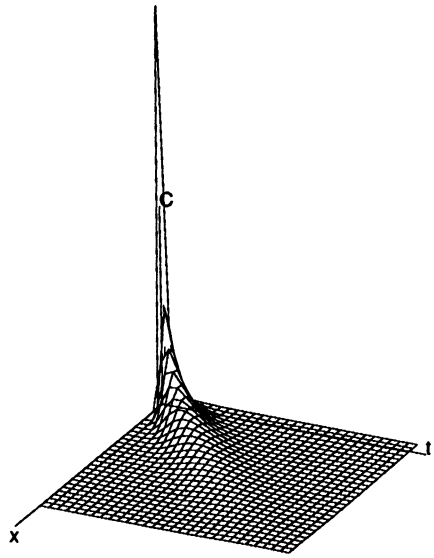


FIG. 2(b). *Evolution of zero-order concentration with initial delta-function data and parameter values* $\gamma = .3$, $\alpha = .1$ (*mixed boundary condition*).

graphs concentration profiles for two values of the parameter α—a parameter that characterizes the rate of the reaction catalyzed by the pipe walls or damping in the case of transport of heat—corresponding to Neumann ($\alpha = 0$) and mixed ($\alpha > 0$) boundary conditions.

We give now a second example of an explicit solution of (5.5) for the concentration $C_0(x, t)$, which evolves from initial data quite different in kind from the delta function

above. Interest derives from the fact that this initial profile replicates an initially oscillating concentration input. The corresponding evolution of an oscillatory form of solution is a direct consequence of the initial-value problem posed—no additional ad hoc assumptions in the modelling process are necessary. Specifically, we take

$$C_0(x,0) = Ai\,[2^{\frac{1}{3}}(x - x_0)] \tag{5.8}$$

($Ai\,(z)$ is the Airy function of the first kind) and seek an *R-separating solution* of (5.5) as outlined in Morse and Feshback or [29]. Such solutions arise from an enlarged separability definition of solutions, and they have the form

$$C(x,t) = R(u,v)U_\lambda(u)V_\lambda(v), \qquad u = u(x,t)\,, \quad v = v(x,t),$$

depending upon a parameter λ. When we carry out the analysis for generating such solutions we find in our case that

(5.9)

$$C_0(x,t) = 2^{\frac{1}{3}}\, e^{-2\alpha t}\, e^{-2\gamma t(x-Ut+2\gamma^2t^2-x_0)}\, e^{-2x_0\gamma t-\frac{4}{3}\gamma^3t^3}\, Ai\,[2^{\frac{1}{3}}(x - Ut + 2\gamma^2t^2 - x_0)].$$

From standard references on special functions we have that $Ai(z)$ for $z < 0$ is oscillatory and, asymptotically,

$$Ai(z) \simeq (4\pi)^{-\frac{1}{2}} z^{-\frac{1}{4}} \exp\left(-\frac{2z^{\frac{3}{2}}}{3}\right)$$

for $z \to +\infty$. Thus, the initial input $C_0(x,0)$ in (5.8) oscillates for values of x up to x_0 and it then decays exponentially to zero for x beyond x_0. We note that the resulting concentration distribution (5.9) is of the form

$$C_0(x,t) = e^{-2\alpha t}h(x - Ut, t).$$

Zeros in the oscillatory region lie on parallel parabolas $x - Ut + 2\gamma^2t^2 - x_0 =$ constant, while for $\xi = x - Ut > x_0$ we observe the exponentially decaying behavior

$$C_0(x,t) \simeq t^{-\frac{1}{2}} e^{2\gamma t(x-Ut-x_0)-\alpha t}$$

for all α. The evolution of the solutions (5.9) for $C_0(x,t)$ in xt-space is depicted in Fig. 3 for both zero and nonzero α, values which correspond to circumstances of nonreacting (insulating) or reacting (noninsulating) pipe boundaries, respectively.

Next, we address the problem for W_0, the first transient term in the expansion for $C(\overline{x},t)$. For this, when the initial data is axisymmetric, we employ the finite Hankel transform of the second kind over the interval $[0,1]$ (cf. [23]). Although (r,θ) are local coordinates, the Laplace operator transforms from laboratory rectangular coordinates into these coordinates in exactly the same way as for the usual polar coordinates. Thus, the problem for W_0 is

$$\begin{aligned} \frac{\partial^2 W_0}{\partial r^2} + \frac{1}{r}\frac{\partial W_0}{\partial r} + \frac{1}{r^2}\frac{\partial^2 W_0}{\partial \theta^2} &= \frac{\partial W_0}{\partial \tau} \\ \frac{\partial W_0}{\partial r}(x,1,\theta,\tau) &= 0 \\ W_0(x,r,\theta,0) &= f(x,r,\theta) - \overline{f}(x), \end{aligned} \tag{5.10}$$

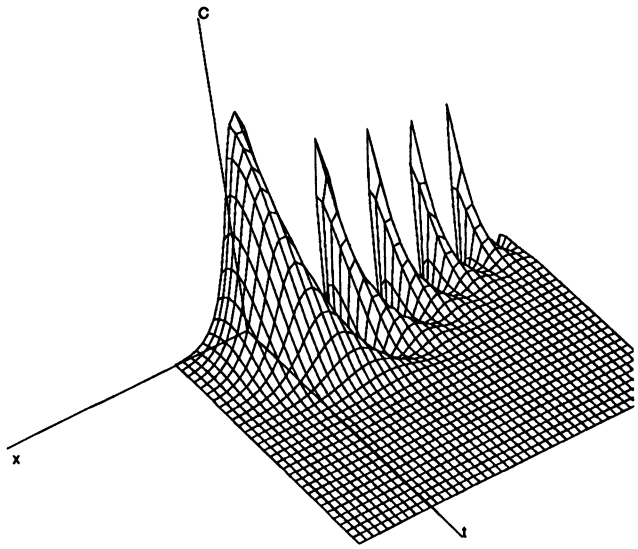


FIG. 3(a). *Evolution of zero-order concentration with initial Airy-function data and parameter values* $\gamma = .3$, $\alpha = 0$ (*Neumann boundary conditions*).

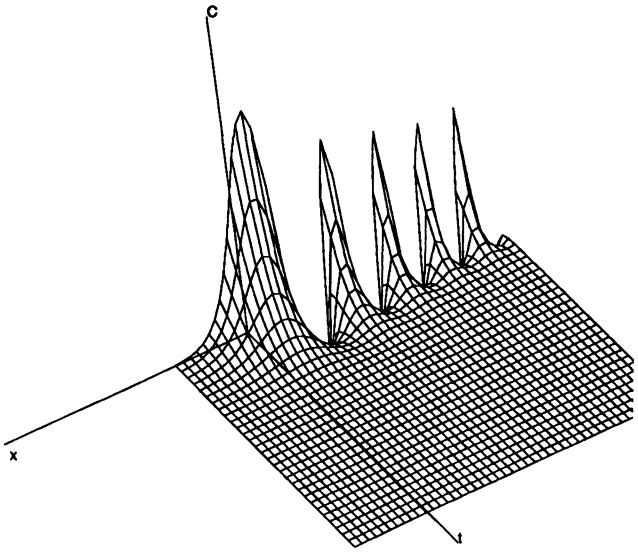


FIG. 3(b). *Evolution of zero-order concentration with initial Airy-function data and parameter values* $\gamma = .3$, $\alpha = 2$ (*mixed boundary conditions*).

where $f(x, r, \theta)$ is the initial data for $C(\overline{x}, t)$. When this initial data is axisymmetric we extract the usual θ-dependence via a Fourier decomposition and then, upon noting the property

$$\phi_{rr} + \frac{1}{r}\phi_r - \frac{m^2}{r^2}\phi \longmapsto -\lambda^2\Phi(\lambda) - \frac{2}{\pi\lambda}\phi'(1) \tag{5.11}$$

for a Hankel transform pair (ϕ, Φ), are left with a first-order ordinary differential equation in τ which is easily solved. The solution of the initial-boundary problem (5.10) for W_0 is

(5.12)
$$W_0(x,r,\theta,\tau) = \sum_{n=1}^{\infty}\sum_{m=-\infty}^{\infty} 2J_m(\xi_n r)\, e^{-\xi_n^2\tau} e^{im\theta} \int_0^1 [f(x,u) - \overline{f}(x)] u J_m(\xi_n u)\, du.$$

(Here, the ξ_n are the zeros of $J_m'(\xi_n) = 0$.) At large times the transverse variations in concentration represented by the first transient term W_0 are much smaller than the axial and transverse variations represented by C_0 and C_1. Moreover, the series in (5.12) converges very rapidly so that for large τ it may be possible to usefully retain only the first term in approximations. When the initial data is not axisymmetric we must proceed differently. In this case a Laplace transform with respect to τ is called for, followed by a Green's function approach. (See the calculation for W_1.)

We direct our attention now to calculating the $O(\epsilon)$ quantities C_1 and W_1, beginning with the initial-boundary value problem for C_1. We recall the form (3.19) of the solution for C_1:

$$C_1(x,r,\theta,t) = m(x,t)\psi(x,r,\theta,t)\frac{\partial C_0}{\partial x} + \Psi(x,r,\theta,t) + C_1^h(x,t).$$

The three terms comprising the right side of the representation for C_1 are dealt with separately. The term involving the derivative of C_0 is known at this stage: $\partial C_0/\partial x$ is given, for example, from (5.6) or (5.8), depending on the imposed initial conditions. The function ψ depends only on the dynamics of the flow—here, it is

$$\psi(x,r,\theta,t) = \frac{1}{24}(-3r^4 + 6r^2 - 2), \tag{5.13}$$

while $m(x,t)$ is given in (5.4). The calculation of the remaining two terms is a lengthy and more involved process. As for the initial-value problem for $C_1^h(x,t)$, we note that the operator in the equation for C_1^h is the same as the one for C_0, so we may employ the previously solved-for delta-function solution (5.7) as a Green's function and use it to express the solution of the problem

$$\begin{aligned} \frac{\partial C_1^h}{\partial t} - \gamma\frac{\partial^2 C_1^h}{\partial x^2} + \frac{\mu}{8}\frac{\partial C_1^h}{\partial x} + 2\alpha C_1^h &= \frac{1}{\pi}\zeta(x,t), \\ C_1^h(x,0) &= -\overline{W}_1(x,0) \end{aligned} \tag{5.14}$$

as an integral. Before we can do this, however, we must carry out two preliminary tasks. They are as follows: providing at this stage a representation for the initial data $\overline{W}_1(x,0)$ and providing an expression for the driving term $\zeta(x,t)$ as given by (3.25). As for the first task, the data $\overline{W}_1(x,0)$ is given by substitution of the expressions in (5.3) into equation (4.5):

(5.15)
$$\begin{aligned} \overline{W}_1(x,0) = {} & \frac{\mu^2\pi}{32}\int_0^{\infty}\int_0^{2\pi}\int_0^1 (1-r^2)\Big[\frac{\partial W_0}{\partial x}(x,r,\theta,\tau) - a\sin(x-\theta)\frac{\partial W_0}{\partial r}(x,r,\theta,\tau) \\ & + \frac{1}{r}a\cos(x-\theta)\frac{\partial W_0}{\partial \theta}(x,r,\theta,\tau)\Big] r\,dr\,d\theta\,d\tau \\ & - \int_0^{\infty}\int_0^{2\pi} a\sin(x-\theta)\frac{\partial W_0}{\partial x}(x,1,\theta,\tau)\,d\theta\,d\tau, \end{aligned}$$

with W_0 obtained from (5.12) when the initial concentration $C(\overline{x},0)$ is axisymmetric. $\overline{W}_1(x,0)$ is recovered most easily, however, not from this last expression but from the auxiliary functions P, Q, and R, defined just before (4.6) in the preceding section. Here, these functions are solutions of problems of the form

$$\frac{\partial^2 g}{\partial r^2}+\frac{1}{r}\frac{\partial g}{\partial r}+\frac{1}{r^2}\frac{\partial^2 g}{\partial \theta^2}=\tilde{\rho}(x,r,\theta), \tag{5.16}$$

$$\frac{\partial g}{\partial r}(x,1,\theta)=0, \qquad \int_0^{2\pi}\int_0^1 g(x,r,\theta)\,r\,dr\,d\theta=0,$$

where $\tilde{\rho}(x,r,\theta)$ is known from the prescribed data $C(\overline{x},0)=f(\overline{x})$; for example, $\tilde{\rho}=\overline{f}_x-f_x$ in the equation for P. When f is axisymmetric and possesses a Fourier series in conventional polar coordinates, one may also obtain g as a Fourier series in a standard way. In any case, the solution when the initial data is arbitrary can be expressed as

$$g(x,r,\theta)=\int_{\Omega_x} N\tilde{\rho}\,dA+\overline{g},$$

where $\overline{g}$ is a constant equal to the average value of g over Ω_x. In this last expression N is a Neumann function for

$$\nabla^2 N=\delta(r,\theta;r',\theta')-\left(\int_{\Omega_x} dA\right)^{-1},$$

$$\frac{dN}{dn}=0.$$

Further progress toward explicit representations can be made in special cases. For example, a separable form $f(x,r,\theta)=a(x)b(r,\theta)$ for the initial data is often assumed. Then a separable form of solution for g can be found in the standard way. Furthermore, it has been shown (see [11] or [16]) that a radially distributed pulse, for example, can be generalized to a prescribed distribution at $x=0$. Alternatively, if we initially take a delta function profile for $C(\overline{x},0)$, the calculations simplify once more.

As Ψ is involved in the construction of $\zeta(x,t)$ we must—in order to complete the second task in specifying a problem for C_1^h—address the problem for Ψ. This problem is

$$\nabla^2\Psi=-2\alpha C_0(x,t), \tag{5.17}$$

$$\frac{d\Psi}{dn}=a\sin(x-\theta)\frac{\partial C_0}{\partial x}-2\alpha C_0(x,t),$$

using the additional condition

$$\int_{\Omega_x}\Psi\,dA=0 \tag{5.18}$$

to fix the arbitrary constant. Define

$$b(x,t,\eta)=\int_0^\eta \left.\frac{d\Psi}{dn}\right|_{\partial\Omega_x} d\eta'=a\frac{\partial C_0}{\partial x}[\cos(x-\eta)-\cos x]-\alpha C_0\eta.$$

We then have the representation

$$\Psi(x,r,\theta,t)=-\frac{r}{\pi}\int_{-\pi}^{\pi} b(x,t,\eta)\frac{\sin(\theta-\eta)}{1-2r\cos(\theta-\eta)+r^2}\,d\eta+\text{const}$$

for Ψ. By now writing $\theta' = \theta - \eta$, this last expression becomes

$$\Psi(x,r,\theta,t) = \frac{-C_0}{8} - \frac{r}{\pi}\int_{-\pi}^{\pi} \frac{[a(\partial C_0/\partial x)\cos(x-\theta+\theta') + \alpha C_0]\sin\theta'}{1-2r\cos\theta'+r^2}\,d\theta' \tag{5.19}$$

after a direct calculation of the condition in (5.18) has been incorporated to fix the constant. Via contour integration, or otherwise, the last integral can be explicitly evaluated to give

$$\Psi(x,r,\theta,t) = a\frac{\partial C_0}{\partial x} r\sin(x-\theta) + \frac{1}{2}\alpha C_0(1-r^2) - \frac{1}{4}\alpha C_0. \tag{5.20}$$

After somewhat lengthy but straightforward calculations for the evaluation of $\zeta(x,t)$ we find that

$$\zeta(x,t) = \pi\left[-a^2\frac{\partial^2 C_0}{\partial x^2} + \left(\frac{a^2\mu}{8} - \frac{\alpha\mu}{96}\right)\frac{\partial C_0}{\partial x} + \frac{\alpha^2}{2}C_0\right]. \tag{5.21}$$

Returning to the equation (5.14) for C_1^h we see that a fundamental solution for

$$LC_1^h = \frac{1}{\pi}\zeta = a_2\frac{\partial^2 C_0}{\partial x^2} + a_1\frac{\partial C_0}{\partial x} + a_0 C_0 \tag{5.22}$$

(L defined by (5.14)) is just the previously determined $C_0(x,t)$ for initial delta function data given above in (5.7). Alternatively, with the change in variables

$$C_0 = he^{-2\alpha t}, \qquad x = \xi + Ut,$$

the problem for h is

$$\begin{aligned} \frac{\partial h}{\partial t} - \gamma\frac{\partial^2 h}{\partial \xi^2} &= e^{2\alpha t}\zeta(\xi,t), \\ h(\xi,0) &= -\overline{W}_1(\xi,0). \end{aligned} \tag{5.23}$$

With the aid of a standard Green's function we write

$$h(\xi,t) = \int_0^t\int_0^\infty g(\xi,t|\xi',t')e^{2\alpha t'}\zeta(\xi',t')\,dt'\,d\xi' + \int_0^\infty g(\xi,t|\xi',0)\,[-\overline{W}_1(\xi',0)]\,d\xi' \tag{5.24}$$

and express the solution for C_1^h as

$$C_1^h(x,t) = e^{-2\alpha t}h(x-Ut,t). \tag{5.25}$$

We are finally in a position to write down the solution of the equation for C_1. When we collect (5.13), (5.20), and (5.25) we have

$$\begin{aligned} C_1(x,r,\theta,t) = {} & \frac{\partial C_0}{\partial x}\left[ar\sin(x-\theta) - \frac{1}{192}(3r^4 - 6r^2 + 2)\right] \\ & + C_0\left[\frac{1}{2}\alpha(1-r^2) - \frac{1}{4}\alpha\right] + C_1^h(x,t) \end{aligned} \tag{5.26}$$

altogether. The paraboloidal term $C_0[\frac{1}{2}\alpha(1-r^2) - \frac{1}{4}\alpha]$ is most prominent for small r. For $r \to 1$ the torsion term $(\partial C_0/\partial x)ar\sin(x-\theta)$ becomes appreciable. All terms are

proportional to C_0 or $(\partial C_0/\partial x)$ and inherit their decay properties. The expression $(1/192)(3r^4 - 6r^2 + 2)(\partial C_0/\partial x)$ may be interpreted in the following way. For an initially concentrated release of solute

$$C_1(x, r, \theta, 0) = 2\pi\delta(x)\delta(r)\delta(\theta)$$

at a point (x, r, θ) of a cross-section Ω_x this expression is the average of the term $u_0(\partial C_0/\partial x)$ in (3.17). It has the effect of displacing the center of mass of the concentration distribution radially and longitudinally proportional to $u_0(\partial C_0/\partial x)$.

Further comment is warranted concerning individual terms in the above solution for C_1. The middle two at an axial station are multiples of C_0 and $(\partial C_0/\partial x)$, which are both homogeneous solutions of (5.22) and may be responsible for observed resonances of C_0 in oscillatory axial profiles as given in Dravid et al. [8]. In a numerical experimental study these authors exhibit computed axial profiles which "display a startling characteristic feature: the axial temperature or heat flux profiles show large amplitude oscillations that decay and damp out in the fully developed region." They explain the rather complex first cycle in the oscillations and speculate that the second and subsequent oscillations are resonances of the first. Secondly, the torsion term in $\sin(x - \theta)$ is of interest in that it exhibits first effects on the concentration due to the torsion of the convecting velocity. It is clear that the minimum point with respect to θ in the concentration profile will rotate with the axial variable x. These authors found this to be of interest in their numerically computed development of the temperature field since it was in contradiction to the general belief. Thirdly, the "very peculiar temperature profile in the fully developed region" computed numerically in [8] can be qualitatively verified by our solution as follows. The computation takes place at a fixed θ and x so it can be done where $\sin(x - \theta) = 0$. Then the critical points of the radial profile are found to be at $r = 0$, and at

$$r = \frac{-\alpha C_0 + \sqrt{(\alpha C_0)^2 + \left(\frac{1}{8}\frac{\partial C_0}{\partial x}\right)^2}}{\frac{1}{8}\frac{\partial C_0}{\partial x}} < 1.$$

Dravid et al. provide further experimentally obtained quantitative tests to substantiate qualitative arguments of the kind presented above to explain their computed profiles.

We conclude our example of dispersion in a spiralling circular pipe with remarks on the developed mean C_d. This quantity satisfies the equation

$$\begin{aligned} &\frac{\partial C_d}{\partial t} - \left(\gamma + \frac{1}{48}\epsilon\left(\frac{\mu}{8}\right)^2\right)\frac{\partial^2 C_d}{\partial x^2} + \left(\frac{\mu}{8} + \epsilon\frac{1}{96}\mu\alpha\right)\frac{\partial C_d}{\partial x} + 2\alpha C_d \\ &\quad = -\epsilon\left[a^2\frac{\partial^2 C_0}{\partial x^2} - \frac{a^2\mu}{8}\frac{\partial C_0}{\partial x} - \frac{\alpha^2}{2}C_0\right] + O(\epsilon^2) \\ &\quad \equiv \epsilon G(x, t) + O(\epsilon^2). \end{aligned} \tag{5.27}$$

Apart from the constants that are adjusted by $O(\epsilon)$, we see that C_d satisfies an inhomogeneous equation with a differential operator of the same form as that governing C_0. Thus, we can use the solution (5.7) for C_0 for delta function initial data with the new coefficients as a fundamental solution for the above equation. Make the change of variables

$$C_d = he^{-2\alpha t}, \qquad x = \xi + U_d t, \tag{5.28}$$

so that h satisfies

$$\frac{\partial h}{\partial t} - \gamma_d \frac{\partial^2 h}{\partial \xi^2} = \epsilon G(\xi + U_d t, t) + O(\epsilon^2). \tag{5.29}$$

Here, γ_d and U_d are the adjusted γ, U of the C_0 equation. With h given by

(5.30)

$$h(\xi,t) = \epsilon \left[\int_0^t \int_0^\infty \frac{e^{-2\alpha t'}}{\sqrt{4\pi\gamma_d t'}} e^{-\frac{(\xi' - U_d t' - x_0)^2}{4\gamma_d t'}} G(\xi', t')\, d\xi' dt' \right] + h(\xi, 0) + O(\epsilon^2),$$

we have

$$C_d(x,t) = e^{-2\alpha t} h(x - U_d t, t).$$

This representation for $C_d(x,t)$, of course, requires $C_d(x,0)$ which, in turn, requires $h(\xi,0)$ in order that h be determined from (5.30). The correct choice of $C_d(x,0)$ is

(5.31)

$$\begin{aligned} C_d(x,0) \;=\; & \overline{f}(x) + \frac{\mu}{4} \int_0^\infty \int_0^{2\pi} \int_0^1 (1 - r^2) \Big[\frac{\partial W_0}{\partial x}(x,r,\theta,\tau) - a \sin(x-\theta) \frac{\partial W_0}{\partial r}(x,r,\theta,\tau) \\ & + \frac{1}{r} a \cos(x-\theta) \frac{\partial W_0}{\partial \theta}(x,r,\theta,\tau) \Big] r\, dr\, d\theta\, d\tau \\ & - \int_0^\infty \int_0^{2\pi} a \sin(x-\theta) \frac{\partial W_0}{\partial x}(x,1,\theta,\tau)\, d\theta\, d\tau. \end{aligned}$$

6. Discussion and related problems. In this paper we have directed our attention to extending the theory and giving a method for analyzing diffusion of a scalar field $C(\overline{x}, t)$ in pipe flows with slowly varying cross-sections for which the convecting flow is different from, but near, Poiseuille flow. Throughout, we have used most frequently the terminology appropriate to dispersion of solute concentration for the scalar field C. In such applications the value $\alpha = 0$ is certainly most likely in the boundary condition $(dC/dn) + \alpha C = 0$. A nonzero value of α arises, however, if solute is catalyzed at the pipe walls. Moreover, when applications of interest include transfer of heat through conducting pipe boundaries, the influence of the wall conductance α becomes important. In this case, in order to investigate dependence of the effective diffusion coefficient D_e on α, the transformation of the governing partial differential equation (4.3) into the pure diffusion equation is involved and this merits a separate study. In these transformations the introduction of a new time scale permits an expression of $C_d(x,t)$ in terms of the mean of the initial concentration in an integral form similar to (5.30), containing a delay-type kernel in the integrand (cf., [9] for the uniform case). The solution in this form provides a potentially alternative attack on problems of dispersion in oscillatory flows where the convecting flow results from a harmonic longitudinal pressure gradient $\partial p/\partial x = P_0 \cos \omega t$ along the pipe (see, e.g., [6], [20], [21], and [28]). Methods similar to those employed in our work here make consideration of such problems possible for pipes of variable cross-sections. Moreover, in at least one instance [20], a delay-diffusion type of partial differential equation has been employed in the modelling of the dispersion process in the uniform situation.

Analytical and experimental results in the case of flow in a pipe of uniform (circular) cross-section verify that the later stages of dispersion of the diffusing substance

will be governed by an augmented apparent diffusion coefficient. At earlier times, however, the dispersion process shows sensitivity to time of release of solute during a cycle so that contraction of the distribution after flow reversals may occur. This implies the apparent longitudinal coefficient is negative. A relevant example is given in [13] of a steady, three-dimensional flow where the $O(\epsilon)$ correction to the underlying flow exhibits regions of reversal. Taken together with the remarks noted previously concerning the sign of the term (4.12) in D_e, this would suggest special attention to the possibility of contraction of the distribution for variable cross-section pipes: there is evidence [28] that effects of steady and oscillatory flow can be additive.

The noticeable effects on dispersion—in particular, on the rate at which the diffusing substance is spreading transversally—underscores the importance of the inclusion of transverse convective terms from the very beginning of the modelling process. Previous work has usually focussed on averaged properties of dispersion (over a period of an imposed oscillation, for example) so that several potentially important oscillatory effects are thereby excluded. The approach taken in this work avoids the early use of averaged quantities.

Finally, we mention briefly two other quite different kinds of flows in which three-dimensional effects seem to be important for dispersion enhancement. These are the following: (i) impulsively started flows and (ii) moving boundary flows. In (i), start-up effects on the dispersion can be quite prolonged and very significantly reduce the transverse spread of the distribution over that observed in fully developed flow. In (ii), it has been shown [25], in the nonviscous case at least, that for laterally moving pipe boundaries dispersion enhancement is possible, but the methods employed in that work are restricted to two dimensions.

Acknowledgments. The author acknowledges the valuable comments of the referees. The author also wishes to thank Professor A. J. Roberts for bringing to his attention references that employ center manifold theory for the study of models based on the slowly varying approximation.

REFERENCES

[1] R. ARIS, *On the dispersion of solute in a fluid flowing through a tube*, Proc. Roy. Soc. A, 235 (1956), pp. 67–77.

[2] S. A. BERGER, L. TALBOT, AND L.-S. YAO, *Flow in curved pipes*, Ann. Rev. Fluid Mech., 15 (1983), pp. 461–512.

[3] G. W. BLUMAN, *On the transformation of diffusion processes into the Weiner process*, SIAM J. Appl. Math., 39 (1980), pp. 238–247.

[4] G. F. CARRIER, *On the diffusive convection in tubes*, Quart. Appl. Math., 14 (1956), pp. 108–112.

[5] P. C. CHATWIN, *The approach to normality of the concentration distribution of solute in in a solvent flowing along a straight pipe*, J. Fluid Mech., 43 (1970), pp. 321–352.

[6] ———, *On the longitudinal dispersion of passive contaminant in oscillatory flows in tubes*, J. Fluid Mech., 71 (1975), pp. 513–527.

[7] ———, *The initial development of diffusion in Poiseuille flow*, J. Fluid Mech., 80 (1977), pp. 33–48.

[8] A. N. DRAVID, K. A. SMITH, E. W. MERRILL, AND P. L. BRIAN, *Effect of secondary fluid motion on laminar flow heat transfer in helically coiled tubes*, AIChE J., 17 (1971), pp. 1114–1122.

[9] P. C. FIFE AND K. R. NICHOLES, *Dispersion in flow through small tubes*, Proc. Roy. Soc. A, London, 344 (1975), pp. 131–145.

[10] W. N. GILL AND R. SANKARASUBRAMANIAN, *Exact analysis of unsteady convection*, Proc. Roy. Soc. A, London, 316 (1970), pp. 341–350.

[11] W. N. GILL AND R. SANKARASUBRAMANIAN, *Unsteady convective diffusion with interphase mass transfer*, Proc. Roy. Soc. A, 333 (1973), pp. 115–132.
[12] W. P. KOTORYNSKI, *Slowly varying channel flows in three dimensions*, J. Inst. Math. Appl., 24 (1979), pp. 71–80.
[13] ———, *Steady laminar flow in a twisted pipe of elliptical cross-section*, Comput. Fluids, 14 (1986), pp. 433–444.
[14] M. J. LIGHTHILL, *Initial development of diffusion in Poiseuille flow*, J. Inst. Math. Appl., 2 (1966), pp. 97–108.
[15] E. M. LUNGU AND H. K.MOFFATT, *The effect of wall conductance on heat diffusion in duct flow*, J. Engrg. Math., 16 (1982), pp. 121–135.
[16] G. N. MERCER AND A. J. ROBERTS, *A centre manifold description of contaminant dispersion in channels with varying flow properties*, SIAM J. Appl. Math., 50 (1990), pp. 1547–1565.
[17] ———, *A complete model of shear dispersion in pipes*, preprint.
[18] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[19] A. J. ROBERTS, *Boundary conditions for approximate differential equations*, J. Austral. Math. Soc. Ser. B, 34 (1992), pp. 54–80.
[20] R. SMITH, *A delay–diffusion description for contaminant dispersion*, J. Fluid Mech., 105 (1981), pp. 469–486.
[21] ———, *The contraction of contaminant distributions in reversing flows*, J. Fluid Mech., 129 (1983), pp. 137–151.
[22] ———, *Diffusion in shear flows made easy: the Taylor limit*, J. Fluid Mech., 175 (1987), pp. 201–214.
[23] I. H. SNEDDON, *The Use of Integral Transforms*, McGraw–Hill, New York, 1972.
[24] G. I. TAYLOR, *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. A, London, 219 (1953), pp. 186–203.
[25] S. TSANGARIS, *Longitudinal dispersion in a duct with moving walls*, Z. Angew. Math. Phys., 37 (1986), pp. 895–909.
[26] E. VARLEY AND B. SEYMOUR, *A method for obtaining exact solutions to partial differential equations with variable coefficients*, Stud. Appl. Math., 88 (1988), pp. 183–225.
[27] M. D. VAN DYKE, *Slow variations in continuum mechanics*, Adv. Appl. Mech., 25 (1987), pp. 1–45.
[28] E. J. WATSON, *Diffusion in oscillatory pipe flow*, J. Fluid Mech., 133 (1983), pp. 233–244.
[29] K. B. WOLF, *Integral Transforms in Science and Engineering*, Plenum Press, New York, 1979.
[30] W. R. YOUNG AND S. JONES, *Shear Dispersion*, Phys. Fluids A, 3 (1991), pp. 1087–1101.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 941–947, May 1994

007

ON A LOCAL EXISTENCE THEOREM FOR A SIMPLIFIED ONE-DIMENSIONAL HYDRODYNAMIC MODEL FOR SEMICONDUCTOR DEVICES*

BO ZHANG†

Abstract. A simplified hydrodynamic model for semiconductor devices, where the energy equation is replaced by a pressure-density relationship, is studied. The system of Euler–Poisson equations is changed to a quasilinear wave equation in Lagrangian mass coordinates. The local existence of a smooth solution of the Euler–Poisson equations is then obtained by using a known result for the quasilinear wave equation.

Key words. hydrodynamic model, Euler–Poisson equations, Lagrangian mass coordinates, quasilinear wave equation

AMS subject classifications. 35L70, 35Q60, 76X05

1. Introduction and main result.

The hydrodynamic model for semiconductor devices is a generalization of the standard drift-diffusion model. For steady-state subsonic electron flow in the hydrodynamic model, Gardner Jerome, and Rose [6] proved the existence of solutions and convergence of Newton's method. In this paper we investigate a simplified hydrodynamic model in which the pressure is a given function of the particle density only. This assumption is commonly used in gas dynamics for isentropic or isothermal flows [2].

After appropriate scaling, the one-dimensional time-dependent system in the case of one carrier type (e.g., electrons) reads [7]

$$\rho_t + (\rho u)_x = 0, \tag{1.1}$$

$$u_t + uu_x + \frac{1}{\rho}(p(\rho))_x = \phi_x - \frac{u}{\tau}, \tag{1.2}$$

$$\phi_{xx} = \rho - D, \tag{1.3}$$

where $\rho(x,t)$, $u(x,t)$, and $\phi(x,t)$ denote the electron density, velocity, and electrostatic potential, respectively. The pressure function, $p = p(\rho)$, has the property that $\rho^2 p'(\rho)$ is strictly increasing from $[0,\infty)$ onto $[0,\infty)$. A commonly used hypothesis [2] is

$$p(\rho) = k\rho^\gamma, \quad \gamma \geq 1, \quad k > 0.$$

The positive function $\tau = \tau(\rho, u)$ is the momentum relaxation time. The device domain is the x-interval $I \equiv (0,1)$. The given function $D = D(x)$ is the doping profile.

The system (1.1)–(1.3) is a set of one-dimensional Euler–Poisson equations with an electric field term and a momentum relaxation time. For this simplified hydrodynamic model, there has recently been some mathematical analysis. Markowich and Degond [7] proved the existence of a unique smooth solution in the stationary subsonic one-dimensional hydrodynamic model that is characterized by an assumption on the current flow through the device that it is sufficiently small. Gamba [4] studied

* Received by the editors January 14, 1992; accepted for publication (in revised form) January 25, 1993.

† Center of Applied Mathematics, Department of Mathematics, Purdue University, West Lafayette, Indiana 47907. Present address, Department of Mathematics, Northwestern University, Evanston, Illinois 60208-2730.

stationary transonic solutions for this simplified model. However, all of the above results are for the steady-state hydrodynamic model. In this paper we study the time-dependent system (1.1)–(1.3). The equations (1.1)–(1.3) are prescribed by the following initial-boundary value conditions:

$$(\rho, u)|_{t=0} = (\rho_0(x), u_0(x)), \quad 0 \le x \le 1, \tag{1.4}$$

$$u(0,t) = 0, \quad u(1,t) = 0, \quad t \ge 0, \tag{1.5}$$

$$\phi(0,t) = \phi_1(t), \quad \phi(1,t) = \phi_2(t), \quad t \ge 0, \tag{1.6}$$

or more general boundary value conditions:

$$\begin{cases} u(0,t) = u_1(t), \quad u(1,t) = u_2(t), \quad t \ge 0, \\ \phi(0,t) = \phi_1(t), \quad \phi(1,t) = \phi_2(t), \quad t \ge 0, \end{cases} \tag{1.7}$$

where ϕ_1 and ϕ_2 are the applied bias. In the next section we will see that there is a difference between (1.5) and (1.7).

Due to the formation of shocks, the initial-boundary value problem (1.1)–(1.6) does not generally have a global smooth solution, no matter how smooth the initial-boundary data and given functions are. For example, a steady-state electron shock wave in a submicrometer semiconductor device was first simulated by Gardner [5]. At best, we should aim at establishing the existence of a local smooth solution on a maximal time interval. For the existence of a physical global weak solution of (1.1)–(1.6), see [11].

As usual, $C^m([0,T];X)$ stands for the space of m-times (strongly) continuously differentiable functions from $[0,T]$ to a Banach space X. We take X as the Sobolev space $H^m \equiv W^{m,2}(I)$ for nonnegative integers m, and set $H_0^1 \equiv W_0^{1,2}(I)$. In order to obtain a solution with the desired regularity we must assume that (1.4)–(1.6) and the given functions satisfy certain natural compatibility conditions of higher order. We need the following assumptions.

(A_1) Regularity of given functions: $\tau(\rho,u) \in C^3(\mathbb{R}_+ \times \mathbb{R})$, $0 < \tau_0 \le \tau(\rho,u) \le M$, and $D(x) \in C^3[0,1]$.

(A_2) Regularity of initial-boundary value conditions: $\phi_i(t) \in C^3[0,T]$, $i = 1,2$, for a constant $T > 0$; $\rho_0 \ge m > 0$, $v_0 \equiv \frac{1}{\rho_0}$, $u_0 \in H^3$.

(A_3) Compatibility conditions: Let

$$w_0(x) = \int_0^x v_0 dx - x \int_0^1 v_0 dx,$$

$$w_1(x) = u_0(x),$$

$$w_2(x) = \frac{k\gamma}{v_0^{\gamma+1}} v_0' - \frac{w_1}{\tau(v_0^{-1}, w_1)} + b,$$

$$w_3(x) = \frac{k\gamma}{v_0^{\gamma+1}} u_0'' - \frac{k\gamma(\gamma+1)}{v_0^{\gamma+2}} u_0' v_0' - \frac{w_2}{\tau} + \frac{w_1}{\tau^2}\left(w_2\tau_2' - \frac{u_0'}{v_0^2}\tau_1'\right) - \int_0^1 D u_0' dx + c,$$

where

$$b = \frac{\phi_2(0) - \phi_1(0) - \int_0^1 (\int_0^x v_0 dx)(1 - Dv_0)dx}{\int_0^1 v_0 dx},$$

$$c = \frac{\phi_2'(0) - \phi_1'(0) - \int_0^1 u_0(1 - Dv_0)dx - \int_0^1 u_0'(\int_0^x u_0 dx)dx}{\int_0^1 v_0 dx},$$

and assume that $w_i \in H^{4-i} \cap H_0^1$, $i = 1, 2, 3$.

The main result of this paper is as follows.

THEOREM 1.1. *Assume that* (A_1), (A_2), *and* (A_3) *hold. Then, for sufficiently small* $T \in (0, \infty)$, *the initial-boundary value problem* (1.1)–(1.6) *has a solution* (ρ, u, ϕ) *such that* $\rho > 0$, $\rho \in \cap_{m=0}^3 C^m([0,T]; H^{3-m})$, $u \in \cap_{m=0}^3 C^m([0,T]; H^{4-m})$, *and* $\phi \in \cap_{m=0}^3 C^m([0,T]; H^{5-m})$.

We will use Lagrangian mass coordinates to reformulate (1.1)–(1.3) into a quasilinear wave equation that is equivalent to the original system. By a result of Dafermos and Hrusa [3], we obtain the local existence and uniqueness of a smooth solution for the quasilinear wave equation.

2. Proof of the local existence theorem. The solution of the Poisson equation (1.3) for boundary data (1.6) is given uniquely by the following equation

$$\phi = \int_0^1 G(x,z)(\rho - D)dz + \phi_1 + x(\phi_2 - \phi_1), \tag{2.1}$$

where $G(x,z)$ is the Green's function defined by

$$G(x,z) = \begin{cases} x(z-1), & x < z, \\ z(x-1), & x > z. \end{cases}$$

Substituting the derivative of (2.1) for ϕ_x in (1.2), we have

$$\rho_t + (\rho u)_x = 0, \tag{2.2}$$

$$u_t + uu_x + \frac{1}{\rho}(p(\rho))_x = \int_0^1 G_x(x,z)(\rho - D)dz - \frac{u}{\tau} + \phi_2 - \phi_1. \tag{2.3}$$

Let (ρ, u) be a solution of (2.2)–(2.3) such that $\rho \in \cap_{m=0}^3 C^m([0,T]; H^{3-m})$ and $u \in \cap_{m=1}^3 C^m([0,T]; H^{4-m})$. The relation between the Euler coordinates (x,t) and Lagrangian mass coordinates (y,t) is given by

$$y = \int_{x(t)}^x \rho(z,t)dz, \tag{2.4}$$

where $x(t)$ is a well-defined particle path satisfying the following ordinary differential equation:

$$x'(t) = u(x(t), t),$$

$$x(0) = 0.$$

This transformation $(x,t) \mapsto (y(x,t),t)$ satisfies the following equations:

$$y_x = \rho(x,t),$$

$$y_t = -\rho(x,t)u(x,t).$$

It follows from (2.2) that the transformation $(x,t) \mapsto (y(x,t),t)$ is consistent.

By this transformation, the system (1.1)–(1.3) can be conveniently reformulated as

$$v_t - u_y = 0, \tag{2.5}$$

$$u_t + p\left(\frac{1}{v}\right)_y = \frac{\phi_y}{v} - \frac{u}{\tau}, \tag{2.6}$$

$$\left(\frac{\phi_y}{v}\right)_y = 1 - Dv, \tag{2.7}$$

where $v = \frac{1}{\rho}$. The initial-boundary value conditions (1.4)–(1.6) are translated into

$$(v,u)|_{t=0} = (v_0(y), u_0(y)), \qquad 0 \le y \le 1, \tag{2.8}$$

$$u(0,t) = 0, \quad u(1,t) = 0, \qquad t \ge 0, \tag{2.9}$$

$$\phi(0,t) = \phi_1(t), \quad \phi(1,t) = \phi_2(t), \qquad t \ge 0. \tag{2.10}$$

Here, we assume that $\int_0^1 \rho_0 dx = 1$.

The problem (1.1)–(1.3) with boundary data (1.7) can be reformulated as the following free boundary problem:

$$\begin{cases} (2.5)-(2.7), \quad y_1(t) < y < y_2(t), \quad t > 0, \\ y_1'(t) = -\dfrac{u_1(t)}{v(y_1(t),t)}, \quad y_1(0) = 0, \\ y_2'(t) = -\dfrac{u_2(t)}{v(y_2(t),t)}, \quad y_2(0) = 1, \\ (v,u)|_{t=0} = (v_0(y), u_0(y)), \quad 0 \le y \le 1, \\ u(y_1(t),t) = u_1(t), \quad u(y_2(t),t) = u_2(t), \quad t \ge 0, \\ \phi(y_1(t),t) = \phi_1(t), \quad \phi(y_2(t),t) = \phi_2(t), \quad t \ge 0, \\ \displaystyle\int_0^1 v_0(y)dy + \int_0^t (u_2(t) - u_1(t))dt \ge m_0 > 0. \end{cases} \tag{2.11}$$

For concreteness, we are concerned with the initial-boundary value problem (2.5)–(2.10). Solving for $\frac{\phi_y}{v}$ from (2.7) and (2.10), we have

$$\frac{\phi_y}{v} = \int_0^y (1 - Dv)dy + f, \tag{2.12}$$

where

$$f = \left(\phi_2 - \phi_1 - \int_0^1 v\left(\int_0^y (1 - Dv)dy'\right)dy\right)\left(\int_0^1 v_0\,dy\right)^{-1}.$$

Then, (2.5)–(2.7) reduces to

$$v_t - u_y = 0, \tag{2.13}$$

$$u_t + p\left(\frac{1}{v}\right)_y = \int_0^y (1 - Dv)dy - \frac{u}{\tau} + f. \tag{2.14}$$

Since the region $I_T \equiv I \times (0,T)$ is simply connected, (2.13) implies the existence of a function w such that

$$v = w_y, \qquad u = w_t. \tag{2.15}$$

Equation (2.14) then becomes

$$w_{tt} - \frac{k\gamma}{w_y^{\gamma+1}} w_{yy} = \widetilde{f} - \frac{w_t}{\tau} \tag{2.16}$$

where

$$\widetilde{f} = \int_0^y (1 - Dw_y)dy + \left(\phi_2 - \phi_1 + \int_0^1 w_y\left(\int_0^y (1 - Dw_y)dy'\right)dy\right)\left(\int_0^1 v_0\,dy\right)^{-1}.$$

The initial-boundary value conditions (2.8) and (2.9) become

$$\begin{cases} w|_{t=0} = \displaystyle\int_0^y v_0(y)dy + C_1, \quad w_t|_{t=0} = u_0(y), \quad 0 \le y \le 1, \\ w(0,t) = C_2, \quad w(1,t) = C_3, \quad t > 0, \end{cases} \tag{2.17}$$

where $\{C_i : i = 1,2,3\}$ is a set of constants such that $C_2 = C_1$ and $C_3 = \int_0^1 v_0(y)dy + C_1$. Without loss of generality, we assume $C_1 = 0$.

LEMMA 2.1. *Assume that* $w \in \cap_{m=0}^4 C^m([0,T]; H^{4-m})$ *and* $w_y(0,t) = \frac{1}{\rho(0,t)} > 0$ *for* $0 \le t \le T$. *Then,* $w_y(y,t)$ *is positive and there exists a constant* $C > 0$ *such that*

$$\int_0^1 w_t^2\,dy \le C,$$

$$\int_0^1 \frac{1}{w_y^{\gamma-1}} dy \le C \quad \text{if } \gamma > 1, \text{ or}$$

$$\int_0^1 \ln w_y\,dy \le C \quad \text{if } \gamma = 1.$$

Proof. Let $\alpha(y) = w_y$. Then, (2.16) can be written as an ordinary differential equation

$$\frac{1}{\alpha(y)}\frac{d}{dy}\alpha(y) = \frac{1}{k\gamma}\alpha^\gamma(y)\left(w_{tt} - \widetilde{f} + \frac{w_t}{\tau}\right).$$

It follows that

$$\alpha(y) = \frac{1}{\rho(0,t)} \exp\left\{ \frac{1}{k\gamma} \int_0^y \alpha^\gamma \left(w_{tt} - \widetilde{f} + \frac{w_t}{\tau} \right) dy \right\}.$$

Then

$$w_y(y,t) \geq \frac{1}{\sup\limits_{0\leq t\leq T} \rho(0,t)} \exp\left\{ \frac{-1}{k\gamma} \sup_{(y,t)\in \overline{I}_T} v^\gamma \left(|u_t| + |\widetilde{f}| + \frac{|u|}{\tau_0} \right) \right\} > 0.$$

From (2.5) and (2.9), we have

$$\int_0^1 w_y \, dy = \int_0^1 v dy = \int_0^1 v_0 dy \equiv \overline{v}_0.$$

Then

$$|\widetilde{f}| \leq 2 \left(1 + \sup_{0\leq y\leq 1} |D|\overline{v}_0 \right) + \sup_{0\leq t\leq T} (|\phi_1(t)| + |\phi_2(t)|)\overline{v}_0 \equiv \widetilde{M}.$$

Multipling (2.16) by w_t and integrating with respect to y over $(0,1)$, we obtain

$$\frac{d}{dt} \left(\frac{1}{2} \int_0^1 w_t^2 \, dy + \frac{k}{\gamma - 1} \int_0^1 \frac{1}{w_y^{\gamma-1}} dy \right) + \int_0^1 \frac{w_t^2}{\tau} dy = \int_0^1 \widetilde{f} w_t \, dy \quad \text{if } \gamma > 1,$$

or

$$\frac{d}{dt} \left(\frac{1}{2} \int_0^1 w_t^2 \, dy + k \int_0^1 \ln w_y \, dy \right) + \int_0^1 \frac{w_t^2}{\tau} dy = \int_0^1 \widetilde{f} w_t \, dy \quad \text{if } \gamma = 1.$$

Then,

$$\frac{1}{2} \int_0^1 w_t^2 \, dy + \frac{k}{\gamma - 1} \int_0^1 \frac{1}{w_y^{\gamma-1}} dy + \frac{1}{2} \int_0^t \int_0^1 \frac{w_t^2}{\tau} dy \, dt \leq \widetilde{M} \int_0^t \int_0^1 \tau \, dy \, dt \quad \text{if } \gamma > 1,$$

or

$$\frac{1}{2} \int_0^1 w_t^2 \, dy + k \int_0^1 \ln w_y \, dy + \frac{1}{2} \int_0^t \int_0^1 \frac{w_t^2}{\tau} dy \, dt \leq \widetilde{M} \int_0^t \int_0^1 \tau \, dy \, dt \quad \text{if } \gamma = 1.$$

It follows that

$$\int_0^1 w_t^2 \, dy \leq C,$$

$$\int_0^1 \frac{1}{w_y^{\gamma-1}} dy \leq C \quad \text{if } \gamma > 1,$$

$$\int_0^1 \ln w_y \, dy \leq C \quad \text{if } \gamma = 1. \qquad \square$$

From Lemma 2.1, (2.16) is a quasilinear wave equation where the speed of propagation, $\sqrt{-p'}$, depends on $\rho = \frac{1}{w_y}$. In the following we change the nonhomogeneous boundary conditions to homogeneous boundary conditions. Then let

$$\overline{w} = w - y\overline{v}_0.$$

Then, the initial-boundary value problem (2.16)–(2.17) becomes

$$
(2.18)\qquad \begin{cases} \overline{w}_{tt} - \dfrac{k\gamma}{(\overline{w}_y + \overline{v}_0)^{\gamma+1}}\overline{w}_{yy} = \widetilde{f} - \dfrac{w_t}{\tau}, & (y,t) \in I_T, \\ \overline{w}|_{t=0} = \displaystyle\int_0^y v_0 dy - \overline{v}_0 y, & 0 \le y \le 1, \\ \overline{w}_t|_{t=0} = u_0(y), & 0 \le y \le 1, \\ \overline{w}(0,t) = \overline{w}(1,t) = 0, & 0 \le t \le T. \end{cases}
$$

LEMMA 2.2. *Assume that* (A_1), (A_2), *and* (A_3) *hold. Then, for sufficiently small* $T \in (0,\infty)$, *the initial-boundary value problem* (2.18) *has a unique solution*

$$
(2.19)\qquad \overline{w} \in \bigcap_{m=0}^{4} C^m([0,T]; H^{4-m}).
$$

Proof. By the assumptions (A_1), (A_2), and (A_3), and Lemma 2.1, all conditions of Theorem 5.1 of [3] are satisfied, so that the result follows. For details, see [3]. □

Proof of Theorem 1.1. By Lemma 2.2, the unique solution $\overline{w}$ of (2.18) satisfies (2.19). It follows from the Sobolev embedding theorem [1] that $\overline{w} \in C^2(\overline{I}_T)$. From (2.15), we have that $\rho = \frac{1}{\overline{w}_y + \overline{v}_0}$, $u = \overline{w}_t$, and ϕ is given by (2.1). The existence of a smooth solution of the initial-boundary value problem (1.1)–(1.6) is equivalent to that of (2.18). Therefore, the result is true. □

Acknowledgment. The author is grateful to Professors Jim Douglas, Jr. and Peter Markowich for their guidance, encouragement, and support.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] R. COURANT, AND K. O. FRIEDRICHS, *Supersonic Flow and Shock-Waves*, Interscience, J. Wiley and Sons, New York, 1967.
[3] C. M. DAFERMOS AND W. J. HRUSA, *Energy methods for quasilinear hyperbolic initial-boundary value problems: application to elastodynamics*, Arch. Rational Mech. Anal., 87 (1985), pp. 267–292.
[4] I. M. GAMBA, *Stationary transonic solutions for a one-dimensional hydrodynamic model for semiconductors*, Comm. Partial Differential Equations, 17 (1992), pp. 553–577.
[5] C. L. GARDNER, *Numerical simulation of a steady-state electron shock wave in a submicron semiconductor device*, IEEE Trans. Electron Devices, 38 (1991), pp. 392–398.
[6] C. L. GARDNER, J. W. JEROME AND D. J. ROSE, *Numerical methods for the hydrodynamic device model: subsonic flow*, IEEE Trans. Computer-Aided Design, 8 (1989), pp. 501–507.
[7] P. A. MARKOWICH AND P. DEGOND, *On a one-dimensional steady-state hydrodynamic model for semiconductors*, Appl. Math. Letters, 3 (1990), pp. 25–29.
[8] J. SMOLLER, *Shock Waves and Reaction–Diffusion Equations*, Springer-Verlag, New York, 1983.
[9] E. THOMANN AND F. ODEH, *Remarks on the well-posedness of the hydrodynamic model for semiconductors devices*, Proceedings of the Sixth International Nasecode Conference, J. H. Miller, ed., Boole Press Ltd., 1989, pp. 22–27.
[10] D. H. WAGNER, *The transformation from Eulerian to Lagrangian coordinates for solutions with discontinuities*, Nonlinear Hyperbolic Problems, C. Carasso, P.-A. Raviart, and D. Serre, eds., Springer-Verlag, New York, 1986.
[11] B. ZHANG, *Convergence of the Godunov scheme for a simplified one-dimensional hydrodynamic model for semiconductor devices*, Comm. Math. Phys., 157 (1993), pp. 1–22.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 948-961, May 1994

008

STABILITY OF SOME TEST EQUATIONS WITH DELAY*

V. B. KOLMANOVSKII†, L. TORELLI‡, AND R. VERMIGLIO§

Abstract. The authors propose some techniques to obtain stability conditions for certain differential equations with delay. These techniques are applied to three concrete test situations. In the first and second cases, they consider equations without instantaneous dissipative terms. This is important because in real situations, such as control theory, finite time is necessary to measure characteristics of the object and to treat the results of the measurements in order to create control action. In the last section, general results are presented for the chemostat model. These test equations are interesting examples for a forthcoming investigation concerning stability of numerical methods.

Key words. stability, test equation, delay differential equation, Liapunov method, chemostat model

AMS subject classifications. 34D05, 34K20

1. Introduction.

Many real phenomena can be modeled by equations with delay, and the general theory of such equations has been considerably developed in the last years. Several authors have investigated the analytical aspect, and the references on this subject are very ample. The reader is referred to the following books for the basic topics: Burton [7], Bellman and Cooke [4], Hale [18], Kolmanovskii and Nosov [22], Hino, Murakami, and Naito [20], and Gripenberg, Londen, and Staffans [13]. Recently the interest in and necessity for numerical schemes have been increasing, since the possibility of obtaining the analytic solution is very poor even for the simplest cases (see Barwell [1], Bellen [2], [3], Hairer, Norsett, and Wanner [17], Meinardus and Nurnberger [23], Torelli and Vermiglio [24], Zennaro [27], and the references therein).

One of the most important problems in the study of dynamical models and their applications is that of describing the nature of the solutions for large positive values of the independent variable. From a numerical point of view, the stability properties of the approximation schemes must also be studied. The usual approach to fulfill such requirements is to have a set of test equations which are as general as possible and for which explicit analytic stability conditions can be given, and to characterize the numerical methods which, when applied to the test equation, show the same asymptotic behavior under the same conditions.

In this part of our investigation we propose more general test equations and we give some sufficient explicit conditions for the asymptotical stability of the solutions. Namely, we consider scalar linear equations with variable coefficients and arbitrary (distributed and discrete) delays and a two-dimensional system of nonlinear equations of chemostat. In general, in the theory of stability for delay equations two approaches are widely used, i.e., the Laplace transform together with the study of the corresponding characteristic equation and the generalizations of Liapunov's direct method [4], [22]. Here both of these approaches are used. In the forthcoming second part we shall present numerical procedures whose stability properties will be checked, according to

*Received by the editors September 28, 1992; accepted for publication (in revised form) March 24, 1993. This research was supported by Ministero dell' Università della Ricerca Scientifica e Tecnologica.

†Department of Cybernetics, Moscow Institute of Electronics and Mathematics, 109028 Moscow, Russia. The work of this author while at the University of Trieste, Trieste, Italy, was supported by Consiglio Nazionale delle Ricerche.

‡Dipartimento di Scienze Matematiche, Università di Trieste, I-34100 Trieste, Italy.

§Dipartimento di Matematica ed Informatica, Università di Udine, I-33100 Udine, Italy.

the philosophy described above, by means of test equations studied in this paper.

Let us now quote some results and definitions from general stability theory of functional differential equations that are used in this paper. For all the details the reader is referred to some of the books mentioned above.

Consider the initial-value problem (IVP) for the general delay differential equation (DDE)

$$x'(t) = f(t, x_t), \qquad t \geq t_0 \geq 0, \tag{1.1}$$

$$x_{t_0} = \varphi(\theta), \qquad \theta \leq 0, \tag{1.2}$$

where $x(t) \in \mathbf{R}^n$ and $x_t := x(t+\theta)$, $\theta \leq 0$, $f : [0, +\infty) \times M \to \mathbf{R}^n$, M denotes a complete metric space of continuous functions mapping the interval $(-\infty, 0]$ into $\mathbf{R}^n$ with a metric ρ and the initial function $\varphi \in M$. We recall that if $-\tau \leq \theta \leq 0$, equation (1.1) is a DDE with *bounded delay* and $M = C[-\tau, 0]$ with $\|\varphi\| = \max_{-t \leq \theta \leq 0} |\varphi(\theta)|$.

If $-\infty < \theta \leq 0$, equation (1.1) is a DDE with *unbounded delay*. In this case we have several possibilities for the choice of the space M and metric ρ, which could be taken properly for the problem (see [20], [12], and [22]). For instance, we can take the space of continuous and bounded functions on $(-\infty, 0]$ with the norm $\|\varphi\| = \sup_{\theta \leq 0} |\varphi(\theta)|$ (see Krisztin [21]).

Let us suppose f and φ are such that there exists a unique solution to (1.1) and (1.2) and denote such a solution by $x(t, t_0, \varphi)$. Without loss of generality we can assume that $f(t, 0) \equiv 0$ for $t \geq 0$, and thus we can investigate the stability of the trivial solution.

DEFINITION 1.1. The trivial solution of (1.1) is called *stable* if for each $\epsilon > 0$ there exists $\delta = \delta(\epsilon, t_0)$ such that the solution $x(t, t_0, \varphi)$ of the problem (1.1) and (1.2) satisfies inequality $|x(t, t_0, \varphi)| \leq \epsilon$ for $t \geq t_0$ if $\varphi \in B(0, \delta) = \{\varphi | \rho(0, \varphi) \leq \delta\}$.

DEFINITION 1.2. If the solution $x(t, t_0, \varphi)$ of (1.1) and (1.2) is stable and in addition for each t_0 there exists $\Delta = \Delta(t_0) > 0$ such that $x(t, t_0, \varphi)$ vanishes when $t \to \infty$ for all initial functions $\varphi \in B(0, \Delta)$, then the trivial solution is called *asymptotically stable*. The set $\Omega(t_0) \subseteq M$ of all initial functions φ such that $x(t, t_0, \varphi)$ vanishes when $t \to \infty$ is the *domain of attraction* of the trivial solution at the initial moment t_0.

If $\Omega(t_0)$ in Definition 1.2 coincides with the space of continuous functions M, then the trivial solution is *globally asymptotically stable*.

If the same definitions hold independently of $t_0 \geq 0$, we speak of *uniform stability*, *uniform asymptotic stability*, and so on.

We recall the following lemma which we shall use in the next section.

LEMMA 1.3. *Let $\alpha(t)$ be any nonnegative, uniformly continuous function on $[t_0, \infty)$. If $\int_{t_0}^{\infty} \alpha(t)dt < \infty$, then $\alpha(t)$ vanishes as $t \to \infty$.*

2. Linear equations with variable delays. In this section we are able to derive sufficient conditions for stability of the zero solution for certain "pure" delay differential equations, i.e., equations where the derivative depends only on the past values of the solution. Our approach is based on the application of suitable degenerate Liapunov functionals. These conditions are easy to verify, and the comparison with the existing results are presented in what follows.

We consider the following real scalar DDE:

$$x'(t) = -b(t)x(t - \tau(t)), \qquad t \geq 0 \tag{2.1}$$

and its generalization

$$x'(t) = -\sum_{i=1}^{N} b_i(t)x(t-\tau_i(t)), \qquad t \geq 0, \tag{2.2}$$

which contains N delay functions. In (2.1) (respectively, (2.2)) we always assume that

(2.3a) M is the space of continuous and bounded functions on $[0,+\infty)$ with the norm $\|\varphi\| = \sup_{\theta\leq 0}|\varphi(\theta)|$;

(2.3b) b (respectively, b_i): $[0,+\infty) \to \mathbf{R}$ is continuous;

(2.3c) τ (respectively, τ_i): $[0,+\infty) \to [0,+\infty)$ is continuous, differentiable, and $\tau'(t)$ (respectively, $\tau_i'(t)) \leq \mathbf{R} < 1$ for $t \geq 0$.

Remark 2.1. The hypothesis (2.3b) ensures that $t-\tau(t)$ (respectively, $t-\tau_i(t)$) has a differentiable inverse function denoted by $g(t)$ (respectively, $g_i(t)$).

Observe that we allow the delay function to be unbounded. In the particular case of bounded delays, several authors have investigated the asymptotic behavior of the solutions of (2.1) and (2.2) as well as more general cases (see Yoneyama [25], [21], Yorke [26], Burton and Haddok [8], Cooke [9], [10], Gyori [14], and all the references therein). For (2.1) it is well known that if

(2.4a) $b : [0,+\infty) \to [0,+\infty)$ is continuous;

(2.4b) $b(t) \leq \beta$ for $t \geq 0$;

(2.4c) $\tau : [0,+\infty) \to [0,q]$ is continuous;

then the zero solution is uniformly stable if $\beta q \leq \frac{3}{2}$ and asymptotically stable if

$$\beta q < \frac{3}{2}. \tag{2.5}$$

Furthermore, the upper bound $\frac{3}{2}$ is the best possible for (2.1); in fact, if $\beta q > \frac{3}{2}$ there are equations with unbounded solutions. This result can be obtained by means of the fundamental work of Yorke [26] for general, scalar, nonlinear equations (1.1), where one of the main assumptions is the boundedness of the delay function. Yoneyama [25] removed (2.4b) and proved that if (2.4a) and (2.4c) hold and

$$\sup_{t\geq 0}\int_t^{t+q} |b(s)|ds \leq \frac{3}{2}, \tag{2.6}$$

then the zero solution of (2.1) is uniformly stable. More recently, Krisztin [21] gave a generalization of Yorke's theorem, which is flexible for more delays. For (2.2) we can easily get that if

$$b_i : [0,+\infty) \to [0,+\infty) \quad \text{is continuous for } i = 1,\dots,N;$$
$$b_i(t) \leq \beta_i \quad \text{for } t \geq 0, \quad i = 1,\dots,N;$$
$$\tau_i : [0,+\infty) \to [0,q_i] \quad \text{is continuous;}$$

then the zero solution is uniformly stable if $\sum_{i=1}^{N} \beta_i q_i \leq 1$ and uniformly asymptotically stable if $\sum_{i=1}^{N} \beta_i q_i < 1$ (for $N \geq 2$ the estimate 1 is the best possible; see [21]).

Other sharp results for (2.1) are given if $b(t)$ and/or $\tau(t)$ are small as $t \to +\infty$ (see [8]–[10], [14], and [15]).

It is important to point out that, concerning stability analysis, (2.1) and (2.2) cannot be considered as special cases of

$$x'(t) = -a(t)x(t) + \sum_{i=1}^{N} b_i(t)x(t - \tau_i(t)), \qquad t \geq 0, \tag{2.7}$$

where we assume (2.3) and $a : [0, +\infty) \to [0, +\infty)$ is continuous. The DDE (2.7) with $a(t) \not\equiv 0$, for $t \geq 0$, has been studied, for instance, in [7], [11], [18], [22], and [16].

In fact, by applying the Liapunov direct method (see [22]) to the following positive-definite functional,

$$V(t, x_t) := x^2(t) + \sum_{i=1}^{N} \int_t^{g_i(t)} |b_i(s)| x^2(s - \tau_i(s)) ds \quad \text{for } t \geq 0,$$

where $g_i(t)$, $i = 1, \ldots, N$, is defined in Remark 2.1, we can easily see that the trivial solution of (2.7) with initial condition (1.2) is uniformly asymptotically stable if $\sum_{i=1}^{N} b_i(t)$ is bounded and the following conditions hold:

$$\inf_{t \geq 0} \left[2a(t) - \sum_{i=1}^{N} (|b_i(t)| + |b_i(g_i(t))| g_i'(t)) \right] = A > 0, \tag{2.8a}$$

$$\sup_{t \geq 0} \sum_{i=1}^{N} \int_t^{g_i(t)} |b_i(s)| ds < +\infty. \tag{2.8b}$$

Conditions (2.8) show that in order to stabilize (2.7), we must have that the instantaneous feedback, defined by the positive function $a(t)$, dominates the delay effects. The same considerations arise by analyzing the results of the authors mentioned above. Thus for the analysis of the asymptotic behavior of pure delay differential equations (2.1) and (2.2), we suggest the use of suitable degenerate Liapunov functionals with a negative definite derivative.

THEOREM 2.2. *Consider the* DDE (2.1) *and assume* (2.3). *If*

$$\inf_{t \geq 0} b(t) = B > 0 \tag{2.9a}$$

and

$$\sup_{t \geq 0} \int_t^{g(t)} b(s) ds = \beta < 1, \tag{2.9b}$$

the trivial solution of (2.1) *with initial condition* (1.2) *is uniformly stable. Moreover, if* $b(t)$ *is bounded, then the trivial solution is uniformly asymptotically stable.*

Proof. Introduce the following nonnegative functional

$$V(t, x_t) := \left(x(t) - \int_t^{g(t)} b(s)x(s - \tau(s)) ds \right)^2 + \int_t^{g(t)} b(g(s))g'(s) \int_s^{g(t)} b(u)x^2(u - \tau(u)) du\, ds.$$

By computing its derivative V' along the trajectories, by (2.1), by the Cauchy inequality (i.e., $a^2+b^2 \geq 2ab$), and by (2.9b), which means that the average value of b must be less than 1, we get for $t \geq 0$,

$$\begin{aligned}
V' &= -2b(g(t))g'(t)x(t)\left(x(t) - \int_t^{g(t)} b(s)x(s-\tau(s))ds\right) \\
&\quad + b(g(t))g'(t)\left(-\int_t^{g(t)} b(s)x^2(s-\tau(s))ds + x^2(t)\int_t^{g(t)} b(g(s))g'(s)ds\right) \\
&\leq -x^2(t)b(g(t))g'(t)\left(2 - \int_t^{g(t)} b(s)ds - \int_{g(t)}^{g(g(t))} b(s)ds\right) \\
&\leq -2(1-\beta)b(g(t))g'(t)x^2(t) \\
&\leq -Cx^2(t),
\end{aligned}$$

where C is a suitable positive constant given by (2.9a) and $\inf_{t\geq 0} g'(t) > 0$. By integrating we have, for $t \geq 0$,

$$V(t,x_t) - V(0,x_0) = \int_0^t V'(s,x_s)ds \leq -C\int_0^t x^2(s)ds,$$

and, since $V(t,x_t)$ is a nonnegative functional, we get

$$C\int_0^t x^2(s)ds \leq V(0,x_0) \quad \text{for } t \geq 0.$$

Thus,

$$\int_0^\infty x^2(s)ds < \infty.$$

Moreover, by the same consideration, it follows that

$$2(1-\beta)\int_0^t b(g(s))g'(s)x^2(s)ds \leq V(0,x_0) \quad \text{for } t \geq 0,$$

and so

$$\int_0^\infty b(g(s))g'(s)x^2(s)ds < \infty.$$

Now by the definition of V, the Cauchy inequality, and (2.9b), we have

$$\begin{aligned}
V &\geq x^2(t) - 2x(t)\int_t^{g(t)} b(s)x(s-\tau(s))ds \\
&\quad + \int_t^{g(t)} b(g(s))g'(s)\int_s^{g(t)} b(u)x^2(u-\tau(u))du\,ds \\
&\geq x^2(t)\left(1 - \int_t^{g(t)} b(s)ds\right) - \int_t^{g(t)} b(s)x^2(s-\tau(s))\left(1 - \int_t^s b(g(u))g'(u)du\right)ds \\
&\geq x^2(t)(1-\beta) - \int_t^{g(t)} b(s)x^2(s-\tau(s))ds.
\end{aligned}$$

Since $V' \leq 0$, we obtain

$$x^2(t)(1-\beta) \leq V(0,x_0) + \int_t^{g(t)} b(s)x^2(s-\tau(s))ds,$$

which by (2.9) gives the stability of the trivial solution.

The boundedness of the function $b(t)$ ensures that the derivative $|x'(t)|$ is bounded. Thus we can apply Lemma 1.3 to $\alpha(t) = x^2(t)$ to get the thesis. □

THEOREM 2.3. *Consider the* DDE (2.2) *and assume* (2.3). *If*

$$\sup_{t\geq 0} \int_t^{g_i(t)} \beta(u)du = B_1 < 1 \quad \textit{for each } i = 1,\dots,N,$$

$$\sup_{t\geq 0} \sum_{i=1}^{N} \int_t^{g_i(t)} b_i(u)du = B_2 < 1, \tag{2.10}$$

$$\inf_{t\geq 0} \beta(t) = B_3 > 0,$$

where $\beta(t) := \sum_{i=1}^{N} b_i(g_i(t))g_i'(t)$, *the trivial solution is uniformly stable. Moreover, if* $\sum_{i=1}^{N} b_i(u)$ *is bounded, then the trivial solution is uniformly asymptotically stable.*

Proof. It is possible to prove this result by using the same technique in Theorem 2.2 defining the nonnegative functional

$$V(t,x_t) = \left(x(t) - \sum_{i=1}^{N} \int_t^{g_i(t)} b_i(s)x(s-\tau_i(s))ds\right)^2$$
$$+ \sum_{i=1}^{N} \int_t^{g_i(t)} \beta(s) \int_s^{g_i(t)} b_i(u)x^2(u-\tau_i(u))du\,ds. \qquad \square$$

Remark 2.4. Since (2.1) and (2.2) are linear, the trivial solution in Theorems 2.2 and 2.3 is globally uniformly asymptotically stable.

Remark 2.5. The boundedness of the solution $b(t)$ in (2.1) is essential to applying Lemma 1.3 in order to get asymptotic stability. This hypothesis can be substituted, for instance, by

$$\int_{t_1}^{t_2} b(s)ds \leq L(t_2 - t_1) \quad \text{for each } t_1, t_2 \geq 0, t_2 \geq t_1.$$

Remark 2.6. Observe that if the delay is bounded or it is constant, the conditions (2.5) and (2.6) are sharper than (2.9) and (2.10).

3. Equations with arbitrary unbounded delay. In this section we consider the scalar real equation

$$x'(t) = a(t)x(t) - b(t)\int_0^{\infty} x(t-s)dk(s), \qquad t \geq 0 \tag{3.1}$$

with the initial condition (1.2). We assume that

(3.2a) $k(t)$ is a function in $[0,\infty)$ with bounded variation;

(3.2b) $\varphi(\theta)$ is such that $\int_0^{\infty} \varphi(t-s)dk(s)$ exists and is finite;

(3.2c) $a : [0,+\infty) \to \mathbf{R}$ is continuous and bounded;

(3.2d) $b : [0,+\infty) \to \mathbf{R}$ is continuous and bounded.

Under the conditions (3.2) we know that (3.1) and (1.2) have a unique solution.

Concerning the bounded variation kernel $k(t)$, it is well known that there exist three uniquely determined functions $k^{(1)}(t)$, $k^{(2)}(t)$, and $k^{(3)}(t)$ whose sum is $k(t)$, i.e.,

$$k(t) = k^{(1)}(t) + k^{(2)}(t) + k^{(3)}(t) \tag{3.3}$$

and such that $k^{(1)}(t)$ is absolutely continuous, $k^{(2)}(t)$ is piecewise constant, and $k^{(3)}(t)$ is a Cantor step function, i.e., an increasing function with derivative almost everywhere equal to zero (see Halmos [19]). So $k^{(1)}(t)$ has the form $\int_0^t f_1(s)ds$, where f_1 is an integrable function (density), $k^{(2)}(t)$ has the form $\sum_{i:\tau_i \le t} \alpha_i$ where, for each i, α_i is the jump in the point h_i and $\sum_{i=1}^{\infty} |\alpha_i| < \infty$.

If in (3.3) $k^{(1)}(t) \equiv 0$, $k^{(3)}(t) \equiv 0$, we can rewrite (3.1) and obtain

$$x'(t) = a(t)x(t) - b(t)\sum_{i=1}^{\infty} x(t-\tau_i)\alpha_i. \tag{3.4}$$

Observe that (3.4) is a particular case of (2.2) and it reduces to the test equation in [20] for $i = 1$.

If in equation (3.3) $k^{(2)}(t) \equiv 0$, $k^{(3)}(t) \equiv 0$, then we get

$$x'(t) = a(t)x(t) - b(t)\int_0^{\infty} x(t-s)f_1(s)ds.$$

Equation (3.1) was studied, for instance, in [16] and [21]. We want to stress that here we do not require the coefficient $a(t)$ to be nonnegative. The following theorem is the main result of this section.

THEOREM 3.1. *Consider* (3.1) *and assume* (3.2). *Define*

$$\gamma_1(t) := \int_0^{\infty} b(t+s)dk(s); \tag{3.5a}$$

$$\gamma_2(t) := \int_0^{\infty} \int_{t-s}^{t} |b(u+s)|du|dk(s)|; \tag{3.5b}$$

$$\gamma_3(t) := \int_0^{\infty} \left[|b(t+s)| \int_{t-s}^{t} |a(v+s) - \gamma_1(v+s)|dv\right] |dk(s)|. \tag{3.5c}$$

If

$$\sup_{t\ge 0}[2(a(t) - \gamma_1(t)) + |a(t) - \gamma_1(t)|\gamma_2(t) + \gamma_3(t)] = -A < 0, \tag{3.6a}$$

$$\sup_{t\ge 0} \gamma_2(t) = B < 1, \tag{3.6b}$$

then the trivial solution of (3.1) *is globally uniformly asymptotically stable* (*see Remark* 2.4).

Proof. To define the functional V, we shall construct two functionals, V_1 and V_2. Let

$$V_1(t, x_t) := \left(x(t) - \int_0^{\infty} \int_{t-s}^{t} b(u+s)x(u)du\, dk(s)\right)^2,$$

so

$$V_1' = 2(a(t) - \gamma_1(t))x^2(t) - 2(a(t) - \gamma_1(t))x(t) \int_0^\infty \int_{t-s}^t b(u+s)x(u)du\, dk(s).$$

If we apply the Cauchy inequality we get

$$\begin{aligned} V_1' &\le 2(a(t) - \gamma_1(t))x^2(t) + |a(t) - \gamma_1(t)| \int_0^\infty \int_{t-s}^t |b(u+s)|(x^2(t) + x^2(u))du|dk(s)| \\ &= (2(a(t) - \gamma_1(t)) + |a(t) - \gamma_1(t)|\gamma_2(t))x^2(t) \\ &\quad + |a(t) - \gamma_1(t)| \int_0^\infty \int_{t-s}^t |b(u+s)|x^2(u)du|dk(s)|. \end{aligned}$$

Define another functional, $V_2(t, x_t)$:

$$V_2(t, x_t) := \int_0^\infty \int_{t-s}^t \left[|a(v+s) - \gamma_1(v+s)| \int_v^t |b(u+s)|x^2(u)du \right] dv|dk(s)|,$$

so

$$V_2' = \gamma_3(t)x^2(t) - |a(t) - \gamma_1(t)| \int_0^\infty \int_{t-s}^t |b(u+s)|x^2(u)du|dk(s)|.$$

Now we are in a position to define the final functional $V(t, x_t)$:

$$V(t, x_t) := V_1(t, x_t) + V_2(t, x_t).$$

Observe that it is nonnegative for each t.

$$V' = V_1' + V_2' \le (2(a(t) - \gamma_1(t)) + |a(t) - \gamma_1(t)|\gamma_2(t) + \gamma_3(t))x^2(t).$$

By condition (3.6a) we have that

$$V' \le -Ax^2(t),$$

and so, for each t,

$$V(t, x_t) - V(0, x_0) \le -A \int_0^t x^2(s)ds, \tag{3.7}$$

$$A \int_0^t x^2(s)ds \le V(0, x_0) \tag{3.8}$$

because V is nonnegative for each t. From (3.8) we get

$$\int_0^{+\infty} x^2(s)ds < \infty. \tag{3.9}$$

Now, from (3.7) we have the inequality $V(t, x_t) \le V(0, x_0)$ for each t and from the definition of V, and in particular because V_2 is nonnegative, and using the Cauchy inequality, we get

$$(1 - \gamma_2(t))x^2(t) \le V(0, x_0) + \int_0^\infty \int_{t-s}^t |b(u+s)|x^2(u)du|dk(s)|.$$

By the boundedness of the function b, by (3.9), (3.6b), and by the hypothesis on the function k, we have that there exists a positive constant A_1 such that $x^2(t) \leq A_1 V(0, x_0)$, and this implies that the solution $x(t)$ is bounded, i.e., $|x(t)| < \infty$ for each t. Now, by (3.1), the boundedness of $x(t)$, and the coefficient $a(t)$, we get $|x'(t)| < \infty$ for each t. So, by applying Lemma 1.3 with $\alpha(t) = x^2(t)$, we get the thesis. □

THEOREM 3.2. *Consider the equation*

$$x'(t) = a(t)x(t) - \sum_{i=1}^{N} b_i(t) \int_0^\infty x(t-s)dk_i(s), \qquad t \geq 0. \tag{3.10}$$

Assume (3.2b–c) *and that* $k_i(s)$ *and* $b_i(s), i = 1, \ldots, N$, *verify, respectively,* (3.2a) *and* (3.2d). *If* (3.6) *hold with*

$$\gamma_1(t) := \sum_{i=1}^{N} \int_0^\infty b_i(t+s)dk_i(s);$$

$$\gamma_2(t) := \sum_{i=1}^{N} \int_0^\infty \int_{t-s}^{t} |b_i(u+s)|du|dk_i(s)|;$$

$$\gamma_3(t) := \sum_{i=1}^{N} \int_0^\infty \left[|b_i(t+s)| \int_{t-s}^{t} |a(v+s) - \gamma_1(v+s)|dv \right] |dk_i(s)|,$$

then the trivial solution of (3.10) *is globally, uniformly, asymptotically stable.*

Proof. It is possible to prove this result by using the same technique in Theorem 3.1 defining the nonnegative functionals

$$V_1(t, x_t) := \left(x(t) - \sum_{i=1}^{N} \int_0^\infty \int_{t-s}^{t} b_i(u+s)x(u)du\, dk_i(s) \right)^2;$$

$$V_2(t, x_t) := \sum_{i=1}^{N} \int_0^\infty \int_{t-s}^{t} \left[|a(v+s) - \gamma_1(v+s)| \int_v^t |b_i(u+s)|x^2(u)du \right] dv|dk_i(s)|. \quad \square$$

Remark 3.3. We remember that conditions (3.6) do not imply that the coefficient $a(t)$ must be negative. For instance, we can put $a(t) \equiv 0$ and obtain the same conditions (3.6) with

$$\gamma_3(t) = \int_0^\infty \left[|b(t+s)| \int_{t-s}^{t} |\gamma_1(v+s)|dv \right] |dk(s)|.$$

Remark 3.4. If in (3.1) the coefficients $a(t)$ and $b(t)$ are constant, i.e., $a(t) \equiv a$, $b(t) \equiv b$, (3.5) reduces to

$$\gamma_1 = b \int_0^\infty dk(s), \quad \gamma_2 = |b|\beta_1, \quad \gamma_3 = |b||a - \gamma_1|\beta_1,$$

where $\beta_1 = \int_0^\infty s|dk(s)|$. Observe that γ_1, γ_2, γ_3 are not depending on t. Conditions (3.6) yield

$$(a - \gamma_1) + |b|\ |a - \gamma_1|\beta_1 < 0 \tag{3.11a}$$

$$|b|\beta_1 < 1. \tag{3.11b}$$

It is easy to prove that (3.11a) is equivalent to (3.11b) provided that $(a-\gamma_1)<0$. If, in addition, $k(t)$ is equal to 0 for $t<\tau$ and equal to 1 for $t\geq\tau$ $(\tau>0)$, we obtain the equation

$$x'(t)=ax(t)-bx(t-\tau)$$

and conditions $|b|\tau<1$, $a-b<0$. Note that these conditions are also valid if the coefficient a is nonnegative.

Remark 3.5. If in (3.1) we have no delay or, equivalently, if $k(t)$ is equal to 0 for $t<0$ and equal to 1 for $t\geq 0$, we get the equation

$$x'(t)=a(t)x(t)-b(t)x(t),$$

$\gamma_1(t)=b(t)$, $\gamma_2(t)\equiv 0$, $\gamma_3(t)\equiv 0$, and the well-known condition

$$\sup_{t\geq 0}(a(t)-b(t))=-A<0.$$

4. Stability of chemostat equations with distributed delay. In this section we consider an open system where nutrient is consumed by a micro-organism and partially recycled. Under some assumptions such a system can be modeled in terms of chemostat equations of the form (see Beretta and Bischi [5] and Beretta and Fasano [6]) for $t\geq 0$,

$$N_1'(t)=u-a_{12}U(N_1(t))N_2(t)+be_2\int_{-\infty}^{t}f_2(t-s)N_2(s)ds, \tag{4.1a}$$

$$N_2'(t)=N_2(t)\left[-e_2+\gamma_2\int_{-\infty}^{t}f_1(t-s)U(N_1(s))ds\right], \tag{4.1b}$$

where N_1 is the concentration of the limiting nutrient, N_2 is the concentration of a species of micro-organism, $u>0$ is a constant nutrient supply, $a_{12}>0$ is the maximum uptake rate, $e_2>0$ is the death rate of micro-organism, $\gamma_2>0$ is the maximum growth rate, and $b\in(0,1)$ is the fraction of the dead biomass that is recycled as a new nutrient. The function $U(N)$ is the quantity of nutrient consumed by the species: it is a continuous, bounded, increasing function of $N\in[0,\infty)$ with

$$U(0)=0,\quad U'(N)=\frac{dU(N)}{dN}>0,\quad \lim_{N\to\infty}U(N)=1.$$

Integral terms in (4.1) reflect the influence of the previous values of N_1, N_2 on their current alterations. The functions $f_i(s)$, $i=1,2$ are nonnegative, square integrable on $[0,\infty)$ and such that $\int_0^\infty f_i(s)ds=1$, $i=1,2$. If

$$e_2<\gamma_2,\qquad a_{12}>b\gamma_2, \tag{4.2}$$

the system (4.1) has a unique, positive stationary solution N_1^0, N_2^0 given by

$$N_1^0=U^{-1}\left(\frac{e_2}{\gamma_2}\right),\qquad N_2^0=\frac{u}{e_2(a_{12}/\gamma_2-b)}. \tag{4.3}$$

In the sequel both inequalities (4.2) are assumed to be fulfilled.

By following the usual linearization procedure, the study of the local properties of the equilibrium (4.3) of (4.1) can be reduced to the investigation of the asymptotic

stability of the trivial solution with respect to the initial disturbances in the uniform norm of the following linearized equations:

$$x_1'(t) = -a_{12}N_2^0U'(N_1^0)x_1(t) - a_{12}\frac{e_2}{\gamma_2}x_2(t) + be_2\int_{-\infty}^{t} f_2(t-s)x_2(s)ds, \tag{4.4a}$$

$$x_2'(t) = \gamma_2 N_2^0 U'(N_1^0)\int_{-\infty}^{t} f_1(t-s)x_1(s)ds. \tag{4.4b}$$

In turn a necessary and sufficient condition for the asymptotic stability of the system (4.4) (i.e., the trivial solution if asymptotically stable) is that

$$D(z) := z^2 + az + ae_2F_1(z) - \frac{be_2\gamma_2}{a_{12}}aF_1(z)F_2(z) \neq 0 \quad \text{when } \mathrm{Re}(z) \geq 0, \tag{4.5}$$

where $a = a_{12}N_2^0U'(N_1^0)$. Here

$$F_i(z) = \int_0^{\infty} e^{-zs}f_i(s)ds, \qquad i = 1,2$$

for $i = 1,2$ are the Laplace transform of the kernels f_i, $i = 1,2$. $D(z)$ is defined as a *characteristic equation* of (4.4).

Note that under our assumptions the functions $F_i(z)$, generally speaking, are not even defined at the half plane $\mathrm{Re}(z) < 0$.

Also note that

$$\begin{aligned} F_1(z)F_2(z) &= \int_0^{\infty} f_1(s)\int_0^{\infty} f_2(s_1)e^{-z(s+s_1)}ds_1ds \\ &= \int_0^{\infty} f_1(s)\int_s^{\infty} f_2(s_2-s)e^{-zs_2}ds_2ds \\ &= \int_0^{\infty} e^{-zs_2}\int_0^{s_2} f_1(s)f_2(s_2-s)ds\,ds_2. \end{aligned}$$

Hence by defining

$$f(s) := ae_2\left[f_1(s) - \frac{b\gamma_2}{a_{12}}\int_0^{s} f_1(s_1)f_2(s-s_1)ds_1\right], \tag{4.6}$$

we can rewrite the characteristic equation of (4.4) in the following form:

$$D(z) = z^2 + az + \int_0^{\infty} e^{-zs}fd(s)ds. \tag{4.7}$$

The latter equation represents a characteristic equation for the following scalar system of the second order

$$x''(t) + ax'(t) + \int_0^{\infty} x(t-s)f(s)ds = 0,$$

which is equivalent to the system of two differential equations of the first order ($t \geq 0$)

$$x'(t) = y(t), \tag{4.8a}$$

$$y'(t) = -ay(t) - \int_0^{\infty} x(t-s)f(s)ds. \tag{4.8b}$$

Consequently, asymptotic stability conditions for (4.8) will be simultaneously sufficient for the local stability of stationary solution (4.3) of system (4.1). Obtain these conditions using the generalization of Liapunov direct method for functional differential equations (see [22]).

THEOREM 4.1. *Consider the system* (4.1) *with the assumption* (4.2). *If*

$$\beta_0 > 0, \quad \beta_1 < a, \quad \beta_2 < +\infty, \tag{4.9}$$

where $\beta_0 = \int_0^\infty f(s)ds$, $\beta_1 = \int_0^\infty s|f(s)|ds$, $\beta_2 = \int_0^\infty s^2|f(s)|ds$, *and* f *is given by* (4.6) *and* $a = a_{12}N_2^0U'(N_1^0)$, *then the positive stationary solution* (4.3) *of* (4.1) *is locally stable.*

Proof. Consider the functional

$$\begin{aligned} V &:= 4\beta_0 x^2(t) + y^2(t) + V_0^2(t) \\ &\quad + \int_0^\infty |f(s)| \int_{t-s}^t \int_{t_1}^t (y^2(t_2) + \beta_0 x^2(t_2))dt_2 dt_1 ds, \end{aligned} \tag{4.10}$$

where

$$V_0(t) := y(t) + ax(t) - \int_0^\infty f(s) \int_{t-s}^t x(t_1)dt_1 ds.$$

By the assumption $\beta_0 > 0$, $\beta_2 < \infty$, the functional (4.10) is positive definite and has upper infinitesimal limit. Find the derivative V' of this functional along the trajectories of (4.8). Note that

$$\begin{aligned} \frac{d}{dt}y^2(t) &= 2y(t)\left[-ay(t) - \int_0^\infty x(t-s)f(s)ds\right] \\ &= 2\left[-ay^2(t) + y(t)\int_0^\infty f(s)\int_{t-s}^t x'(s_1)ds_1 ds - \beta_0 x(t)y(t)\right] \\ &\le -2ay^2(t) - 2\beta_0 x(t)y(t) + y^2(t)\beta_1 + \int_0^\infty |f(s)| \int_{t-s}^t y^2(s_1)ds_1 ds. \end{aligned} \tag{4.11}$$

Further we get, since $\beta_0 > 0$,

$$\begin{aligned} (V_0^2)' &= -2V_0\beta_0 x(t) = -2\beta_0 x(t)\left[y(t) + ax(t) - \int_0^\infty f(s)\int_{t-s}^t x(t_1)dt_1 ds\right] \\ &\le \beta_0\left[-2x(t)y(t) - 2ax^2(t) + \beta_1 x^2(t) + \int_0^\infty |f(s)| \int_{t-s}^t x^2(t_1)dt_1 ds\right]. \end{aligned} \tag{4.12}$$

Finally, relations (4.9)–(4.11) give us

$$\begin{aligned} V' &\le 4\beta_0 x(t)y(t) - 2ay^2(t) - 2\beta_0 x(t)y(t) + \beta_1 y^2(t) + \int_0^\infty |f(s)| \int_{t-s}^t y^2(s_1)ds_1 ds \\ &\quad -2\beta_0 x(t)y(t) - 2a\beta_0 x^2(t) + \beta_1\beta_0 x^2(t) + \beta_0 \int_0^\infty |f(s)| \int_{t-s}^t x^2(s_1)ds_1 ds \\ &\quad - \int_0^\infty |f(s)| \int_{t-s}^t [y^2(s_1) + \beta_0 x^2(s_1)]ds_1 ds + (y^2(t) + \beta_0 x^2(t))\beta_1 \\ &\le -2(a - \beta_1)[\beta_0 x^2(t) + y^2(t)]. \end{aligned}$$

By (4.9) we have that the system (4.8) is asymptotically stable. □

Remark 4.2. Note that inequality $\beta_0 > 0$ is necessary for asymptotic stability of (4.8). Really if $\beta_0 \leq 0$, then characteristic equation (4.7) will have nonnegative real roots because the function (4.7) is such that $D(0) \leq 0$ and $D(z) \to \infty$ as $\mathrm{Re}(z) \to \infty$, $\mathrm{Im}(z) = 0$.

Note also that for an ordinary differential equation of second order,

$$x''(t) + ax'(t) + \beta_0 x(t) = 0$$

(which arises if $f(s) = \beta_0 \delta(s)$, where $\delta(s)$ is a delta function), inequality (4.12) goes into $a > 0$, $\beta_0 > 0$, which are necessary and sufficient stability conditions in this case.

REFERENCES

[1] V. K. BARWELL, *Special stability problems for functional differential equations*, BIT, 15 (1975), pp. 130–135.

[2] A. BELLEN, *One-step collocation for delay differential equations*, J. CAM, 10 (1984), pp. 275–283.

[3] ———, *A Runge–Kutta–Nystrom method for delay differential equations*, Progr. Sci. Comput., 5 (1985), pp. 271–283.

[4] R. BELLMAN AND L. K. COOKE, *Differential Difference Equations*, Academic Press, London, 1963.

[5] E. BERETTA AND G. I. BISCHI, *Stability and Hopf bifurcation in some nutrient species models with nutrient cycling and time lags—Biomedical modelling and simulation*, J. Eisenfeld and D. S. Levine, eds., IMACS, (1989), pp. 175–181.

[6] E. BERETTA AND A. FASANO, *A mathematical model for the dynamics of a phytoplankton population*, Proc. of Claremont Conference on Differential Equations and Application to Biology and Medicine, 1990.

[7] T. A. BURTON, *Volterra Integral and Differential Equations*, Academic Press, New York, 1983.

[8] T. A. BURTON AND J. R. HADDOCK, *On the delay-differential equations* $x'(t) + a(t)f(x(t - r(t))) = 0$ *and* $x''(t) + af(x(t - r(t))) = 0$, J. Math. Anal. Appl., 54 (1976), pp. 37–48.

[9] K. L. COOKE, *Asymptotic theory for the delay-differential equation* $u'(t) = -au(t - r(u(t)))$, J. Math. Anal. Appl., 19 (1967), pp. 160–173.

[10] ———, *Linear functional differential equations of asymptotically autonomous type*, J. Differential Equations, 7 (1970), pp. 154–174.

[11] ———, *Stability of non-autonomous delay differential equations by Liapunov functionals*, Lecture Notes in Math., 1076 (1983), pp. 41–52.

[12] C. CORDUNEANU AND V. LAKSHMIKANTHAM, *Equations with unbounded delay: a survey*, Nonlinear Anal. Th. Meth. Appl., 4 (1980), pp. 831–877.

[13] G. GRIPENBERG, S. LONDEN, AND O. J. STAFFANS, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, 1991.

[14] I. GYORI, *Necessary and sufficient conditions in an asymptotically ordinary delay differential equation*, Differential Integral Equations, 6 (1993), pp. 225–239.

[15] J. R. HADDOCK, *On the asymptotic behaviour of solutions* $x'(t) = -a(t)f(x(t - r(t)))$, SIAM J. Math. Anal., 5 (1974), pp. 569–573.

[16] J. R. HADDOCK AND Y. KUANG, *Asymptotic theory for a class of nonautonomous delay differential equations*, J. Math. Anal. Appl., 168 (1992), pp. 147–162.

[17] E. HAIRER, S. P. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations* I, Springer-Verlag, Berlin, 1987.

[18] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[19] P. R. HALMOS, *Measure Theory*, Van Nostrand, New York, 1951.

[20] Y. HINO, S. MURAKAMI, AND T. NAITO, *Functional differential equations with infinite delay*, Lecture Notes in Math. 1473, 1991.

[21] T. KRISZTIN, *On the stability properties for one-dimensional functional differential equations*, Funck. Ekvac., 34 (1991), pp. 241–256.

[22] V. B. KOLMANOVSKII AND V. R. NOSOV, *Stability of Functional Differential Equations*, Academic Press, New York, London, 1986.

[23] G. MEINARDUS AND G. NURNBERGER, EDS., *Delay Equations, Approximation and Application*, International Symposium at the University of Manheim, October 8–11, 1984, Birkhauser-Verlag, Basel, Switzerland.

[24] L. TORELLI AND R. VERMIGLIO, *On the stability of continuous quadrature rules for differential equations with several constant delays*, IMA J. Numer. Anal., 13 (1993), pp. 291–302.

[25] T. YONEYAMA, *On the 3/2 stability theorem for one-dimensional delay-differential equations*, J. Math. Anal. Appl., 125 (1987), pp. 161–173.

[26] J. A. YORKE, *Asymptotic stability for one dimensional differential-delay equations*, J. Differential Equations, 7 (1970), pp. 189–202.

[27] M. ZENNARO, *P-stability properties of Runge–Kutta methods for delay differential equations*, Numer. Math., 49 (1986), pp. 305–318.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 962–975, May 1994

009

HERMITE INTERPOLATION ON THE LATTICE $\mathbb{Z}^d$*

KURT JETTER†, SHERMAN D. RIEMENSCHNEIDER‡ AND ZUOWEI SHEN§

Abstract. This paper deals with the interpolation of derivative data on the integer lattice $\mathbb{Z}^d$ by means of spaces generated by the lattice translates of several functions. The derivatives to be interpolated can be of a general form, given by a set of linear constant coefficient differential operators induced by a linearly independent set of polynomials on $\mathbb{R}^d$; for example, successive partial derivatives (Hermite interpolation) or powers of the Laplacian. The method used here is to adapt the generating functions of the interpolation space to the derivatives to be interpolated. This is done by introducing oscillations in the space through multiplication by certain shift invariant (1-periodic) trigonometric polynomials. The resulting interpolation schemes do provide convergence to smooth functions as the mesh size is reduced to zero under suitable restrictions.

Key words. cardinal interpolation, Hermite interpolation, multivariate interpolation, splines, radial basis functions

AMS subject classifications. 41A05, 65D05, 41A15, 65D07, 41A29, 41A63

1. Introduction.

In this paper we take another look at the general problem of interpolating derivative data on the lattice $\mathbb{Z}^d$. So far the following situations were considered in the literature:

- interpolation of data on the lattice from spaces generated by lattice translates of either compactly supported functions (cf. the survey [13]) or radial basis functions (cf. [12]);
- hermite interpolation of derivative values by spaces of box splines [14] or by similarly generated spaces (cf. [13]);
- interpolation on periodic meshes [4].

For many spaces generated by lattice translates of compactly supported functions, the exponential decay of the fundamental solutions for the interpolation is lost in passing from interpolation of function values to interpolation of derivative values when $d > 1$. The reason for this may be that the proper space of interpolating functions has not yet been discovered (or that the univariate model is being forced on the multivariate setting). In this paper, we search for an appropriate generating family that will provide fundamental solutions with exponential decay when such solutions exist for interpolation of function values. The essential idea is to introduce oscillations in the approximating family that are appropriate to the problem at hand.

The general cardinal interpolation problem reads as follows: For given functions

$$\phi_1, \ldots, \phi_r : \mathbb{R}^d \to \mathbb{C}$$

and functionals $\lambda_1, \ldots, \lambda_r$, we want to interpolate (real or complex) data

$$d_k := (d_k(\alpha))_{\alpha \in \mathbb{Z}^d}, \qquad k = 1, \ldots, r,$$

*Received by the editors March 2, 1992; accepted for publication (in revised form) April 7, 1993. This research was supported by National Science and Engineering Research Council of Canada grant A7687.

†Fachbereich Mathematik, Universität Duisburg, 4100 Duisburg 1, Germany. This author's research was partially supported by North Atlantic Treaty Organization grant CRG 900158.

‡Department of Mathematics, University of Alberta, Edmonton, Canada. This author's research was partially supported by North Atlantic Treaty Organization grant CRG 901018.

§Center for Mathematical Sciences, 1308 West Dayton St., Madison, Wisconsin 53706. This author's research was partially supported under North Atlantic Treaty Organization CRG 901018 and National Science Foundation grant DMS-9000053.

on the lattice $\mathbb{Z}^d$; i.e., we want to find sequences

$$c_j = (c_j(\alpha))_{\alpha\in\mathbb{Z}^d}, \qquad j = 1,\dots,r,$$

such that $S(x) := \sum_{\alpha\in\mathbb{Z}^d}\sum_{j=1}^{r} c_j(\alpha)\phi_j(x-\alpha)$ satisfies the interpolation conditions

$$\lambda_k S(\cdot+\beta) = d_k(\beta), \quad \beta\in\mathbb{Z}^d, \quad k = 1,\dots,r. \tag{1.1}$$

For simplicity, we use the following notation for the semidiscrete convolution of a sequence with a function

$$\phi *' c := \sum_{\alpha\in\mathbb{Z}^d} \phi(\cdot-\alpha)c(\alpha).$$

The solution $S(x)$ then comes from the space

$$\mathcal{S} := \left\{ S(x) = \sum_{j=1}^{r} \phi_j *' c_j \,|\, c_j : \mathbb{Z}^d \to \mathbb{C} \right\},$$

and the functions $\phi_1,\dots,\phi_r$ are the *generators* of $\mathcal{S}$. In the references mentioned above, the linear functionals are either point evaluation at the origin with $r = 1$ for the interpolation of function values [12], [13] or point evaluation of successive derivatives [14], [13], or the functionals $\lambda_i : f \mapsto f(\tau_i + \cdot)$ with $\tau_i \in [0,1), i = 1\dots,r$ in [4].

It is well known (cf. [13]) that problem (1.1) can be transformed to a system of linear equations with the use of symbols. Define the (*periodic*) *symbol* of the $\ell_1(\mathbb{Z}^d)$ sequence $(\lambda_k\phi_j(\cdot+\alpha))_{\alpha\in\mathbb{Z}^d}$ as

$$A_{k,j}(\xi) := \sum_{\alpha\in\mathbb{Z}^d} \lambda_k\phi_j(\cdot+\alpha)e^{-i\alpha\cdot\xi}, \quad \xi\in\mathbb{R}^d, \quad 1\le k,j\le r. \tag{1.2a}$$

More generally, the (*complex*) *symbol* of the sequence is the Laurent series

$$a_{k,j}(z) = \sum_{\alpha\in\mathbb{Z}^d} \lambda_k\phi_j(\cdot+\alpha)z^{\alpha}, \tag{1.2b}$$

so that (1.2a) equals (1.2b) on the d-dimensional torus

$$\mathbf{T}^d := \{z = (e^{-i\xi(1)},\dots,e^{-i\xi(d)}) : \xi(m)\in[0,2\pi),\ m = 1,\dots,d\}.$$

The symbol matrix for the interpolation problem is then given by

$$A(\xi) := (A_{k,j}(\xi))_{k,j=1}^{r} \tag{1.3}$$

and (1.1) transforms into the following linear system of equations:

$$\sum_{j=1}^{r} A_{k,j}(\xi)C_j(\xi) = D_k(\xi), \qquad k = 1,\dots,r, \tag{1.4}$$

where

$$C_j(\xi) = \sum_{\alpha} c_j(\alpha)e^{-i\alpha\cdot\xi} \quad \text{and} \quad D_k(\xi) = \sum_{\alpha} d_k(\alpha)e^{-i\alpha\cdot\xi}$$

are the formal Fourier series associated with the sequences c_j and d_k. Hence,

$$\det A(\xi) \neq 0 \quad \text{for all } \xi \in \mathbb{R}^d \tag{1.5}$$

is a necessary condition for the unique solution of (1.4) for the unknown functions C_j. In turn, the properties of these solutions determine whether solving (1.4) is equivalent to solving (1.1).

Of particular importance is the search for the fundamental solutions, where the right-hand sides of (1.4) are chosen successively from the r-dimensional canonical unit vectors, or equivalently, all data sequences in (1.1) are zero, except the one that is the Kronecker sequence: $\delta(0) = 1$ and $\delta(\alpha) = 0$, otherwise. Now if the entries of the symbol matrix are analytic on the torus, and if (1.5) holds, then the solutions C_j will be analytic on the torus as well, and the Fourier coefficients of these will decay exponentially. In this case, (1.4) is equivalent to (1.1) for data sequences d_k of at most polynomial growth.

Our approach to this general cardinal Hermite interpolation problem (CHIP) given in §2 can be described as follows: for given functionals $\lambda_1, \ldots, \lambda_r$ find a suitable basis of generators $\phi_1, \ldots, \phi_r$ so that the symbol matrix is lower triangular with diagonal elements $A_{k,k}(\xi) \neq 0, \xi \in \mathbb{R}^d$. It is shown that this property holds when $\phi_j = \sigma_j \psi_j$, where σ_j is a properly chosen trigonometric polynomial and ψ_j is a function with a nonvanishing symbol. This particular choice of functions seems to be somewhat reminiscent of the construction of a so-called Wilson basis in [7], although they only considered the univariate case. The construction is carried out when the functionals are given by linearly independent constant coefficient partial differential operators. Two important cases, interpolation of successive derivatives as defined by an arbitrary lower set of $\mathbb{Z}_+^d$ and the interpolation to powers of the Laplacian, are given as the main examples. The simplest case occurs when the functions ψ_j are all the same function ϕ (and consequently, the ϕ_j are given by multiplication of ϕ by the appropriate trigonometric polynomials). Concrete examples are given by taking ϕ as a box spline in which case the approximating family is generated by integer translates of compactly supported piecewise exponential polynomials with the smoothness properties of the box spline. In §3 we consider the question of whether the scaled interpolation operators provide the approximation order determined from the Strang–Fix conditions. The Fourier transform of the fundamental functions shows that the algorithms of [8] can be applied for numerical computations.

2. A general cardinal interpolation problem. We consider the problem of interpolating derivative data on the integer lattice $\mathbb{Z}^d$ when the derivatives are given by a linearly independent set of real constant coefficient linear partial differential operators (complex coefficients could also be considered; see the remark below). Let $p_1, \ldots, p_r$ be linearly independent polynomials on $\mathbb{R}^d$. Since we are dealing with a linear interpolation process without loss of generality, we may assume the following: (i) in the representation

$$p_j = \sum_{m=0}^{n_j} f_{j,m}, \qquad f_{j,m} \quad \text{homogeneous of degree } m,$$

the leading terms $f_{j,n_j}, j = 1, \ldots, r$, are nonzero and distinct; (ii) $n_k \leq n_j$ if $k < j$ and the leading homogeneous term of p_j contains a monomial that does not appear in p_k for all $k < j$. (This may be achieved by taking linear combinations and relabelling the polynomials. It should be noted, however, that our choice of generators will depend

on this representation.) The linear functionals for the interpolation problem will be $\lambda_k f := p_k(D)f(0)$.

We associate a homogeneous polynomial q_j with each polynomial p_j as follows: Choose q_j to be that portion of the leading homogeneous term f_{j,n_j} containing those monomials not in p_k, $k < j$. Define

$$\sigma_j(x) := q_j(e^{2\pi i x} - 1), \qquad j = 1, \dots, r, \tag{2.1}$$

where we have adopted the multivariate notation

$$(e^{2\pi i x} - 1)^\alpha := \prod_{m=1}^{d} (e^{2\pi i x(m)} - 1)^{\alpha(m)}. \tag{2.2}$$

We remark that other choices are possible here. For example, we may discard some monomials from q_j, and/or we may replace $e^{2\pi i x} - 1$ in the definition of σ_j by $\sin(2\pi x)$ and sometimes even by $\sin(\pi x)$. These changes require very modest changes in the arguments presented below. Our choice of $e^{2\pi i x} - 1$ makes the proof simpler, but it may also make the interpolatory process complex even when the functions ψ_j and the data are real-valued. However, in the latter case, the imaginary part of the interpolant interpolates the zero data and goes to zero with the mesh size (cf. §3).

Let $\psi_1, \dots, \psi_r$ be given functions in $C^\kappa(\mathbb{R}^d)$ with $\kappa = \max_j n_j$ and for which the sequences

$$\psi_{j|} := \{\psi_j(\alpha)\}_{\alpha \in \mathbb{Z}^d} \in \ell_1(\mathbb{Z}^d). \tag{2.3}$$

Denote their corresponding symbols by

$$\widetilde{\psi}_j(\xi) := \sum_{\alpha \in \mathbb{Z}^d} \psi_j(\alpha) e^{-i\alpha\xi}, \qquad j = 1, \dots, r. \tag{2.4}$$

Define the functions

$$\phi_j := \sigma_j \psi_j, \qquad j = 1, \dots, r. \tag{2.5}$$

We are now in a position to state our main results.

THEOREM 2.1. *In addition to the above definitions and assumptions, assume that the sequences*

$$(p_k(D)\phi_j)_| := \{(p_k(D)\phi_j)(\alpha)\}_{\alpha \in \mathbb{Z}^d} \in \ell_1(\mathbb{Z}^d).$$

Then the symbol matrix $A(\xi) = (A_{k,j}(\xi))_{k,j=1}^r$ *for the linear functionals* $\lambda_k f := p_k(D)f(0)$ *and the generators* $\phi_1, \dots, \phi_r$ *is lower triangular with diagonal entries of the form*

$$A_{j,j}(\xi) = \text{const}_j \widetilde{\psi}_j(\xi),$$

where $\text{const}_j \neq 0$. *In particular,* $\det A(\xi) \neq 0$ *for all* $\xi \in \mathbb{R}^d$ *if and only if all the symbols* $\widetilde{\psi}_j$ *satisfy* $\widetilde{\psi}_j(\xi) \neq 0, \xi \in \mathbb{R}^d$.

The proof is an immediate consequence of the following lemma.

LEMMA 2.2. *We have*

$$(p_k(D)\phi_j)(\alpha) = 0 \quad \text{for all } \alpha \in \mathbb{Z}^d \quad \text{if } k < j. \tag{2.6}$$

When $k = j$,

$$(p_j(D)\phi_j)(\alpha) = (q_j(D)q_j)(0)\psi_j(\alpha), \qquad \alpha \in \mathbb{Z}^d, \tag{2.7}$$

and $(q_j(D)q_j)(0)$ is a nonzero constant.

Proof. Let

$$p_k(x) = \sum_\beta a_{k,\beta}\frac{x^\beta}{\beta!} \quad \text{and} \quad q_j(x) = \sum_\gamma b_{j,\gamma}\frac{x^\gamma}{\gamma!}.$$

Then by our choices of p_k and q_j, whenever $j > k$, given monomials x^β from p_k and x^γ from q_j with nonzero coefficients, there is a component $m = m(\beta,\gamma)$ for which $\beta(m) < \gamma(m)$. Hence,

$$(D^\beta((e^{2\pi i\cdot} - 1)^\gamma\psi_j))(\alpha) = \sum_{\mu\le\beta}\binom{\beta}{\mu}(D^\mu((e^{2\pi i\cdot} - 1)^\gamma)D^{\beta-\mu}\psi_j)(\alpha) = 0 \tag{2.8}$$

$$\forall \alpha \in \mathbb{Z}^d,$$

by the Leibnitz formula since each summand contains a power of $(e^{2\pi i\alpha(m)} - 1)$. This gives relation (2.6).

The same is true if $k = j$ except when $\beta = \gamma$, which is now permitted. In this case, only the term $\mu = \beta = \gamma$ in (2.8) does not give zero when evaluated on the integer lattice; and that term yields $(2\pi i)^{n_j}\gamma!\psi_j(\alpha)$ since $|\gamma| = n_j$. When these terms appear in the evaluation of $p_j(D)$ on $\sigma_j\psi_j$ at the integer lattice, we obtain

$$(p_j(D)\phi_j)(\alpha) = (q_j(D)q_j)(0)\psi_j(\alpha) = \left((2\pi i)^{n_j}\sum_{|\gamma|=n_j}|b_{j,\gamma}|^2/\gamma!\right)\psi_j(\alpha). \qquad \square$$

Remark. To allow for the complex coefficient polynomials p_j, we could take the functionals to be $\gamma_k f := \bar{p}_k(D)f(0)$, in which case the last proof is exactly the same when $p_j(D)$ and $q_j(D)$ are replaced by $\bar{p}_j(D)$ and $\bar{q}_j(D)$, respectively.

We give three bivariate examples where the conditions of the theorem are satisfied:

- for a (centered) 3-directional box spline $\psi_j = M^c_{k,\ell,m}$, the symbol $\widetilde{\psi}_j$ does not vanish on $\mathbb{R}^2$ (even if we use shifted box splines $M^c_{k,\ell,m}(x + \cdot)$ and the shift x is from a certain hexagonal shift region; cf. [3]). Since $M^c_{k,\ell,m} \in C^\kappa(\mathbb{R}^2)$ with $\kappa = k + \ell + m - \mu - 2$ and $\mu := \max\{k,\ell,m\}$, there are many choices available to fit any interpolation problem;
- in case of bivariate polyharmonic splines we may choose

$$\psi_j(x) = (-\nabla)^k(x(1)^2 + x(2)^2)^{k-1}\log(x(1)^2 + x(2)^2)$$

with $(-\nabla)$ the discrete Laplacian

$$\begin{aligned}(-\nabla)f(x(1),x(2)) = 4f(x(1),x(2)) - \{&f(x(1)-1,x(2)) + f(x(1)+1,x(2))\\ &+ f(x(1),x(2)-1) + f(x(1),x(2)+1)\}\end{aligned}$$

and $2 \le k \in \mathbb{N}$ (cf. [6]); here we have $\psi_j \in C^{k-2}$;

- in case of the bivariate Hardy's multiquadric we may choose

$$\psi_j = -\nabla^{3/2}\sqrt{1 + x(1)^2 + x(2)^2}$$

with the fractional discrete Laplacian as defined in [6].

Let $\mathcal{A}$ denote the 2π-periodic functions with absolutely convergent Fourier series and let $\mathcal{E}$ be the 2π-periodic functions with exponentially decaying Fourier coefficients.

COROLLARY 2.3. *Under the assumptions of Theorem* 2.1, *if the functions* ψ_j *are bounded and* $|\widetilde{\psi_j}| > 0, j = 1, \ldots, r$, *then the functions*

$$L_j := \sum_{m=j}^{r} \phi_m *' c_{m,j}, \qquad j = 1, \ldots, r, \tag{2.9}$$

with coefficient sequences from the functions $C_{m,j}(y) = \sum_\alpha c_{m,j} \exp(-i\alpha y) \in \mathcal{A}$ *defined by*

$$C_{m,j} = 0, \qquad m = 1, \ldots, j-1,$$

$$C_{m,j} = \left(\delta_{m,j} - \sum_{\mu=j}^{m-1} A_{m,\mu}(y) C_{\mu,j}(y)\right) \Big/ A_{m,m}(y), \qquad m = j, \ldots, r, \tag{2.10}$$

are bounded fundamental solutions to the CHIP *determined by* $p_1(D), \ldots, p_r(D)$ *and the space* $\mathcal{S}$ *generated by* $\phi_1, \ldots, \phi_r$. *In other words,* L_j *is the unique bounded solution from* $\mathcal{S}$ *of the problem*

$$(p_k(D)S)_| = \delta_{k,j}\delta, \qquad k = 1, \ldots, r.$$

If the functions $\phi_j, j = 1, \ldots, r$, *are also compactly supported, then the functions* L_j *have exponential decay.*

Proof. Each system of equations

$$A(\xi) \begin{pmatrix} C_{1,j}(\xi) \\ \vdots \\ C_{r,j}(\xi) \end{pmatrix} = e_j^T$$

for the unit vectors e_j of $\mathbb{R}^r$ has a unique solution since the coefficient matrix is lower triangular and the diagonal does not vanish on $\mathbb{R}^d$. Then the form (2.10) for the solution is nothing more than the one obtained through forward substitution. Moreover, $C_{m,j} \in \mathcal{A}$, as follows from (2.10) and the fact that each entry of $A(\xi)$ belongs to $\mathcal{A}$, and Wiener's lemma (which implies that $1/\widetilde{\psi_j}$ belongs to $\mathcal{A}$). In addition, if the functions ψ_j are compactly supported, then by a standard argument, $C_{m,j} \in \mathcal{E}$ since each entry in $A(\xi)$ is a trigonometric polynomial, and $1/\widetilde{\psi_j}$ is analytic in a neighborhood of the torus $\mathbf{T^d}$. It then follows that the functions L_j in (2.9) are well defined and have the stated properties. □

Therefore, we may define the cardinal Hermite interpolation operator by

$$\mathcal{J}f := \sum_{j=1}^{r} L_j *' (p_j(D)f)_| \tag{2.11}$$

whenever the semidiscrete convolution is well defined. Obviously, we must have $f \in C^\kappa(\mathbb{R}^d)$, $\kappa = \max n_j$, but we must also limit the growth of the sequences $(p_j(D)f)_|$ depending on the decay of L_j.

Example: classical cardinal Hermite interpolation. In multivariate Hermite interpolation we use a successive series of partial derivatives as our functionals. This can be described using the notion of lower sets

$$A \subseteq \mathbb{Z}_+^d \quad \text{with } \#\Lambda = r, \tag{2.12}$$

where (by definition) $\lambda \in \Lambda$ and $\mu \in Z_+^d$ with $\mu \le \lambda$ implies $\mu \in \Lambda$; here $\le$ is the usual partial ordering of vectors in $\mathbb{R}^d$. With $\lambda \in \Lambda$ we associate the polynomial $p_\lambda(x) = x^\lambda$ and with an abuse of terminology the linear functional

$$\lambda : f \mapsto D^\lambda f(0) := \frac{\partial^{|\lambda|}}{\partial_{x(1)}^{\lambda(1)} \cdots \partial_{x(d)}^{\lambda(d)}} f(0, \ldots, 0),$$

corresponding to $p_\lambda(D) = D^\lambda$. Hence Hermite interpolation of total order $\kappa \in \mathbb{N}_0$ is described by the triangular set $\{\lambda \in \mathbb{Z}_+^d; |\lambda| \le \kappa\}$ while Hermite interpolation of coordinate order $\kappa \in \mathbb{N}_0^d$ is described by the rectangular set $\{\lambda \in \mathbb{Z}_+^d; \lambda \le \kappa\}$. Moreover, every lower set Λ is the union of finitely many rectangular sets.

For the lower set (2.12) let

$$\Lambda = \{\lambda_1 \prec \lambda_2 \prec \cdots \prec \lambda_r\}$$

with $\prec$ the ordering according to increasing total degree and with the lexicographic order when the degrees are equal (hence $\lambda_1 = 0$). (The reader may check that the argument of Lemma 2.7 remains unchanged if the pure lexicographic order were chosen here.) This orders the polynomials $p_j := p_{\lambda_j}$ in a way that is consistent with (i) and (ii). Then $q_j = p_j$ and

$$\sigma_j(x) := (e^{2\pi i x} - 1)^{\lambda_j}.$$

The smoothness requirements on the functions ψ_j are that $\psi_j \in C^\kappa(\mathbb{R}^d)$ with $\kappa = \max\{|\lambda|; \lambda \in \Lambda\} = |\lambda_r|$. Let $\phi_j := \sigma_j \psi_j$; the symbol matrix (1.3) has the entries

$$A_{k,j}(\xi) = \sum_{\alpha \in \mathbb{Z}^d} (D^{\lambda_k} \phi_j)(\alpha) e^{i\alpha \cdot \xi}, \qquad k, j = 1, \ldots, r.$$

The simplest situation is to have a compactly supported function ϕ with nonvanishing symbol $\widetilde{\phi}$ and to take all the ψ_j equal to ϕ. When ϕ is taken as a centered box spline, then we can compare the present method to the approach taken in [14]. In [14], the functions ϕ_j were the derivatives $p_j(D)M$ of a fixed box spline M, where the box spline is required to have linear independent integer translates and a direction set with only even multiplicities (which implies that $\widetilde{M}$ does not vanish on $\mathbb{R}^d$). But even then the symbol matrix had singularities, and it took effort to show that for the expected solution (essentially from Cramer's Rule) the singularities cancelled sufficiently to give bounded fundamental solutions in $L_2(\mathbb{R}^d)$. As far as computing the solutions from [14], one could use the fft to obtain the coefficients but care must be taken because of the singularities (see the methods used in [8]). With the method here, we may take any centered box spline for which cardinal interpolation is correct regardless of multiplicities. Moreover, the box splines here only need smoothness κ as

opposed to at least 2κ in [14]. The latter means that the complexity of computing the trigonometric polynomials comprising the rational functions $C_{m,j}$ is reduced, and since these rational functions now have no singularities, the fft techniques easily generate a sufficient number of coefficients to handle practical problems. The relevant facts about box splines can be found in [2]. Here in Example 2.4 is a concrete bivariate example.

Example 2.4. Consider first-order Hermite interpolation with the centered box spline $M^c_{2,2,1} \in C^1(\mathbb{R}^2)$. Here the functionals are point evaluation and the two first-order partial derivatives and the functions ϕ_j are given by

$$\phi_1(x(1), x(2)) = \phi(x(1), x(2)) = M^c_{2,2,1}(x(1), x(2)),$$

$$\phi_2(x(1), x(2)) = e^{2\pi i x(1)} - 1)\phi(x(1), x(2)),$$

$$\phi_3(x(1), x(2)) = (e^{2\pi i x(2)} - 1)\phi(x(1), x(2)),$$

The nonzero entries of the symbol matrix are given by

$$A_{1,1}(\xi(1), \xi(2)) = \widetilde{\phi}(\xi(1), \xi(2))$$

$$= \frac{1}{24}\{14 + 4\cos\xi(1) + 4\cos\xi(2) + 2\cos(\xi(1) + \xi(2))\},$$

$$A_{2,2}(\xi(1), \xi(2)) = A_{3,3}(\xi(1), \xi(2)) = 2\pi i\widetilde{\phi}(\xi(1), \xi(2)),$$

$$A_{2,1}(\xi(1), \xi(2)) = A_{3,1}(\xi(2), \xi(1))$$

$$= \frac{i}{4}\{-3\sin\xi(1) + \sin\xi(2) - \sin(\xi(1) + \xi(2))\}.$$

The nontrivial rational functions $C_{m,j}$ are simply

$$C_{1,1} = 2\pi i C_{2,2} = 2\pi i C_{3,3} = 1/A_{1,1} \quad C_{2,1} = \frac{-A_{2,1}}{A_{1,1}A_{2,2}}, \quad C_{3,1} = \frac{-A_{3,1}}{A_{1,1}A_{3,3}}.$$

All the denominators are powers of the complex symbol

$$\widetilde{\phi}(z, w) = (14 + 2(z + z^{-1}) + 2(w + w^{-1}) + zw + z^{-1}w^{-1})/24,$$

which vanishes when $z = -w = \rho = .2789\ldots$ and vanishes nowhere in $\rho < |z|, |w| < 1/\rho$ with ρ the solution of $\rho + 1/\rho = \sqrt{2}(1 + \sqrt{3})$. The latter provides information on the decay of the coefficients for the $C_{m,j}$; namely, $|c_{m,j}(\alpha)| = O((\rho + \epsilon)^{|\alpha|})$ for any $\epsilon > 0$ (cf. [2, Chap. 4]).

Example: Hermite interpolation to powers of the Laplacian. The powers of the Laplacian, $\Delta = \sum_{m=1}^{d} D_m^2$, give rise to the functionals

$$\lambda_j : f \mapsto \Delta^{j-1} f(0), \qquad j = 1, \ldots, r,$$

and the corresponding polynomials

$$p_j(x) = \left(\sum_{m=1}^{d} x(m)^2\right)^j.$$

Clearly, the ordering of the polynomials is consistent with (i) and (ii). For a given function $\phi \in C^{2r-2}(\mathbb{R}^d)$ we put

$$\phi_j = \sigma^{j-1}\phi \quad \text{with } \sigma(x(1),\dots,x(d)) = \sum_{m=1}^{d} 2(1-\cos 2\pi x(m)).$$

Note that σ is the symbol of the discrete Laplacian $(-\nabla)$ (with step size 2π). Here we have chosen to take $q_j = p_j$ and to replace $(e^{2\pi i x} - 1)$ by $\sin(\pi x)$. The following lemma replaces Lemma 2.2 for this choice of p_2, ϕ_j (under suitable assumptions on ϕ), and σ_j:

LEMMA 2.5. *With the above notation, we have for* $k, j = 1,\dots,r$,

$$(\Delta^{k-1}\phi_j)(\alpha) = 0 \quad \textit{for all } \alpha \in \mathbb{Z}^d$$

in case $k < j$, *and*

$$(\Delta^{k-1}\phi_k)(\alpha) = c_k\phi(\alpha), \qquad \alpha \in \mathbb{Z}^d.$$

with $c_k = \Delta^{k-1}\sigma^{k-1}(0)$.

Proof. The proof follows the lines of Lemma 2.2 and will be omitted. □

Remark. The constants c_k are given by

$$c_k = (8\pi^2)^{k-1} \sum_{|\alpha|=k-1} \left(\frac{(k-1)!}{\alpha(1)!\cdots\alpha(d)!}\right)^2,$$

and in particular

$$c_k = (8\pi^2)^{k-1}\binom{2(k-1)}{k-1} \quad \text{in case } d = 2.$$

Example 2.6. Taking $\phi_1(x,y) = \phi(x,y) = M^c_{2,2,2}(x,y)$, a centered box spline in $C^2(\mathbb{R}^2)$ together with the functionals

$$\lambda_1 : f \mapsto f(0,0) \quad \text{and} \quad \lambda_2 : f \mapsto (D_1^2 + D_2^2)f(0,0),$$

we have $\phi_2(x,y) = 2(2 - \cos 2\pi x - \cos 2\pi y)\phi(x,y)$, and the nonzero entries of the symbol matrix are

$$A_{1,1}(\xi(1),\xi(2)) = \widetilde{\phi}(\xi(1),\xi(2))$$

$$= \{3 + \cos\xi(1) + \cos\xi(2) + \cos(\xi(1)+\xi(2))\}/6$$

$$A_{2,2}(\xi(1),\xi(2)) = 16\pi^2\widetilde{\phi}(\xi(1),\xi(2))$$

$$A_{2,1}(\xi(1),\xi(2)) = -4 + 2(\cos\xi(1) + \cos\xi(2)).$$

The nontrivial rational functions $C_{m,j}$ are

$$C_{1,1} = 16\pi^2 C_{2,2} = \frac{1}{A_{1,1}} \quad \text{and} \quad C_{2,1} = -A_{2,1}/(A_{1,1}A_{2,2}).$$

Here the complex symbol $\widetilde{\phi}(z,w) = (6 + z + z^{-1} + w + w^{-1} + zw + z^{-1}w^{-1})/12$ vanishes when $z = -w = \rho = .4354\dots$ and nowhere in $\rho < |w|, |z| < 1/\rho$, where $\rho + 1/\rho = 1 + \sqrt{3}$. In particular, the decay rate of the coefficients in this example is not as fast as in Example 2.4.

3. Approximation orders from Hermite interpolants. In this section we observe that Hermite type interpolation can be carried out using our scheme without loss of approximation order over that obtained by cardinal interpolation from the space generated by the function $\phi_1 = \phi$:

$$\mathcal{S}_1 := \{\phi *' c \,|\, c : \mathbb{Z}^d \to \mathbb{C}\}.$$

For this study we make the following assumptions:

(a) $p_1 = 1, p_2, \dots, p_r$ are linearly independent;

(b) $\phi_j = \sigma_j \phi$, where $\phi \in C^\kappa, \kappa = \max_{1 \le j \le r} n_j$, has a nonvanishing symbol, $\widetilde{\phi}(\xi) \neq 0$ for all $\xi \in \mathbb{R}^d$, and satisfies the Strang–Fix conditions of order $s \le \kappa$,

$$\widehat{\phi}(0) = 1, \quad \text{and} \quad D^\beta \widehat{\phi}(2\pi\alpha) = 0 \quad \text{for } |\beta| < s, \quad \alpha \in \mathbb{Z}^d \backslash 0.$$

Under assumption (b) it is well known (see, e.g., [5] and [1]) that for all sufficiently smooth functions the cardinal interpolation operator $\mathcal{L}$ for a compactly supported function ϕ and the space $\mathcal{S}_1$ provide optimal approximation order: if

$$\mathcal{L}_h := \Sigma_{1/h} \mathcal{L} \Sigma_h, \qquad \Sigma_h : f \to f(h \cdot),$$

then for $f \in C^s(\mathbb{R}^d)$ with support in the compact set Ω,

$$\|f - \mathcal{L}_h f\|_{\infty,\Omega} = O(h^s). \tag{3.1}$$

We want a similar result for cardinal Hermite interpolation as discussed in §2, and therefore we would like to admit noncompactly supported functions as well. In order to ensure that the cardinal interpolation operator is well defined on polynomials of sufficient order, it will be necessary to impose some decay as $|x| \to \infty$;

(c) For the given κ and s, there is an $\epsilon, 0 < \epsilon < 1$, and a constant (depending on ϕ, κ and s) such that

$$|D^\beta \phi(x)| \le \text{ const } (1 + |x|)^{-d-s-\epsilon} \quad \text{for all} \quad |\beta| \le \kappa, \quad \text{and} \quad x \in \mathbb{R}^d.$$

Functions of this type have recently played a role in approximation order questions in [11] and [9], and in a certain sense, the result for cardinal interpolation is a special case of the results. We shall make use of the results detailed there, together with a generalization of Wiener's lemma [10] to sketch a proof for cardinal interpolation. The inequality in (c) implies that the symbol $\widetilde{\phi}$ belongs to $F_{d+s+\epsilon}$, where

$$F_\ell = \left\{ g = \sum_\alpha c(\alpha) \exp(-i\alpha \cdot) \in \mathcal{A} : |c(\alpha)| = O(|\alpha|^{-\ell}) \right\}.$$

By an extension of Wiener's lemma [10], $(1°)|g| > 0$ and $g \in F_\ell$ imply $1/g \in F_\ell$, and since $(2°) g_1, g_2 \in F_\ell$ implies $g_1 + g_2, g_1 g_2 \in F_\ell$, we have by Corollary 2.11 that $C_{m,j} \in F_{d+s+\epsilon}$ when ϕ satisfies (c). It therefore follows that

$$\sup |(\phi *' c_{m,j})(x)|(1 + |x|)^{d+s+\epsilon} < \infty \quad \text{for all } 1 \le m, j \le r. \tag{3.2}$$

Consequently, $\phi^\star : f \mapsto \phi *' f_|$ is a well defined map on Π_{s-1} to itself, $1 - \phi^\star$ is degree-reducing on Π_{s-1}, and there is a finitely supported sequence b for which the map $T : f \mapsto (\phi *' b) *' f$ is the identity on Π_{s-1} [9, §2].

The convergence result for cardinal interpolation for functions ϕ satisfying (a), (b), and (c) can now be proven, for example, by the argument of [1, Thm. 3] when the mapping β used there is replaced by the mapping T (see also, the arguments in [11]).

Define the scaled cardinal Hermite interpolation operator $\mathcal{J}_h$ by

$$\mathcal{J}_h := \Sigma_{1/h}\mathcal{J}\Sigma_h, \qquad \Sigma_h : f \to (h\cdot), \tag{3.3}$$

with $\mathcal{J}$ as in (2.11). For the approximation properties of the cardinal Hermite interpolation operator, we have the following.

THEOREM 3.1. *If $p_1, \dots, p_r$ and ϕ satisfy the assumptions* (a), (b), *and* (c) *above, then for all $f \in W_\infty^{\kappa+s}(\mathbb{R})^d$ with support in an arbitrary compact set Ω, the corresponding cardinal Hermite interpolation operator $\mathcal{J}_h$ satisfies*

$$\|f - \mathcal{J}_h f\|_{\infty,\Omega} = O(h^s).$$

Proof. Since ϕ satisfies (b), the fundamental functions L_j are uniquely defined with coefficient sequences $c_{m,j} \in \mathcal{A}$. Now observe that each L_j also satisfies (c) (cf. [11]). Under the assumptions interpolation is unique so that $\mathcal{J}(\mathcal{L}f) = \mathcal{L}f$ since $\mathcal{S}_1 \subset \mathcal{S}$. In particular $\mathcal{J}p = p$ for all polynomials of degree $s-1$ since $\mathcal{L}p = p$, $p \in \Pi_{s-1} \subset \mathcal{S}_1$, when ϕ satisfies the Strang–Fix conditions of order s (cf. [9]). Therefore, for the operators

$$\mathcal{R} := \mathcal{J} - \mathcal{L}, \qquad \mathcal{R}_h := \mathcal{J}_h - \mathcal{L}_h,$$

we have

$$\mathcal{R}_h p = 0, \qquad p \in \Pi_{s-1}. \tag{3.4}$$

Hence,

$$\|f - \mathcal{J}_h f\|_{\infty,\Omega} \le \|f - \mathcal{L}_h f\|_{\infty,\Omega} + \|\mathcal{R}_h f\|_{\infty,\Omega},$$

and it is sufficient to show that

$$\|\mathcal{R}_h f\|_{\infty,\Omega} = O(h^s).$$

We first obtain a representation for $\mathcal{R}f$. From Corollary 2.3 we have

$$C_{1,1} = 1/A_{1,1} = 1/\widetilde{\phi}$$

and

$$L_1 = \phi *' c_{1,1} + \sum_{m=2}^{r} \phi_m *' c_{m,1}.$$

But it is well known that $L_0 := \phi *' c_{1,1}$ with this choice of coefficients is the fundamental solution for cardinal interpolation from $\mathcal{S}_1$. Thus, from (2.11) and Corollary 2.3,

$$\begin{aligned}\mathcal{J}f &= \sum_{j=1}^{r} \left(\sum_{m=j}^{r} \phi_m *' c_{m,j} \right) *' (p_j(D)f)_| \\ &= \mathcal{L}f + \sum_{j=1}^{r} \left(\sum_{\substack{m=j \\ m\neq 1}}^{r} \phi_m *' c_{m,j} \right) *' (p_j(D)f)_|.\end{aligned}$$

From this we obtain

$$\begin{aligned}\mathcal{R}f &= \sum_{j=1}^{r}\left(\sum_{\substack{m=j\\ m\neq 1}}^{r}\phi_m *' c_{m,j}\right) *' (p_j(D)f)_{|}\\ &= \sum_{j=1}^{r}\sum_{\substack{m=j\\ m\neq 1}}^{r}\sigma_m\left(\phi *' c_{m,j} *' (p_j(D)f)_{|}\right),\end{aligned} \tag{3.5}$$

with the last equality obtained by the shift invariance of σ_j.

Let g be sufficiently smooth and P_w be the Taylor polynomial of total degree $s-1$ for g at w. Then the remainder formula

$$g(x) - P_w(x) = \sum_{|\alpha|=s}\frac{(x-w)^\alpha}{\alpha!}\int_0^1 t^{s-1}D^\alpha g(x+t(w-x))dt$$

yields the estimate

$$|p_j(D)(g-P_w)(\beta)| \leq \text{ const } \max_{\substack{\eta\leq\max(\alpha,\kappa)\\ |\alpha|=s}}\{|(\beta-w)^{\alpha-\eta}|\}\max_{\substack{|\gamma|\leq n_j\\ |\alpha|=s}}\|D^{\alpha+\gamma}g\|_\infty \tag{3.6}$$

for some constant depending only on p_j.

The estimate on the approximation order can now be completed as follows. For $x\in\Omega$, we take $g=\Sigma_h f$ and $w=x/h$ and estimate

$$\mathcal{R}_h f(x) = \mathcal{R}g(x/h) = \mathcal{R}(g-P_w)(x/h).$$

By (c), the fact that the coefficient sequences $c_{m,j}$ belong to $F_{d+s+\epsilon}$ and (3.6), we have

$$|\phi *' c_{m,j} *' (p_j(D)(g-P_w))_{|}(x/h)| = \left|\sum_{\alpha,\beta}\phi(x/h-\alpha)c_{m,j}(\alpha-\beta)(p_j(D)(g-P_w))(\beta)\right|$$

$$\leq \text{ const}_1\sum_{\alpha,\beta}(1+|x/h-\alpha|)^{-d-s-\epsilon}(1+|\alpha-\beta|)^{-d-s-\epsilon}|\beta-x/h|^s\max_{\substack{|\gamma|\leq n_j\\ |l|=s}}\|D^{l+\gamma}g\|_\infty$$

$$\leq \text{ const}_3\, h^s\max_{\substack{|\gamma|\leq n_j\\ |l|=s}}\|D^{l+\gamma}f\|_\infty.$$

The estimate now follows from the representation (3.5). □

Example 3.2. As a last example we note that our results seem new even in the univariate setting. To illustrate this, we interpolate the function values and those of the second derivative for the function

$$F(x) = \begin{cases} -(x^2-9)^4(e^{-.3(x+1.2)(x-1.6)} - 2e^{-.10(x-.5)^2})/300 & \text{if } -3\leq x\leq 3,\\ 0 & \text{otherwise,}\end{cases}$$

using $\phi_1=\phi$, the centered cubic cardinal B-spline, and $\phi_2=(\exp(2\pi i\cdot)-1)^2\phi$. These two functions are plotted in Fig. 1. In this case the matrix $A(\xi)$ takes the form

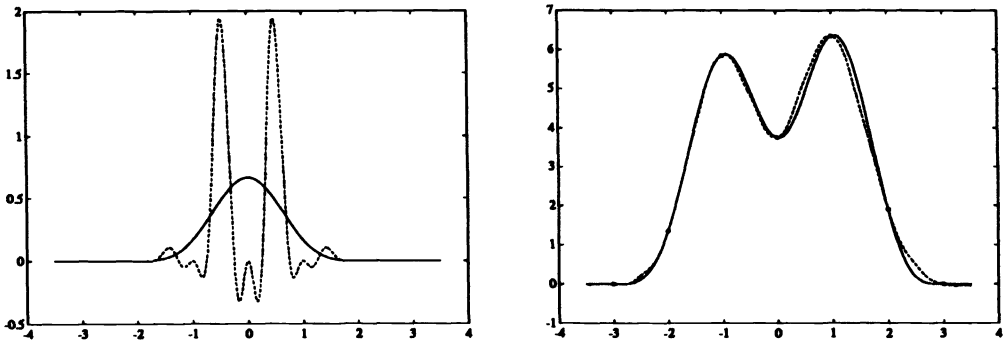


FIG. 1(a). *ϕ_1 and ϕ_2 for the cubic spline. The function F with its cardinal interpolant ($h = 1$).*

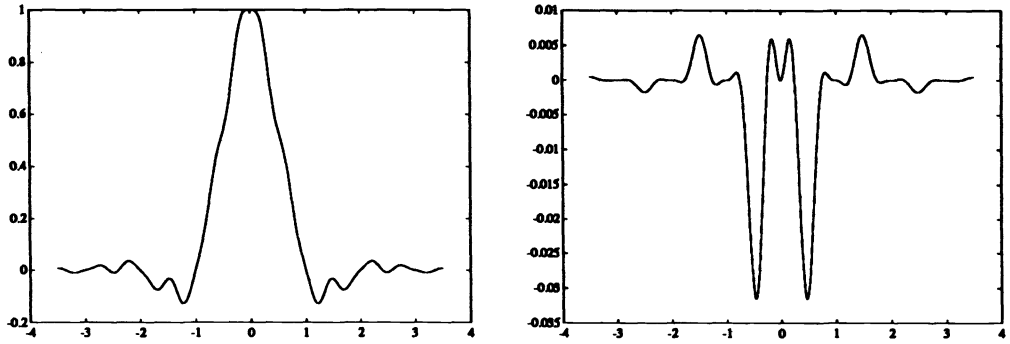


FIG. 1(b). *The fundamental functions L_1 and L_2 for interpolation of F and F''.*

$$A(\xi) = \begin{bmatrix} (2+\cos(\xi))/3 & 0 \\ -2+2\cos(\xi) & 2(2\pi i)^2(2+\cos(\xi))/3 \end{bmatrix}.$$

Consequently,

$$C_{1,1}(\xi) = \frac{3}{2+\cos(\xi)}, \quad C_{2,1}(\xi) = \frac{9(\cos(\xi)-1)}{4\pi^2(2+\cos(\xi))^2}, \quad C_{2,2} = \frac{3}{8\pi^2(2+\cos(\xi))}.$$

The fundamental functions are also shown in Fig. 1. Finally, the error in approximating F on $[-3.5, 3.5]$ as estimated on a step size .01 using $h = 1, .5, .25, .125$, respectively, rounded in 6 places is

$$[0.470411 \quad 0.053301 \quad 0.003088 \quad 0.000185],$$

which already illustrates $O(h^4)$ convergence. Only the cardinal interpolant is shown in Fig. 1, the scale not permitting the other approximations to be distinguished from the curve. (Only the real parts are shown in the figure, the imaginary parts are neglected although they are included in calculation of the error.)

Acknowledgment. An essential part of this work was completed while Kurt Jetter was visiting the University of Alberta, Edmonton, Canada, in August 1991. The support from the members of the approximation theory group there is gratefully acknowledged. We thank the two referees for comments that helped us improve the paper.

REFERENCES

[1] C. DE BOOR, K. HÖLLIG, AND S. D. RIEMENSCHNEIDER, *Fundamental solutions for multivariate difference equations*, Amer. J. Math., 111 (1989), pp. 403–415.

[2] ———, *Box Splines*, Springer-Verlag, New York, 1993.

[3] P. BINEV, AND K. JETTER, *Cardinal interpolation with shifted 3-directional box splines*, Proc. Roy. Soc. Edinburgh, 122A (1992), pp. 205–220.

[4] M. BUHMANN, AND C. A. MICCHELLI, *On radial basis approximation on periodic grids*, Proc. Cambridge Philos. Soc., 112 (1992), pp. 317–334.

[5] C. K. CHUI, K. JETTER, AND J. D. WARD, *Cardinal interpolation with multivariate splines*, Math. Comp., 48 (1987), pp. 711–724.

[6] ———, *Cardinal interpolation with differences of tempered functions*, Comput. Math. Appl., 24 (1992), pp. 35–48.

[7] I. DAUBECHIES, S. JAFFARD, AND J. L. JOURNÉ, *A simple Wilson orthonormal basis with exponential decay*, SIAM J. Math. Anal., 22 (1991), pp. 554–573.

[8] K. JETTER, AND J. STÖCKLER, *Algorithms for cardinal interpolation using box splines and radial basis functions*, Numer. Math., 60 (1991), pp. 97–114.

[9] R.-Q. JIA AND JUNJIANG LEI, *Approximation by multiinteger translates of functions having global support*, J. Approx. Theory, 72 (1993), pp. 2–23.

[10] J. LEI, *L_p-approximation by certain projection operators*, preprint.

[11] W. A. LIGHT AND E. W. CHENEY, *The Strang–Fix theory for functions having non-compact support*, Constr. Approx., 8 (1992), pp. 35–48.

[12] M. J. D. POWELL, *The theory of radial basis function approximation in* 1990, in Advances in Numerical Analysis, Vol. II: Wavelets, Subdivision Algorithms, and Radial Basis Functions, W. Light, ed., Oxford University Press, Oxford, 1992, pp. 105–210.

[13] S. D. RIEMENSCHNEIDER, *Multivariate cardinal interpolation*, in Approximation Theory VI, C. Chui, L. Schumaker, and J. Ward, eds., Academic Press, New York, pp. 561–580, 1989.

[14] S. D. RIEMENSCHNEIDER AND K. SCHERER, *Cardinal Hermite interpolation with box splines* I, II, Constr. Approx., 3 (1987), pp. 223–238; Numer. Math., 58 (1991), pp. 591–602.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 976–990, May 1994

010

CHARACTERIZATIONS OF ORTHOGONAL POLYNOMIALS SATISFYING DIFFERENTIAL EQUATIONS*

K. H. KWON†, L. L. LITTLEJOHN‡ AND B. H. YOO†

Abstract. In 1938, H. L. Krall found a necessary and sufficient condition for an orthogonal polynomial set $\{P_n(x)\}_0^\infty$ to satisfy a linear differential equation of the form

$$\sum_{i=0}^{N} \ell_i(x) y^{(i)}(x) = \lambda_n y(x).$$

Here the authors give a new simple proof of Krall's theorem as well as some other characterizations of such orthogonal polynomial sets based on the symmetrizability of the differential operator. In particular it is shown that such orthogonal polynomial sets are characterized by a certain Sobolev-type orthogonality, which generalizes Hahn's charaterization of classical orthogonal polynomials.

Key words. orthogonal polynomials, differential equations, Sobolev-type orthogonality

AMS subject classification. 33C45

1. Introduction. Let us consider a linear differential equation of order $N \geq 1$ of the form

$$\sum_{i=0}^{N} \ell_i(x) y^{(i)}(x) = \lambda y(x), \tag{1.1}$$

where $\ell_i(x)$, $i = 0, \ldots, N$, are real-valued smooth functions on the real line with $\ell_N(x) \not\equiv 0$ and λ is a real parameter, and ask this question: When does the differential equation (1.1) have an orthogonal polynomial set as solutions?

It is easy to see that if the differential equation (1.1) has polynomial solutions $P_n(x)$ of degree n for $n = 0, 1, \ldots, N$, then it must be of the form

$$L_N(y) = \sum_{i=0}^{N} \ell_i(x) y^{(i)}(x) = \sum_{i=0}^{N} \sum_{j=0}^{i} \ell_{ij} x^j y^{(i)}(x) = \lambda_n y(x), \tag{1.2}$$

where ℓ_{ij} are real constants and the eigenvalue parameter is given by

$$\lambda_n = \ell_{00} + \ell_{11} n + \cdots + \ell_{NN} n(n-1) \cdots (n-N+1). \tag{1.3}$$

In 1929, Bochner [2] (see also Krall and Frink [15]) proved that there are essentially (that is, up to a linear change of variable) only four distinct orthogonal polynomial sets satisfying the differential equation (1.2) for $N = 2$. They are now called the classical orthogonal polynomials of Jacobi, Laguerre, Hermite, and Bessel. He also implicitly imposed the problem of classifying all orthogonal polynomials satisfying the differential equation (1.2) for $N > 2$.

The classifying problem itself is not resolved yet in general except for $N = 2$ (due to Bochner [2]) and for $N = 4$ (due to Krall [13]). However, Krall [12] found

* Received by the editors August 29, 1992; accepted for publication (in revised form) April 19, 1993.

† Department of Mathematics, Korea Advanced Institute of Science and Technology, Taejon, Korea. The work of first author is partially supported by The Korea Science and Engineering Foundation (90-08-00-02) and the Global Analysis Research Center.

‡ Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322-3900.

a remarkable theorem (cf. Theorem 2.1) characterizing all differential equations of the form (1.2) which have an orthogonal polynomial set as solutions. Its proof in [12] is based on the notion of a dual equation to the differential equation (1.2), which is developed by Sheffer [25]. Later a second, simpler proof using the generating functions of orthogonal polynomials was found by Krall and Sheffer [16].

In this paper, we present a third proof of Krall's theorem as well as some other equivalent characterizations. The advantage that this third proof has over the other two proofs lies in its simplicity. Furthermore, this new proof makes use of the symmetry equations associated with the differential expression $L_N[\cdot]$ defined in (1.2); see Remark 1 following Lemma 2.2 below as well as the discussion in §4. These symmetry equations and their importance to differential equations and to the theory of orthogonal polynomials was first observed by Littlejohn in [21] and [22]. Indeed, if these equations have a simultaneous nontrivial distributional solution, it is an orthogonalizing weight functional for the corresponding sequence of polynomial solutions to (1.2). This constructive technique was recently successfully used [17] to produce the first known example of an orthogonalizing signed measure of bounded variation for the Bessel polynomials.

This new proof is based on the fact that if the differential equation (1.2) has orthogonal polynomial solutions, then it must be *symmetrizable on polynomials*; see §2 for further details.

As a consequence of this new proof, we shall show that an orthogonal polynomial set satisfies the differential equation (1.2) if and only if it has a certain Sobolev-type orthogonality (see Theorem 3.2 below). Much work is currently being done on the general theoretic properties of polynomials that are orthogonal with respect to some Sobolev inner product. Our Theorem 3.2 generalizes the well-known Sonine–Hahn characterization (see [6] and [26]) of classical orthogonal polynomials which can be restated as follows.

THEOREM 1.1. *Suppose $\phi_n(x)$ is a real polynomial of degree n $(n = 0, 1, \ldots)$ and $\{\phi_n(x)\}_{n=0}^{\infty}$ are simultaneously orthogonal with respect to the two bilinear quadratic forms defined by*: (i) $(p, q)_0 := \int_R p(x)q(x)d\mu_0(x)$,

(ii) $(p, q)_1 := (p, q)_0 + \int_R p'(x)q'(x)d\mu_1(x)$,

where μ_0 and μ_1 are (signed) measures on the Borel sets of the real line R, each having finite moments of all orders. Then, up to a linear change of variable, $\{\phi_n(x)\}_{n=0}^{\infty}$ is one of the following systems:

(a) *Jacobi polynomials,*

(b) *Laguerre polynomials,*

(c) *Hermite polynomials,*

(d) *Bessel polynomials.*

We believe that our proof sheds new light on the important problem of identifying orthogonal polynomial solutions of the equation (1.2) as eigenfunctions of an operator which is self-adjoint in some Hilbert (when the symmetry equations produce a positive-definite weight functional) or Krein space (when the symmetry equations produce a quasi-definite weight functional, as in the case of the Bessel polynomials). The reader is referred to the contributions [5],[8],[9],[11], and [18].

This work continues the contribution [19] in which the case $N = 4$ is discussed.

2. New proof of Krall's theorem. All polynomials in the following are assumed to be real polynomials in one variable and we let $\mathcal{P}$ denote the space of all real polynomials. We call any linear functional σ on $\mathcal{P}$ a moment functional and

$$\sigma_n := \langle \sigma, x^n \rangle, \qquad n = 0, 1, \ldots \tag{2.1}$$

the moments of σ. We denote the degree of a polynomial $\phi(x)$ by $\deg\phi$ with the convention $\deg 0 = -1$. By a polynomial set, we mean a sequence of polynomials $\{\phi_n(x)\}_0^\infty$ with $\deg\phi_n = n$, $n = 0, 1, \cdots$. Any polynomial set $\{\phi_n(x)\}_0^\infty$ determines a moment functional σ, called a canonical moment functional of $\{\phi_n(x)\}_0^\infty$, by the conditions

$$\langle\sigma, \phi_0\rangle \neq 0 \quad \text{and} \quad \langle\sigma, \phi_n\rangle = 0, \qquad n = 1, 2, \ldots. \tag{2.2}$$

Note that a canonical moment functional of a polynomial set $\{\phi_n(x)\}_0^\infty$ is uniquely determined by $\{\phi_n(x)\}_0^\infty$ up to a nonzero constant multiple.

DEFINITION 2.1. *A polynomial set $\{P_n(x)\}_0^\infty$ is called an orthogonal polynomial set* (OPS *in short*) *if there is a moment functional σ such that*

$$\langle\sigma, P_m(x)P_n(x)\rangle = K_n\delta_{mn}, \qquad m \text{ and } n = 0, 1, \ldots, \tag{2.3}$$

where K_n are nonzero real constants. In this case, we call $\{P_n(x)\}_0^\infty$ an OPS relative to σ.

Note that if $\{P_n(x)\}_0^\infty$ is an OPS relative to σ, then σ must be a canonical moment functional of $\{P_n(x)\}_0^\infty$.

The main goal of this section is to provide a new, simple, and illuminating proof of the following theorem due to Krall [12] as well as some other equivalent formulations.

THEOREM 2.1. *Let $\{P_n(x)\}_0^\infty$ be a polynomial set and $\{\sigma_n\}_0^\infty$ the moments of any canonical moment functional σ of $\{P_n(x)\}_0^\infty$. Then $\{P_n(x)\}_0^\infty$ is an* OPS *satisfying the differential equation* (1.2) *for each $n = 0, 1, \ldots$ if and only if $\{\sigma_n\}_0^\infty$ satisfy*

(i) $\Delta_n := \det[\sigma_{i+j}]_{i,j=0}^n \neq 0$, $n = 0, 1, \ldots$,

(ii) $S_k(m) := \sum_{i=2k+1}^N \sum_{j=0}^i \binom{i-k-1}{k} P(m-2k-1, i-2k-1)\ell_{i,i-j}\sigma_{m-j} = 0$ *for $k = 0, 1, \ldots, [\frac{N-1}{2}]$ and $m = 2k+1, 2k+2, \ldots$, where $[x]$ is the integer part of a real number x and*

$$P(n,k) = \begin{cases} 0, & n = 0 \\ n(n-1)\cdots(n-k+1), & n = 1, 2, \ldots. \end{cases}$$

Furthermore, N must be even, say, $N = 2r$ for some $r \in \{1, 2, \ldots\}$.

We begin by recalling a few well-known facts on the symmetrizability of linear differential operators of the form

$$L := L(x, D) = \sum_0^N a_i(x)D^i, \tag{2.4}$$

where $D = d/dx$, $a_i(x)$ are real-valued functions in $C^i(I)$, $a_N(x) \not\equiv 0$, and I is an open interval on the real line. The formal adjoint of L is a differential operator L^* defined by

$$L^*(y) = \sum_0^N (-1)^i (a_i y)^{(i)}, \qquad y(x) \text{ in } C^N(I). \tag{2.5}$$

The operator L is called symmetric if $L = L^*$. It is called symmetrizable if there is a real-valued function $s(x) \not\equiv 0$ in $C^N(I)$ such that sL is symmetric. Then we call $s(x)$ a symmetry factor of L.

It follows easily from the above definition that (i) any symmetric differential operator must be of even order.

(ii) The sum of any two symmetric differential operators is also symmetric.

With these two elementary facts (see [4] and [14]) we have that the most general symmetric differential operator of order $N = 2r$ must be of the form

$$L(y) = \sum_{0}^{r} (-1)^i (f_i y^{(i)})^{(i)}$$

or

$$L(y) = \sum_{i=0}^{r} f_i y^{(2i)} + \sum_{i=1}^{r} \sum_{j=0}^{i-1} \binom{2i}{2j+1} \frac{2^{2i-2j}-1}{i-j} B_{2i-2j} f_i^{(2i-2j-1)} y^{(2j+1)},$$

where B_{2i} are the Bernoulli numbers defined by

$$\frac{x}{e^x - 1} = 1 - \frac{x}{2} + \sum_{i=1}^{\infty} \frac{B_{2i}}{(2i)!} x^{2i}.$$

Now a symmetry factor of the differential operator L in (2.4) can be characterized as in the next lemma.

LEMMA 2.2 [21],[23]. *Let the Nth-order differential expression $L[\cdot]$ be as defined in* (2.4) *with the previously mentioned conditions on the coefficients $a_i(x)$ $(i = 0, 1, \ldots, N; x \in I)$. For any real-valued function $s(x) \not\equiv 0$ in $C^N(I)$, the following are equivalent.*

(i) *$s(x)$ is a symmetry factor for $L[\cdot]$ on I; that is, $sL = (sL)^*$.*

(ii) *$s(x)$, $x \in I$, satisfies the $N + 1$ equations*

$$\sum_{i=k}^{N} (-1)^i \binom{i}{k} (a_i s)^{(i-k)} = a_k s, \qquad k = 0, 1, \ldots, N. \tag{2.6}$$

(iii) *$s(x)$ satisfies the $r := [\frac{N+1}{2}]$ equations*

$$R_k(s) := \sum_{\ell=k}^{r} \sum_{j=0}^{2\ell-2k+1} \binom{2\ell}{2k-1} \binom{2\ell-2k+1}{j} \frac{2^{2\ell-2k+2}-1}{\ell-k+1} \cdot B_{2\ell-2k+2} a_{2\ell}^{(2\ell-2k+1-j)} s^{(j)} - a_{2k-1} s = 0, \qquad k = 1, 2, \ldots, r. \tag{2.7}$$

(iv) *$s(x)$ satisfies the $r := [\frac{N+1}{2}]$ equations*

$$\widetilde{R}_k(s) := \sum_{i=2k+1}^{N} (-1)^i \binom{i-k-1}{k} (a_i s)^{(i-2k-1)} = 0, \qquad k = 0, 1, \ldots, r-1. \tag{2.8}$$

(v) *There are $r + 1$ real-valued functions $f_i(x)$ in $C^{2i}(I)$, $i = 0, 1, \ldots, r = [\frac{N+1}{2}]$, with $f_r(x) \not\equiv 0$, and*

$$(sLy)(x) = \sum_{i=0}^{r} (-1)^i [f_i(x) y^{(i)}(x)]^{(i)}, \qquad y \in C^N(I). \tag{2.9}$$

Furthermore, if any of the above equivalent conditions holds, then $N = 2r$ must be even.

Remark 1. The equations $\widetilde{R}_k(s) = 0$ $(k = 0, 1, \ldots, r-1)$, given in (2.8), are called the symmetry equations associated with the expression $L[\cdot]$.

Remark 2. We may add another equivalent condition to those listed in Lemma 2.2. Indeed, this condition is as follows.

(vi) For any two real-valued functions $y(x)$ and $z(x)$ in $C^N(I)$, one of which has compact support in I, it is the case that

$$\langle sL[y], z\rangle := \int_I z(x)(sLy)(x)\,dx = \int_I y(x)(sLz)(x)\,dx := \langle y, sL[z]\rangle. \tag{2.10}$$

Indeed, this is quite often the standard definition of a formally symmetric differential expression.

We point out here that $\langle\cdot,\cdot\rangle$ is, in general, only a bilinear quadratic form and not always a positive-definite inner product. We shall continue to use this notation throughout this paper.

In order to adapt Lemma 2.2 to our situation we need the following simple formal calculus on moment functionals.

For any moment functional σ and any polynomial $\phi(x)$, we define two new moment functionals σ', the derivative of σ, and $\phi\sigma$, multiplication of σ by a polynomial $\phi(x)$, by

$$\langle\sigma', \psi(x)\rangle = -\langle\sigma, \psi'(x)\rangle, \tag{2.11}$$

$$\langle\phi\sigma, \psi(x)\rangle = \langle\sigma, \phi(x)\psi(x)\rangle \tag{2.12}$$

for ψ in $\mathcal{P}$. Then we have the Leibnitz rule:

$$(\phi\sigma)' = \phi'\sigma + \phi\sigma'. \tag{2.13}$$

Finally, we need the following simple fact.

LEMMA 2.3. *Let $\{P_n(x)\}_0^\infty$ be an* OPS *relative to a moment functional σ. Then we have that follows:* (i) *For any polynomial $\phi(x)$, $\phi\sigma \equiv 0$ if and only if $\phi(x) \equiv 0$.*

(ii) *For any other moment functional τ, $\langle\tau, P_n\rangle = 0$, $n \geq k+1$ for some integer $k \geq 0$ if and only if $\tau = \phi\sigma$ for some polynomial $\phi(x)$ of degree $\leq k$.*

Proof. (i) Assume that $\phi\sigma \equiv 0$ but $\phi(x) \not\equiv 0$. Write $\phi(x)$ as

$$\phi(x) = \sum_0^n c_j P_j(x),$$

where $n = \deg\phi(\geq 0)$ and c_j are constants with $c_n \neq 0$. Then by the orthogonality of $\{P_n(x)\}_0^\infty$ relative to σ, we have

$$0 = \langle\phi\sigma, P_n\rangle = \langle\sigma, \phi P_n\rangle = c_n\langle\sigma, P_n^2\rangle$$

and so $c_n = 0$, which is a contradiction. The converse is trivial.

(ii) Consider a moment functional $\tilde{\tau}$ given by

$$\tilde{\tau} = \left(\sum_0^k c_j P_j(x)\right)\sigma$$

where c_j are real constants to be determined. Then we have

$$\langle \tilde{\tau}, P_n \rangle = \sum_0^k c_j \langle \sigma, P_j P_n \rangle = \begin{cases} c_n \langle \sigma, P_n^2 \rangle, & n \le k, \\ 0, & n > k. \end{cases} \tag{2.14}$$

Assume $\langle \tau, P_n \rangle = 0$ for $n > k$. Then the equation (2.14) shows that if we take

$$c_j = \langle \tau, P_j \rangle \langle \sigma, P_j^2 \rangle^{-1}, \qquad j = 0, 1, \ldots, k,$$

then $\langle \tau, P_n \rangle = \langle \tilde{\tau}, P_n \rangle$ for all $n \ge 0$ so that $\tau = \tilde{\tau}$ since $\{P_n(x)\}_0^\infty$ is a polynomial set. Conversely, if $\tau = \phi\sigma$ for some polynomial $\phi(x)$ of degree $\le k$, then $\langle \tau, P_n \rangle = \langle \sigma, \phi P_n \rangle = 0$ for $n > k$. □

Now we are ready to give a new proof of Theorem 2.1. In fact, we shall prove the following which is equivalent to Theorem 2.1 and is of interest in itself.

THEOREM 2.4. *Let $\{P_n(x)\}_0^\infty$ be an* OPS, *σ a canonical moment functional of $\{P_n(x)\}_0^\infty$, and $\{\sigma_n\}_0^\infty$ the moments of σ. Then the following statements are all equivalent.*

(i) *For each $n = 0, 1, \ldots$, $P_n(x)$ satisfies the differential equation* (1.2).

(ii) *σL_N is symmetric on polynomials in the sense that*

$$\langle L_N(\phi)\sigma, \psi \rangle = \langle L_N(\psi)\sigma, \phi \rangle \tag{2.15}$$

for all polynomials $\phi(x)$ and $\psi(x)$.

(iii) *σ satisfies the $r := [\frac{N+1}{2}]$ equations (with R_k as in* (2.7))

$$R_k \sigma = 0, \qquad k = 1, 2, \ldots, r. \tag{2.16}$$

(iv) *σ satisfies the $r := [\frac{N+1}{2}]$ equations (with $\widetilde{R}_k$ as in* (2.8))

$$\widetilde{R}_k \sigma = 0, \qquad k = 0, 1, \ldots, r-1. \tag{2.17}$$

(v) *$\{\sigma_n\}_0^\infty$ satisfies the $r := [\frac{N+1}{2}]$ recurrence relations (with $S_k(m)$ as in Theorem* 2.1)

$$S_k(m) = 0, \quad k = 0, 1, \ldots, r-1 \quad \text{and} \quad m = 2k+1, 2k+2, \ldots.$$

(vi) *$\{\sigma_n\}_0^\infty$ satisfies the $r := [\frac{N+1}{2}]$ recurrence relations*

$$\begin{aligned} T_k(m) &:= \sum_{i=k}^{r} \sum_{j=0}^{2i} \binom{2i}{2k-1} P(m-2k+1, 2i-2k+1) \frac{2^{2i-2k+2}-1}{i-k+1}, \\ & B_{2i-2k+2} \ell_{2i,j} \sigma_{m-2i+j} + \sum_{j=0}^{2k-1} \ell_{2k-1,j} \sigma_{m-2k+1+j} = 0, \end{aligned} \tag{2.18}$$

$k = 1, 2, \ldots, r$ and $m = 2k-1, 2k, \ldots$.

Furthermore, if any of the above equivalent conditions holds, then $N = 2r$ must be even.

Proof. For all polynomials $\phi(x)$ and $\psi(x)$, we have from the Leibnitz rule (2.13) that

$$\begin{aligned}\langle L_N(\phi)\sigma,\psi\rangle &= \left\langle \sum_{i=0}^{N} \ell_i \phi^{(i)}\sigma,\psi \right\rangle = \left\langle \sum_{i=0}^{N}(-1)^i(\psi\ell_i\sigma)^{(i)},\phi \right\rangle \\ &= \left\langle \sum_{i=0}^{N}\sum_{k=0}^{i}(-1)^i\binom{i}{k}\psi^{(k)}(\ell_i\sigma)^{(i-k)},\phi \right\rangle \\ &= \left\langle \sum_{k=0}^{N}\sum_{i=k}^{N}(-1)^i\binom{i}{k}\psi^{(k)}(\ell_i\sigma)^{(i-k)},\phi \right\rangle.\end{aligned}$$

Hence, the condition (2.15) is equivalent to

$$\sum_{i=k}^{N}(-1)^i\binom{i}{k}(\ell_i\sigma)^{(i-k)} = \ell_k\sigma, \qquad k=0,1,\dots,N. \tag{2.19}$$

Therefore, the equivalences of the conditions (ii), (iii), and (iv) follow immediately from Lemma 2.2. Now, assume that the condition (ii) holds. Equivalently, it means that σ satisfies the $N+1$ equations in (2.19). Since $L_N(P_n) = \sum_0^N \ell_i P_n^{(i)}$ is a polynomial of degree $\le n$, we may write it as

$$L_N(P_n) = \sum_0^N \ell_i P_n^{(i)} = \sum_0^n c_j P_j,$$

where c_j are constants depending on n. Then for $k=0,1,\dots,n$, we have by (2.19)

$$\begin{aligned}c_k\langle\sigma,P_k^2\rangle &= \left\langle \sigma,\sum_{i=0}^{N}\ell_i P_n^{(i)}P_k \right\rangle = \sum_{i=0}^{N}(-1)^i\left\langle (P_k\ell_i\sigma)^{(i)},P_n \right\rangle \\ &= \sum_{j=0}^{N}\sum_{i=j}^{N}(-1)^i\binom{i}{j}\langle P_k^{(j)}(\ell_i\sigma)^{(i-j)},P_n\rangle = \sum_{j=0}^{N}\langle P_k^{(i)}\ell_j\sigma,P_n\rangle \\ &= \sum_{j=0}^{N}\langle\sigma,P_k^{(i)}\ell_j P_n\rangle = \begin{cases} 0 & \text{if } k<n, \\ \lambda_n\langle\sigma,P_n^2\rangle & \text{if } k=n. \end{cases}\end{aligned}$$

Hence, we have $c_k = 0$, $k<n$ and $c_n = \lambda_n$ so that $L_N(P_n) = \lambda_n P_n$.

Conversely, assume that the condition (i) holds. Multiplying $L_N P_n = \lambda_n P_n$ by P_k and applying σ we obtain

$$\begin{aligned}\left\langle \sigma,P_k\sum_0^N \ell_i P_n^{(i)} \right\rangle &= \left\langle \sum_0^N(-1)^i(P_k\ell_i\sigma)^{(i)},P_n \right\rangle \\ &= \lambda_n\langle\sigma,P_kP_n\rangle = \begin{cases} 0 & \text{if } k\neq n, \\ \lambda_n\langle\sigma,P_n^2\rangle & \text{if } k=n. \end{cases}\end{aligned} \tag{2.20}$$

If we set

$$v_k := \sum_0^N(-1)^i(P_k\ell_i\sigma)^{(i)},$$

then the equation (2.20) implies $\langle v_k, P_n\rangle = 0$ for $k > n$ so that by Lemma 2.3 we have

$$v_k = \sum_{j=0}^{k} \langle v_k, P_j\rangle \langle \sigma, P_j^2\rangle^{-1} P_j(x)\sigma = \lambda_k P_k(x)\sigma, \qquad k = 0, 1, \ldots. \tag{2.21}$$

On the other hand, we have

$$v_k = \sum_{i=0}^{N} (-1)^i (P_k \ell_i \sigma)^{(i)} = \sum_{j=0}^{N} P_k^{(j)} \sum_{i=j}^{N} (-1)^i \binom{i}{j} (\ell_i \sigma)^{(i-j)} = \sum_{j=0}^{N} P_k^{(j)} u_j \tag{2.22}$$

where

$$u_j := \sum_{i=j}^{N} (-1)^i \binom{i}{j} (\ell_i \sigma)^{(i-j)}, \qquad j = 0, 1, \ldots, N.$$

Hence, we have from (2.21) and (2.22) that

$$v_k = \lambda_k P_k(x)\sigma = \sum_{j=0}^{N} P_k^{(j)} u_j = \sum_{j=0}^{k} P_k^{(j)} u_j, \qquad k = 0, 1, \ldots, N. \tag{2.23}$$

Now we claim that $u_j = \ell_j(x)\sigma$, $j = 0, 1, \ldots, N$ so that the condition (2.19), i.e., (2.15) holds. For $j = 0$, $v_0 = \lambda_0 P_0(x)\sigma = P_0 u_0$ and so $u_0 = \lambda_0 \sigma = \ell_0 \sigma$. Assume that $u_j = \ell_j(x)\sigma$, $j = 0, 1, \ldots, k$ for some $k \le N - 1$. Then from (2.23) we have

$$v_{k+1} = \lambda_{k+1} P_{k+1}\sigma = \sum_{j=0}^{k} P_{k+1}^{(j)} u_j + P_{k+1}^{(k+1)} u_{k+1}$$

and so

$$\begin{aligned} P_{k+1}^{(k+1)} u_{k+1} &= \lambda_{k+1} P_{k+1}\sigma - \sum_{j=0}^{k} P_{k+1}^{(j)} u_j \\ &= \left(\lambda_{k+1} P_{k+1} - \sum_{j=0}^{k} P_{k+1}^{(j)} \ell_j\right)\sigma \\ &= \left(\sum_{j=0}^{N} \ell_j P_{k+1}^{(j)} - \sum_{j=0}^{k} P_{k+1}^{(j)} \ell_j\right)\sigma = \ell_{k+1} P_{k+1}^{(k+1)}\sigma. \end{aligned}$$

Hence, $u_{k+1} = \ell_{k+1}(x)\sigma$.

Finally, the condition (v) (respectively, (vi)) is just a restatement of the condition (iv) (respectively, (iii)) in terms of the moments $\{\sigma_n\}_0^\infty$ of σ (cf. [22]). □

Remark. The equivalence of two moment relations $S_k(m) = 0$ in (v) and $T_k(m) = 0$ in (vi) was first observed by Littlejohn [22] in which he gave the precise connection between them (see equation (5.5) in [22]).

Now Theorem 2.1 follows directly from Theorem 2.4, since a polynomial set $\{P_n(x)\}_0^\infty$ is an OPS if and only if the moments $\{\sigma_n\}_0^\infty$ of $\{P_n(x)\}_0^\infty$ satisfy the condition (i) in Theorem 2.1.

The rising interest in OPSs satisfying differential equations of the form (1.2) lies partly in the fact that they provide good examples of realizing the general Weyl–Titchmarsh theory of higher order differential equations (see [5],[8],[9] and [11]). In this sense, the equivalence of conditions (i) and (ii) in Theorem 2.4 is quite interesting.

3. Sobolev-type orthogonality. The condition (v) for the symmetry factor $s(x)$ in Lemma 2.2 has an analogue for any canonical moment functional σ of an OPS satisfying the differential equation (1.2). To be precise we have the following.

THEOREM 3.1. *Let $\{P_n(x)\}_0^\infty$, σ, and r be the same as in Theorem 2.4. Then any of the equivalent conditions* (i)–(vi) *in Theorem 2.4 is also equivalent to the following:*

(vii) *There are $r+1$ moment functionals $\{\tau_i\}_0^r$ such that $\tau_r \not\equiv 0$ and*

$$L_{2r}(\phi)\sigma = \sum_0^r (-1)^i [\phi^{(i)}\tau_i]^{(i)} \tag{3.1}$$

for every polynomial $\phi(x)$.

Moreover, the moment functionals σ and $\{\tau_i\}_0^r$ are related by the equations

$$\ell_k(x)\sigma = \sum_{\left[\frac{k+1}{2}\right]}^{\min(r,k)} (-1)^i \binom{i}{k-i} \tau_i^{(2i-k)}, \qquad k = 0, 1, \cdots, 2r. \tag{3.2}$$

Proof. The proof of the equivalence of the condition (ii) in Theorem 2.4 and (vii) in Theorem 3.1 is essentially the same as the proof of the equivalence of the conditions (v) and (vi) in Lemma 2.2 except for the test functions used (functions in $C^{2r}(I)$ for Lemma 2.2 and polynomials for Theorem 3.1) and the interpretation of the bilinear quadratic form $\langle\ ,\ \rangle$.

To prove (3.2), let us expand the right-hand side of (3.1). Then we have

$$\begin{aligned}
\sum_{i=0}^r (-1)^i [\phi^{(i)}\tau_i]^{(i)} &= \sum_{i=0}^r (-1)^i \sum_{j=0}^i \binom{i}{j} \phi^{(i+j)} \tau_i^{(i-j)} \\
&= \sum_{i=0}^r (-1)^i \sum_{k=i}^{2i} \binom{i}{k-i} \phi^{(k)} \tau_i^{(2i-k)} \\
&= \sum_{k=0}^{2r} \phi^{(k)} \sigma_{(k)},
\end{aligned}$$

where

$$\sigma_{(k)} = \sum_{\left[\frac{k+1}{2}\right]}^{\min(r,k)} (-1)^i \binom{i}{k-i} \tau_i^{(2i-k)}, \qquad k = 0, 1, \ldots, 2r. \tag{3.3}$$

Since (3.1) holds for every polynomial $\phi(x)$, we have (3.2) by comparing the coefficients of $\phi^{(k)}$ from both sides of (3.1). In particular, we have $\tau_r = (-1)^r \ell_{2r}(x)\sigma$, so that $\tau_r \equiv 0$ if and only if $\ell_{2r}(x) \equiv 0$ (cf. Lemma 2.3 (i)). This completes the proof of Theorem 3.1. □

Now we are ready to give the main result of this section, that is, the Sobolev-type orthogonality of OPSs satisfying the differential equation (1.2).

THEOREM 3.2. *Let $\{P_n(x)\}_0^\infty$, σ, and r be the same as in Theorem 2.4. Then each of the equivalent conditions* (i)–(vi) *in Theorem 2.4 and* (vii) *in Theorem 3.1 is also equivalent to the following.*

(viii) *There are $r+1$ moment functionals $\{\tau_i\}_0^r$ such that $\tau_r \not\equiv 0$ and*

$$\sum_{i=0}^r \langle \tau_i, P_m^{(i)} P_n^{(i)} \rangle = M_n \delta_{mn}, \qquad m \text{ and } n = 0, 1, \ldots \tag{3.4}$$

where M_n are constants. Moreover we may require $M_n \neq 0$, $n = 0, 1, \ldots$, if necessary.

Proof. Let $\{P_n(x)\}_0^\infty$ be an OPS relative to σ with $\langle \sigma, P_m P_n \rangle = K_n \delta_{mn}$. We first assume that each $P_n(x)$ satisfies the differential equation (1.2) with $N = 2r$ and let $\{\tau_i\}_0^r$ be the moment functionals in Theorem 3.1. Then (3.1) gives

$$\lambda_n K_n \delta_{mn} = \langle \sigma, P_m L_{2r}(P_n) \rangle = \langle L_{2r}(P_n)\sigma, P_m \rangle$$

$$= \left\langle \sum_0^r (-1)^i [P_n^{(i)} \tau_i]^{(i)}, P_m \right\rangle = \sum_0^r \langle \tau_i, P_m^{(i)} P_n^{(i)} \rangle.$$

Hence we have (3.4) with $M_n = \lambda_n K_n$.

Conversely we assume that there are $r+1$ moment functionals $\{\tau_i\}_0^r$, $\tau_r \not\equiv 0$ for which we have (3.4). Then we may rewrite (3.4) with $\sigma_{(k)}$ in (3.3) as

$$\sum_0^{2r} \langle \sigma_{(k)}, P_m^{(k)} P_n \rangle = 0, \qquad m \neq n \tag{3.5}$$

for m and $n = 0, 1, \ldots$, since

$$\sum_0^r \langle \tau_i, P_m^{(i)} P_n^{(i)} \rangle = \sum_0^r \langle P_m^{(i)} \tau_i, P_n^{(i)} \rangle$$

$$= \sum_0^r (-1)^i \langle [P_m^{(i)} \tau_i]^{(i)}, P_n \rangle = \sum_0^{2r} \langle P_m^{(k)} \sigma_{(k)}, P_n \rangle.$$

We now claim that

$$\langle \sigma_{(k)}, P_n \rangle = 0 \tag{3.6}$$

for $n > k$, $k = 0, 1, \ldots, 2r$ so that by Lemma 2.3 (ii)

$$\sigma_{(k)} = \ell_k(x)\sigma \tag{3.7}$$

for some polynomial $\ell_k(x)$ of degree $\leq k$. For simplicity, let us assume that all $P_n(x)$ are monic. For $k = 0$, we have $\sigma_{(0)} = \tau_0$ from (3.3) and so from (3.4)

$$\langle \sigma_{(0)}, P_n \rangle = \langle \tau_0, P_n \rangle = \sum_0^r \langle \tau_i, P_0^{(i)} P_n^{(i)} \rangle = 0$$

for $n > 0$. Hence (3.6) holds for $k = 0$.

Assume that (3.6) holds for $k = 0, 1, \ldots, m-1$ $(1 \leq m < 2r)$. Then we have for $n > m$ from (3.5)

$$0 = \sum_0^{2r} \langle \sigma_{(k)}, P_m^{(k)} P_n \rangle = \sum_0^m \langle \sigma_{(k)}, P_m^{(k)} P_n \rangle$$

$$= \sum_0^{m-1} \langle \sigma_{(k)}, P_m^{(k)} P_n \rangle + \langle \sigma_{(m)}, P_m^{(m)} P_n \rangle$$

$$= \sum_0^{m-1} \langle \sigma, \ell_k P_m^{(k)} P_n \rangle + m! \langle \sigma_{(m)}, P_n \rangle = m! \langle \sigma_{(m)}, P_n \rangle$$

by the induction hypothesis and the fact that $\deg \ell_k P_m^{(k)} \leq m < n$. Hence (3.6) also holds for $k = m$ and our claim is proved inductively. Moreover $\ell_{2r}(x) \not\equiv 0$ since $\tau_r = (-1)^r \ell_{2r}(x)\sigma \not\equiv 0$.

Now consider $L_{2r}(P_n) = \sum_0^{2r} \ell_i P_n^{(i)}$ with $\ell_i(x)$ in (3.7). Since $L_{2r}(P_n)$ is a polynomial of degree $\leq n$, we may write it as

$$L_{2r}(P_n) = \sum_0^n c_j P_j$$

where $\{c_j\}_0^n$ are constants depending on n. From (3.5), (3.7), and the orthogonality of $\{P_n(x)\}_0^\infty$ relative to σ we have

$$\begin{aligned} c_m \langle \sigma, P_m^2 \rangle &= \left\langle \sigma, P_m \sum_0^n c_j P_j \right\rangle = \langle \sigma, P_m L_{2r}(P_n) \rangle \\ &= \langle L_{2r}(P_n)\sigma, P_m \rangle = \sum_0^{2r} \langle P_n^{(i)} \sigma_{(i)}, P_m \rangle \\ &= \sum_0^{2r} \langle \sigma_{(i)}, P_m P_n^{(i)} \rangle = 0 \end{aligned}$$

for $m = 0, 1, \ldots, n-1$. Hence $c_m = 0$ for $m = 0, 1, \ldots, n-1$ and so $L_{2r}(P_n) = c_n P_n = \lambda_n P_n$ by comparing the coefficients of x^n on both sides.

Finally since the term $\ell_0(x)y = \ell_{00}y = \lambda_0 y$ is always common in both sides of the differential equation (1.2) we may take $\ell_0(x)$ arbitrarily so that we may have $\lambda_n \neq 0$ and so $M_n = \lambda_n K_n \neq 0$, $n = 0, 1, \ldots$ by taking ℓ_{00} suitably. □

By taking $\ell_0(x) \equiv 0$, we may have $\tau_0 \equiv 0$ in Theorem 3.2. In particular for $r = 1$, Theorem 3.2 with $\tau_0 \equiv 0$ reduces to the next corollary.

COROLLARY 3.3. *An* OPS $\{P_n(x)\}_0^\infty$ *is classical, that is, it satisfies the differential equation* (1.2) *with* $N = 2$ *if and only if* $\{P_n'(x)\}_1^\infty$ *is a weak orthogonal polynomial set in the sense that there is a nontrivial moment functional* τ *with*

$$\langle \tau, P_m' P_n' \rangle = 0, \qquad m \neq n \tag{3.8}$$

for m *and* $n = 1, 2, \ldots$.

Remark. We refer the reader to §4 where a discussion is made of the representation of the moment functionals σ and $\{\tau_i\}$ of Theorem 3.2 in terms of Stieltjes (signed) measures. The reader can then see that Theorem 3.2 is a generalization of Theorem 1.1. Furthermore, notice that Corollary 3.3 gives an improvement over the Sonine–Hahn classification (Theorem 1.1) in the sense that we may assume that $\{P_n'(x)\}_0^\infty$ is a weak orthogonal polynomial sequence and not necessarily an OPS.

DEFINITION 3.1 (Chihara [3]). *A moment functional* σ *is called symmetric if*

$$\langle \sigma, x^{2n+1} \rangle = 0, \quad n = 0, 1, \ldots. \tag{3.9}$$

An OPS $\{P_n(x)\}_0^\infty$ *relative to* σ *is called symmetric if* σ *is symmetric.*

By definition, it is easy to see that (i) the sum of any two symmetric moment functionals is also symmetric.

(ii) An even order derivative of a symmetric moment functional is also symmetric.

LEMMA 3.4 (Krall and Littlejohn [10]). *If a symmetric* OPS $\{P_n(x)\}_0^\infty$ *satisfies the differential equation* (1.2), *then* $\ell_{ij} = 0$ *when* $i + j$ *is odd.*

Now we have the following characterization theorem of symmetric OPSs.

THEOREM 3.5. *An* OPS $\{P_n(x)\}_0^\infty$ *is symmetric and satisfies the differential equation* (1.2) *of order* $N = 2r$ *if and only if there are* $r+1$ *symmetric moment functionals* $\{\tau_i\}_0^r$ *such that* $\tau_r \not\equiv 0$ *and* (3.4) *holds with* $\tau_0 \not\equiv 0$.

Proof. For $m = n = 0$, (3.4) becomes $\langle \tau_0, P_0^2 \rangle = M_0 \neq 0$ so that $\sigma_{(0)} = \tau_0 = \ell_{00}\sigma$ for some constant $\ell_{00} \neq 0$ (cf. (3.7)). Hence σ is symmetric if τ_0 is symmetric and the sufficiency follows from Theorem 3.2. Conversely, assume that $\{P_n(x)\}_0^\infty$ is a symmetric OPS satisfying the differential equation (1.2) of order $N = 2r$. Then we have (3.4) with the moment functionals $\{\tau_i\}_0^r$ defined by (3.2). We shall show that all τ_i are symmetric by induction on $i = r, r-1, \dots, 0$. First for $i = r$, $\tau_r = (-1)^r \ell_{2r}(x)\sigma$ is symmetric since σ is symmetric and $\ell_{2r}(x)$ has only even powers of x by Lemma 3.4. Next assume that $\{\tau_i\}_{k+1}^r$ are symmetric for some integer k with $0 \leq k \leq r-1$. Then we have from (3.2)

$$\begin{aligned}\ell_{2k}(x)\sigma &= \sum_{k}^{\min(r,2k)} (-1)^i \binom{i}{2k-i} \tau_i^{(2i-2k)} \\ &= (-1)^k \tau_k + \sum_{k+1}^{\min(r,2k)} (-1)^i \binom{i}{2k-i} \tau_i^{(2i-2k)}\end{aligned}$$

so that

$$\tau_k = (-1)^k \ell_{2k}\sigma - \sum_{k+1}^{\min(r,2k)} (-1)^{i+k} \binom{i}{2k-i} \tau_i^{(2i-2k)}. \tag{3.10}$$

Hence τ_k is also symmetric since $\ell_{2k}\sigma$ is symmetric by Lemma 3.4 and all $\tau_i^{(2i-2k)}$, $i \geq k+1$, are symmetric as even order derivatives of symmetric moment functionals. Therefore by induction all τ_i, $i = 0, 1, \dots, r$, are symmetric. Finally we may choose $M_0 \neq 0$ (cf. Theorem 3.2) so that $\tau_0 \not\equiv 0$. □

4. Integral representation of Sobolev-type orthogonality. By a classical theorem of Boas [1] on the Stieltjes moment problem, any moment functional can be represented as a Riemann–Stieltjes integral with respect to some signed Stieltjes measure.

In this respect, it is natural to find integral representations of the moment functionals σ and $\{\tau_i\}_0^r$ in Theorem 3.2.

As briefly discussed in §1, for any OPS $\{P_n(x)\}_0^\infty$ relative to σ satisfying the differential equation (1.2) of order $N = 2r$, an orthogonalizing weight functional $w(x)$ can be constructed by solving a certain overdetermined system of r nonhomogeneous linear differential equations in the space $\mathcal{D}'$ of distributions (see [7],[10],[17] and [22]). More precisely, these r differential expressions are the symmetry expressions $\widetilde{R}_k(s)$ defined in (2.8). However, instead of studying the general distributional solution to $\widetilde{R}_k(s) = 0$ $(k = 0, 1, \dots, r-1)$, we solve $\widetilde{R}_k(s)(x) = g_k(x)$ $(k = 0, 1, \dots, r-1)$ where g_k are suitable *ghost functions*; that is, they have zero moments:

$$\int_{-\infty}^{\infty} x^n g_k(x)dx = 0, \qquad n = 0, 1, \dots.$$

Once $w(x)$ representing σ is found, representations of the moment functionals $\{\tau_i\}_0^r$ in Theorem 3.2 come directly from the equation (3.10) in which σ is replaced by $w(x)$.

The case of classical-type orthogonal polynomials (i.e., $N = 4$) was handled in [19] and we now illustrate it by an OPS satisfying the sixth-order differential equation.

The Krall polynomials [20] are polynomial solutions of the sixth-order differential equation:

(4.1)

$$(x^3-1)^3y^{(6)} + 18x(x^2-1)^2y^{(5)} + [(3AC+3BC+96)x^4 - (6AC+6BC+132)x^2 + (3AC+3BC+36)]y^{(4)}$$

$$+[(24AC+24BC+168)x^3 - (24AC+24BC+168)x]y^{(3)}$$

$$+[(12ABC^2+42AC+42BC+72)x^2 + (12BC-12AC)x - (12ABC^2+30AC+30BC+72)]y''$$

$$+[(24ABC^2+12AC+12BC)x + (12BC-12AC)]y' = \lambda_n y.$$

They are orthogonal with respect to the distribution

$$w(x) = (1/A)\delta(x+1) + (1/B)\delta(x-1) + CH(1-x^2); \tag{4.2}$$

this is the classical Legendre weight function on $(-1,1)$ together with two mass points at $x = \pm 1$. If we let $w_i(x)$ be the representations of moment functionals τ_i, $i = 1,2,3$, then we have from the equation (3.10):

$$w_3(x) = -\ell_6(x)w(x), \tag{4.3}$$

$$w_2(x) = \ell_4(x)w(x) + 3w_3''(x), \tag{4.4}$$

$$w_1(x) = -\ell_2(x)w(x) + w_2''(x). \tag{4.5}$$

By substituting (4.2) into (4.3), (4.4), and (4.5), we can obtain the representations of τ_3, τ_2, and τ_1. For example, we have

$$w_2(x) = C[\ell_4(x) - 18(x^2-1)(5x^2-1)]H(1-x^2)$$

so that

$$\langle \tau_2, \phi(x)\rangle = C\int_{-1}^{1} [\ell_4(x) - 18(x^2-1)(5x^2-1)]\phi(x)\,dx$$

for every polynomial $\phi(x)$.

5. Remark. Let $\{P_n(x)\}_0^\infty$ be an OPS relative to σ satisfying the differential equation (1.2) of order $N = 2r$. Then by Theorem 2.4 (iii) for $k = r$, σ must satisfy

$$r(\ell_{2r}\sigma)' - \ell_{2r-1}\sigma = 0. \tag{5.1}$$

On the other hand, since $\ell_{2r}(x)P_{n+1}'(x)$ is a polynomial of degree $\le n+2r$, we may write it as

$$\ell_{2r}(x)P_{n+1}'(x) = \sum_0^{n+2r} c_j P_j(x)$$

for some constants $\{c_j\}_0^{n+2r}$. Then we have from (5.1)

$$c_k\langle\sigma, P_k^2\rangle = \sum_0^{n+2r} c_j\langle\sigma, P_jP_k\rangle = \langle\sigma, P_k\ell_{2r}P'_{n+1}\rangle$$
$$= -\langle(P_k\ell_{2r}\sigma)', P_{n+1}\rangle = -\left\langle P'_k\ell_{2r}\sigma + \frac{1}{r}P_k\ell_{2r-1}\sigma, P_{n+1}\right\rangle$$
$$= -\left\langle\sigma, \left(P'_k\ell_{2r} + \frac{1}{r}P_k\ell_{2r-1}\right)P_{n+1}\right\rangle$$

for $k = 0, 1, \ldots, n+2r$. Since $\deg\left(P'_k\ell_{2r} + \frac{1}{r}P_k\ell_{2r-1}\right) \le 2r+k-1$, the last term and so c_k is equal to 0 if $k < n-(2r-2)$. Hence we have

$$\ell_{2r}(x)P'_{n+1}(x) = \sum_{n-(2r-2)}^{n+2r} c_jP_j(x) \tag{5.2}$$

which is the differential-difference relation characterizing the semiclassical orthogonal polynomials introduced by Maroni [24]. Therefore, $\{P_n(x)\}_0^\infty$ is semiclassical and $\{P'_{n+1}(x)\}_0^\infty$ is quasi-orthogonal (see Theorem 3.1 in [24]).

REFERENCES

[1] R. P. Boas, *The Stieltjes moment problem for functions of bounded variation*, Bull. Amer. Math. Soc., 45 (1939), pp. 399–404.

[2] S. Bochner, *Über Sturm-Liouvillesche Polynomsysteme*, Math. Z., 29 (1929), pp. 730–736.

[3] T. S. Chihara, *An introduction to orthogonal polynomials*, Gordon and Breach, New York, 1977.

[4] E. A. Coddington and N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, New York, 1955.

[5] W. N. Everitt and L. L. Littlejohn, *Orthogonal polynomials and spectral theory: a survey*, Orthog. Polyn. & their Appl., IMACS Vol. 9, Baltzer (1991), pp. 21–55.

[6] W. Hahn, *Über die Jacobischen Polynome und zwei verwandte Polynomklassen*, Math. Z., 39 (1935), pp. 634–638.

[7] S. S. Kim and K. H. Kwon, *Generalized weights for orthogonal polynomials*, Differential Integral Equations, 4 (1991), pp. 601–608.

[8] A. M. Krall and L. L. Littlejohn, *A singular sixth order differential equation with orthogonal polynomial eigenfunctions*, Lecture Notes in Math. 964, Springer-Verlag, New York, 1983, pp. 435–444.

[9] ———, *Orthogonal polynomials and singular Sturm-Liouville systems* I, Rocky Mountain J. Math., 16 (1986), pp. 435–479.

[10] ———, *On the classification of differential equations having orthogonal polynomial solutions* II, Ann. Mat. Pura Appl., 4 (1987), pp. 77–102.

[11] ———, *Orthogonal polynomials and higher order singular Sturm-Liouville systems*, Acta. Appl. Math., 17 (1989), pp. 99–170.

[12] H. L. Krall, *Certain differential equations for Tchebychev polynomials*, Duke Math. J, 4 (1938), pp. 705–719.

[13] ———, *On orthogonal polynomials satisfying a certain fourth order differential equation*, Penn. State Coll. Studies, No.6, University Park, PA, 1940.

[14] ———, *Self adjoint differential expressions*, Amer. Math. Monthly, 67 (1960), pp. 876–878.

[15] H. L. Krall and O. Frink, *A new class of orthogonal polynomials: The Bessel polynomials*, Trans. Amer. Math. Soc., 65 (1949), pp. 100–115.

[16] H. L. Krall and I. M. Sheffer, *A characterization of orthogonal polynomials*, J. Math. Anal. Appl., 8 (1964), pp. 232–244.

[17] K. H. Kwon, S. S. Kim, and S. S. Han, *Orthogonalizing weights for Tchebychev set of polynomials*, Bull. London Math. Soc., 24 (1992), pp. 361–367.

[18] K. H. Kwon and S. S. Han, *Spectral analysis of Bessel polynomials in Krein space*, Quaestiones Math., 14 (1991), pp. 327–335.

[19] K. H. KWON, L. L. LITTLEJOHN, J. K. LEE, AND B. H. YOO, *Characterization of classical type orthogonal polynomials*, Proc. Amer. Math. Soc., to appear.

[20] L. L. LITTLEJOHN, *The Krall polynomials: A new class of orthogonal polynomials*, Quaestiones Math., 5 (1982), pp. 255–265.

[21] ———, *Symmetry factors for differential equations*, Amer. Math. Monthly, 7 (1983), pp. 462-464.

[22] ———, *On the classification of differential equations having orthogonal polynomial solutions*, Ann. Mat. Pura Appl., 4 (1984), pp. 35–53.

[23] L. L. LITTLEJOHN AND D. RACE, *Symmetric and symmetrizable differential expressions*, Proc. London Math. Soc., 60 (1990), pp. 344–364.

[24] P. MARONI, *Prolégomènes a l'étude des polynomes orthogonaux semi-classiques*, Ann. Mat. Pura Appl., 149 (1987), pp. 165–184.

[25] I. M. SHEFFER, *On the properties of polynomials satisfying a linear differential equation*, Part I, Trans. Amer. Math. Soc., 35 (1933), pp. 184–214.

[26] N. JA. SONINE, *Über die angenäherte Berechnung der bestimmten Integrale und über die dabei vorkommenden ganzen Functionen*, Warsaw Univ. Izv., 18 (1887), pp. 1–76. (In Russian.) Summary in Jbuch. Fortschritte Math. 19, p. 282.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 991–1001, May 1994

011

ON ZEROS OF MULTIVARIATE QUASI-ORTHOGONAL POLYNOMIALS AND GAUSSIAN CUBATURE FORMULAE*

YUAN XU†

Abstract. Zeros of multivariate quasi-orthogonal polynomials are characterized by joint eigenvalues of a family of block Jacobi matrices. The result is used to study Gaussian cubature formula of degree $2n-2$.

Key words. multivariate orthogonal polynomials, quasi-orthogonal polynomials, common zeros, Gaussian cubature

AMS subject classifications. 65D30, 41A05

1. Introduction.

Let w be a nonnegative weight function on $\mathbb{R}$ with infinite support. Let $\{p_n(w)\}_{n=0}^{\infty}$ be a sequence of orthogonal polynomials with respect to the weight function w. For $\rho \in \mathbb{R}$, quasi-orthogonal polynomial of degree n, $q_n(w)$, is defined by

$$q_n(w) = p_n(w) + \rho p_{n-1}(w). \tag{1.1}$$

The polynomial $q_n(w)$ is orthogonal to polynomials of degree $n-2$ or less with respect to w. Quasi-orthogonal polynomials play an important role in the study of quadrature formulae. It is well known that the zeros of nth quasi-orthogonal polynomials are the nodes of a quadrature formula of degree $2n-2$. The case $\rho = 0$ leads to a Gaussian quadrature, which is of degree $2n-1$.

Let Π_n^d be the set of polynomials of total degree n in d variables, and Π^d be the set of all polynomials in d variables. For a nonnegative function W on $\mathbb{R}^d$, a minimal cubature formula of degree m is a linear functional

$$I_N(f) = \sum_{k=1}^{N} \lambda_k f(x_k), \qquad \lambda_k > 0,\ x_k \in \mathbb{R}^d \tag{1.2}$$

where N, the number of the involved nodes x_k, is minimal, such that $\int fW = I_N(f)$ whenever $f \in \Pi_m^d$. It is known [17] that $N \geq \dim \Pi_{[m/2]}^d$ in general. Formulae that attain this lower bound are of the highest precision and are termed Gaussian cubature formulae – of degree $m = 2n-1$, or $m = 2n-2$.

These cubature formulae have been studied by several authors; we refer to [8]–[17], [21]–[23] and their references. The multivariate orthogonal polynomials play an essential role in the study of these cubatures. Let $\mathbb{N}_0$ be the set of nonnegative integers, $\alpha,\ \beta \in \mathbb{N}_0^d$ and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. We denote by $\{\mathbb{P}_\alpha^n\}_{|\alpha|=n}{}_{n=0}^{\infty}$, where $P_\alpha^n \in \Pi_n^d$, the orthonormal polynomials with respect to W, i.e.,

$$\int P_\alpha^n P_\beta^n W dx = \delta_{n,m}\delta_{\alpha,\beta}.$$

For convenience, we assume $\int W(x)dx = 1$ from now on. Mysovskikh [11] characterized the Gaussian cubature of degree $2n-1$: in order such a formula to exist, it is necessary and sufficient that nth degree orthogonal polynomials $\{\mathbb{P}_\alpha^n\}_{|\alpha|=n}$ have

* Received by the editors September 14, 1992; accepted for publication (in revised form) April 1, 1993.

† Department of Mathematics, University of Oregon, Eugene, Oregon 97403-1222.

dim Π_{n-1} common zeros. However, unlike the case of quadrature formula, Gaussian cubature of degree $2n-1$ does not always exist. Möller [8] found an improved lower bound for the minimal number of nodes for the centrally symmetric weight functions. Mysovskikh [11] showed that the minimal number of nodes depends on special properties of the weight function W and its support set, and obtained improved lower bounds for the weight function W that satisfies $W(x) = W(-x)$. From their work it follows that there exist no cubature formulae of degree $2n-1$ for most of the classical weight functions (see [1], [2], [8], [10], [11], and [22]). The only weight functions that are known to admit the Gaussian cubature formula of odd degree for all n are those found very recently in [16]. On the other hand, Schmid [12], [13], Morrow and Patterson [9] studied the cubature formulae of degree $2n-2$, and one surprising result is that for $W(x,y) = (1-x^2)^{1/2}(1-y^2)^{1/2}$ on $[-1,1]^2$ the Gaussian cubature of degree $2n-2$ exists, while the cubature of degree $2n-1$ does not. In general, they showed that a Gaussian cubature of degree $2n-2$ exists if and only if the quasi-orthogonal polynomials $\{Q_\alpha^n\}_{|\alpha|=n}$, where

$$Q_\alpha^n = P_\alpha^n + \sum_{|\beta|=n-1} \gamma_{\alpha,\beta} P_\beta^{n-1}, \qquad |\alpha| = n, \tag{1.3}$$

have dim Π_{n-1}^d distinct real common zeros. The even degree Gaussian cubature formulae are also discussed in [11].

The purpose of this paper is to investigate the zeros of multivariate quasi-orthogonal polynomials, and use the knowledge gained to study the Gaussian cubature formulae. Our main result in §3 characterizes the common zeros of quasi-orthogonal polynomials as common eigenvalues of a family of block Jacobi matrices. For $d \ge 2$, the existence of maximal number of common zeros of $\{Q_\alpha^n\}_{|\alpha|=n}$ is equivalent to that $\{\gamma_{\alpha,\beta}\}$ satisfy certain nonlinear equations involving the coefficients of three-term relations satisfied by the orthogonal polynomials. In §4, we derive these equations for all $d \ge 2$ and study their structures. Finally, in §5, we apply the results in earlier sections to Gaussian cubature formulae, and to interpolation based on the common zeros of quasi-orthogonal polynomials. Our approach is based on our recent study of multivariate orthogonal polynomials [19]–[23], the notation and the preliminaries are given in the next section.

2. Preliminaries. Our treatment of multivariate orthogonal polynomials is based on a three-term relation in vector-matrix form. Let $r_n = r_n^d = \dim \Pi_n^d - \dim \Pi_{n-1}^d$. Let $\{P_\alpha^n\}_{|\alpha|=n} g = \{P_{\alpha_j}\}_{j=1}^{r_n}$ be a sequence of orthonormal polynomials with respect to weight function W, where the elements are rearranged according to the lexicographical order. Introducing vector notation

$$\mathbb{P}_n(x) = \left[P_{\alpha_1}^n(x), P_{\alpha_2}^n(x), \ldots, P_{\alpha_{r_n}}^n(x)\right]^T \tag{2.1}$$

we can express the orthonormal property of $\{P_\alpha^n\}$ by $\int \mathbb{P}_n \mathbb{P}_m^T W = \delta_{m,n} I$, where I is the identity matrix of size $r_n \times r_n$. For our convenience, we call $\{\mathbb{P}_n\}_{n=0}^\infty$ sequence of orthonormal polynomials. Throughout this paper, the notation $A : i \times j$ means that A is a matrix of size $i \times j$. For $x \in \mathbb{R}^d$ we write $x = (x_1, \ldots, x_d)$. From the vector notation it follows that $\{\mathbb{P}_n\}$ satisfies the following.

Three-term relation. For $k \ge 0$, $1 \le i \le d$, there exist matrices $A_{n,i} : r_n \times r_{n+1}$ and $B_{n,i} : r_n \times r_n$, such that

$$x_i \mathbb{P}_n = A_{n,i} \mathbb{P}_{n+1} + B_{n,i} \mathbb{P}_n + A_{n-1,i}^T \mathbb{P}_{n-1}, \qquad 1 \le i \le d, \tag{2.2}$$

where $\mathbb{P}_{-1} = 0$, $\mathbb{P}_0 = 1$, *and* $A_{-1,i}$ *is taken to be zero.*

Rank conditions. For $n \ge 0$, rank $A_{n,i} = r_n$ *for* $1 \le i \le d$, *and*

$$\text{rank } A_n = r_{n+1}, \qquad A_n = (A_{n,1}^T | \ldots | A_{n,d}^T)^T. \tag{2.3}$$

Actually, if $\{\mathbb{P}_n\}_{n=0}^\infty$ is a sequence of polynomial vectors that satisfies three-term relation (2.2) and equation (2.3), then $\{\mathbb{P}_n\}_{n=0}^\infty$ is orthonormal with respect to a square-positive, linear functional Favard theorem [6], [19], [20]. The equation (2.3) also implies that there exist matrices $D_{k,i} : r_{k+1} \times r_k$ such that

$$\sum_{i=1}^{d} D_{k,i} A_{k,i} = I. \tag{2.4}$$

We can take $D_n = (D_{n,1}^T | \cdots | D_{n,d}^T)^T$ as the generalized inverse of A_n. The orthogonality of $\{\mathbb{P}_n\}$ also implies that the coefficient matrices in the three-term relation satisfy the following conditions.

Commuting conditions. For $n \ge 0$, $1 \le i, j \le d$,

$$A_{n,i} A_{n+1,j} = A_{n,j} A_{n+1,i}\ , \tag{2.5}$$

$$A_{n,i} B_{n+1,j} + B_{n,i} A_{n,j} = B_{n,j} A_{n,i} + A_{n,j} B_{n+1,i}\ , \tag{2.6}$$

$$A_{n-1,i}^T A_{n-1,j} + B_{n,i} B_{n,j} + A_{n,i} A_{n,j}^T = A_{n-1,j}^T A_{n-1,i} + B_{n,j} B_{n,i} + A_{n,j} A_{n,i}^T\ . \tag{2.7}$$

These are the conditions that make the linear operators associated with $\{\mathbb{P}_n\}$ formally commuting. The spectral theorem for commuting family of selfadjoint operators have been used to study the integral representation of the linear functional in the Favard theorem [20]. From the three-term relation (2.2) we also have [19] and the next formula.

Christoffel–Darboux formula. For $n \ge 1$, $1 \le i \le d$,

$$K_n(x, y) := \sum_{k=0}^{n-1} \mathbb{P}_k^T(x) \mathbb{P}_k(y) = \frac{[A_{n-1,i} \mathbb{P}_n(x)]^T \mathbb{P}_{n-1}(y) - \mathbb{P}_{n-1}^T(x) [A_{n-1,i} \mathbb{P}_n(y)]}{x_i - y_i}.$$

For $\alpha \in \mathbb{N}_0^d$ and $x \in \mathbb{R}^d$ we write $x^\alpha = x_1^{\alpha_1} \ldots x_d^{\alpha_d}$. For $n \in \mathbb{N}_0$ we denote by x^n the r_n-tuple $\{x^\alpha\}_{|\alpha|=n}$, where the elements are arranged in lexicographical order. In our vector notation, x^n severs the role of monomials, we can write $\mathbb{P}_n$ as

$$\mathbb{P}_n = G_n x^n + G_{n,n-1} x^{n-1} + \cdots, \tag{2.8}$$

where G_n is called leading coefficient matrix of $\mathbb{P}_n$. Let $L_{n,i}$ denote the matrices of size $r_{n-1} \times r_n$, satisfying

$$L_{n,i} x^n = x_i x^{n-1}, \qquad 1 \le i \le d. \tag{2.9}$$

The leading coefficient matrix G_n is invertible, and it is related to the coefficient matrices $A_{n,i}$ in (2.2) by $L_{n,i}$ as follows [23]:

$$A_{n,i} = G_n L_{n+1,i} G_{n+1}^{-1}. \tag{2.10}$$

Other properties of multivariate orthogonal polynomials can be found in [7], [8], and [19]–[23].

3. Characterization of zeros of quasi-orthogonal polynomials. Let $\{\mathbb{P}_n\}_{n=0}^\infty$ be a sequence of orthonormal polynomials. Let the quasi-orthogonal polynomials be defined by (1.3). Using our vector notation we can rewrite the nth quasi-orthogonal polynomial as follows:

$$\mathbb{Q}_n = \mathbb{P}_n + \Gamma \mathbb{P}_{n-1}, \tag{3.1}$$

where $\Gamma = \Gamma_n$ is a matrix of size $r_n \times r_{n-1}$. We consider the common zeros of components of $\mathbb{Q}_n$. For convenience, we call them zeros of $\mathbb{Q}_n$. Our main result in this section characterizes the zeros of $\mathbb{Q}_n$ as eigenvalues of truncated block Jacobi matrices, we now define these matrices.

Let $A_{n,i}$ and $B_{n,i}$ be coefficient matrices in the three-term relation (2.2). The block Jacobi matrices, T_i, associated with $\mathbb{P}_n$ are defined by

$$T_i = \begin{bmatrix} B_{0,i} & A_{0,i} & & \bigcirc \\ A_{0,i}^T & B_{1,i} & A_{1,i} & \\ & A_{1,i}^T & B_{2,i} & \ddots \\ \bigcirc & & \ddots & \ddots \end{bmatrix}, \qquad 1 \le i \le d.$$

These infinite matrices can be considered as linear operators on ℓ^2. For $d = 1$, this is the classical Jacobi matrix. The truncated block Jacobi matrices, $T_{n,i}$, are obtained from T_i by deleting block rows and block columns with numbers $\ge n$. In [22] it is proved that the joint eigenvalues of $T_{n-1,i}$, $1 \le i \le d$, are zeros of $\mathbb{P}_n$, which is well known in the case $d = 1$. For quasi-orthogonal polynomials, we define another set of truncated block Jacobi matrices, $S_{n,i}$. Let Γ be the matrix appeared at (3.1). We define

$$S_{n-1,i} = \begin{bmatrix} B_{0,i} & A_{0,i} & & & & \bigcirc \\ A_{0,i}^T & B_{1,i} & A_{1,i} & & & \\ & \ddots & \ddots & \ddots & & \\ & & & B_{n-2,i} & A_{n-2,i} & \\ \bigcirc & & & A_{n-2,i}^T & B_{n-1,i} - A_{n-1,i}\Gamma & \end{bmatrix}, \qquad 1 \le i \le d.$$

We note that the size of $S_{n-1,i}$ is $\binom{n+d-2}{n-2} \times \binom{n+d-2}{n-2}$, and the size of elements $A_{n,i}$ and $B_{n,i}$ increase with n. We shall denote by S_n the tube of matrices $S_n = (S_{n,1}, \ldots, S_{n,d})$. If there exists a vector x and numbers λ_i such that $S_{n,i}x = \lambda_i x$ for $1 \le i \le d$, then we call $\Lambda = (\lambda_1, \ldots \lambda_d)$ a joint eigenvector of S_n, or simply an eigenvalue of S_n. We now state our main result in this section.

THEOREM 3.1. *Let $\mathbb{Q}_n$ be quasi-orthogonal polynomial defined at* (3.1), *and $S_{n,i}$, $1 \le i \le d$, be the corresponding truncated block Jacobi matrices. Then $\Lambda = (\lambda_1, \ldots, \lambda_d)$ is a zero of $\mathbb{Q}_n$ if and only if Λ is an eigenvalue of S_{n-1}. Moreover, the joint eigenvector is given by $(\mathbb{P}_0^T(\Lambda), \ldots, \mathbb{P}_{n-1}^T(\Lambda))^T$.*

Proof. If $\mathbb{Q}_n(\Lambda) = 0$, then $\mathbb{P}_n(\Lambda) = -\Gamma \mathbb{P}_{n-1}(\Lambda)$. It follows from the three-term relation that

$$\begin{aligned} B_{0,i}\mathbb{P}_0(\Lambda) + A_{0,i}\mathbb{P}_1(\Lambda) &= \lambda_i \mathbb{P}_0(\Lambda), \\ A_{k-1,i}^T \mathbb{P}_{k-1}(\Lambda) + B_{k,i}\mathbb{P}_k(\Lambda) + A_{k,i}\mathbb{P}_{k+1}(\Lambda) &= \lambda_i \mathbb{P}_k(\Lambda), \qquad 1 \le k \le n-2, \\ A_{n-2,i}^T \mathbb{P}_{n-2}(\Lambda) + (B_{n-1,i} - A_{n-1,i}\Gamma)\mathbb{P}_{n-1}(\Lambda) &= \lambda_i \mathbb{P}_{n-1}(\Lambda) \end{aligned}$$

for $1 \le i \le d$. Therefore, we have readily

$$S_{n-1,i}x = \lambda_i x, \qquad x = [\mathbb{P}_0^T(\Lambda), \dots, \mathbb{P}_{n-1}^T(\Lambda)]^T.$$

That is, Λ is the eigenvalue of S_{n-1} with joint eigenvector x. On the other hand, suppose that $\Lambda = (\lambda_1, \dots, \lambda_d)$ is an eigenvalue of S_{n-1} with a joint eigenvector x. Let us write $x = (x_1^T, \dots, x_{n-1}^T)^T$, where $x_j \in \mathbb{R}^{r_j}$. We also define $x_n = 0$. Then we have that $\{x_j\}$ satisfies a three-term relation

$$\begin{aligned} B_{0,i}x_0 + A_{0,i}x_1 &= \lambda_i x_0, \\ A_{k-1,i}^T x_{k-1} + B_{k,i}x_k + A_{k,i}x_{k+1} &= \lambda_i x_k, \qquad 1 \le k \le n-2, \\ A_{n-2,i}^T x_{n-2} + (B_{n-1,i} - A_{n-1,i}\Gamma)x_{n-1} &= \lambda_i x_{n-1} \end{aligned}$$

for $1 \le i \le d$. We can normalize x such that $x_0 = 1 = \mathbb{P}_0$. Let $y_k = \mathbb{P}_k(\Lambda)$ for $0 \le k \le n-1$, and let $y_n = \mathbb{Q}_n(\Lambda)$. From the three-term relation satisfied by $\mathbb{P}_k$ we have that $\{y_k\}_{k=0}^n$ satisfies the same three-term relation that $\{x_k\}_{k=0}^n$ satisfies. Therefore, so does $\{u_k\}_{k=0}^n$, where $u_k = \{x_k - y_k\}$. Since $u_0 = 0$, we can use (2.4) and induction to prove that $u_k = 0$ for $0 \le k \le n$. In particular, we obtain $u_n = y_n = \mathbb{Q}_n(\Lambda) = 0$. □

For $\Gamma = 0$, this theorem is proved in [22] for $d \ge 1$, and the case $d = 1$ is well known. For $\Gamma \ne 0$ and $d = 1$, we actually have

$$q_n = p_n + \rho p_{n-1} = (a_0 \dots a_{n-1})^{-1} \det(xI - S_{n-1}).$$

This formula seems to be unnoticed in the literature. It is clearly very useful both numerically and theoretically. We shall discuss this approach toward the quasi-orthogonal polynomials in a much more general setting in a forthcoming paper. We note that $S_{n-1,i}$ is almost symmetric, only the block in the lower right corner is questionable. It turns out that $S_{n,i}$ is symmetric if we want $\mathbb{Q}_n$ to have maximal numbers of zeros; see Theorem 4.1 below.

4. The condition for maximal number of zeros. Since the size of $S_{n-1,i}$ is $\dim \Pi_{n-1} \times \dim \Pi_{n-1}$, it follows from Theorem 3.1 that $\mathbb{Q}_n$ has at most $\dim \Pi_{n-1}$ zeros. The most interesting case is when $\mathbb{Q}_n$ has these many zeros—the maximal number of zeros, because then a Gaussian cubature formula of degree $2n-2$ exists, and these zeros sever as the nodes of the cubature. For $\mathbb{Q}_n$ to have $\dim \Pi_{n-1}$ zeros, the matrix Γ has to be in a special form and satisfies a nonlinear matrix equation.

THEOREM 4.1. *The polynomial vector $\mathbb{Q}_n$ has $N = \dim \Pi_{n-1}$ distinct zeros if and only if for $1 \le i \le d$*

$$A_{n-1,i}\Gamma = \Gamma^T A_{n-1,i}^T \tag{4.1}$$

and for $1 \le i, j \le d$

$$\begin{aligned} &\Gamma^T(A_{n-1,i}^T A_{n-1,j} - A_{n-1,j}^T A_{n-1,i})\Gamma - (A_{n-1,i}A_{n-1,j}^T - A_{n-1,j}A_{n-1,i}^T) \\ &\quad = (B_{n-1,i}A_{n-1,j} - B_{n-1,j}A_{n-1,i})\Gamma - [(B_{n-1,i}A_{n-1,j} - B_{n-1,j}A_{n-1,i})\Gamma]^T. \end{aligned} \tag{4.2}$$

Proof. From Theorem 3.1 it follows that $\mathbb{Q}_n$ can have at most N distinct zeros, and $\mathbb{Q}_n$ has N zeros if and only if $S_{n,1}, \dots, S_{n,d}$ can be simultaneously diagonalized

by an invertible matrix. Therefore, if $\mathbb{Q}_n$ has N distinct zeros, then $S_{n,i}$ are diagonalizable, thus, commute (cf. [5, p. 52]). On the other hand, if (4.1) holds, then $S_{n,i}$ are symmetric, thus, diagonalizable. Since a family of diagonalizable matrices is simultaneously diagonalizable if and only if it is a commuting family [5 p. 52], we have $S_{n,i}S_{n,j} = S_{n,j}S_{n,i}$ for all $1 \le i, j \le d$. From (2.5), (2.6), and (2.7), these equations are equivalent to

$$\begin{aligned} &B_{n-2,i}A_{n-2,j} + A_{n-2,i}B_{n-1,j} - A_{n-2,i}A_{n-1,j}\Gamma \\ &\quad = B_{n-2,j}A_{n-2,i} + A_{n-2,j}B_{n-1,i} - A_{n-2,j}A_{n-1,i}\Gamma, \end{aligned} \tag{4.3}$$

$$\begin{aligned} &A^T_{n-2,i}B_{n-2,j} + B_{n-1,i}A^T_{n-2,j} - A_{n-1,i}\Gamma A^T_{n-2,j} \\ &\quad = A^T_{n-2,j}B_{n-2,i} + B_{n-1,j}A^T_{n-2,i} - A_{n-1,j}\Gamma A^T_{n-2,i}, \end{aligned} \tag{4.4}$$

and

$$\begin{aligned} &A^T_{n-2,i}A_{n-2,j} + (B_{n-1,i} - A_{n-1,i}\Gamma)(B_{n-1,j} - A_{n-1,j}\Gamma) \\ &\quad = A^T_{n-2,j}A_{n-2,i} + (B_{n-1,j} - A_{n-1,j}\Gamma)(B_{n-1,i} - A_{n-1,i}\Gamma). \end{aligned} \tag{4.5}$$

From (2.5) and (2.6) it follows that (4.3) is always true. Using (2.7) we can rewrite (4.5) as

$$\begin{aligned} &A_{n-1,i}\Gamma B_{n-1,j} + B_{n-1,i}A_{n-1,j}\Gamma + A_{n-1,i}A^T_{n-1,j} - A_{n-1,i}\Gamma A_{n-1,j}\Gamma \\ &\quad = A_{n-1,j}\Gamma B_{n-1,i} + B_{n-1,j}A_{n-1,i}\Gamma + A_{n-1,j}A^T_{n-1,i} - A_{n-1,j}\Gamma A_{n-1,i}\Gamma. \end{aligned}$$

Since $B_{n,i} = \int x_i \mathbb{P}_n \mathbb{P}_n^T W$ is symmetric, it is easily seen that (4.2) follows from this equation and (4.1). Therefore, it remains to prove (4.1). From (2.6) it follows that (4.4) can be simplified to

$$A_{n-1,i}\Gamma A^T_{n-2,j} = A_{n-1,j}\Gamma A^T_{n-2,i}. \tag{4.6}$$

We shall show that (4.6) is equivalent to (4.1). One way is easy: if $A_{n-1,i}\Gamma$ is symmetric, then by (2.5)

$$A_{n-1,i}\Gamma A^T_{n-2,j} = \Gamma^T A^T_{n-1,i}A^T_{n-2,j} = \Gamma^T A^T_{n-1,j}A^T_{n-2,i} = A_{n-1,j}\Gamma A^T_{n-2,i}.$$

On the other hand, let $L_{n,i}$ be matrices defined at (2.9) and G_n be the leading coefficient matrix of $\mathbb{P}_n$, then it follows from (2.10) that (4.6) is equivalent to

$$L_{n,i}H_n L^T_{n-1,j} = L_{n,j}H_n L^T_{n-1,i}, \qquad 1 \le i, j \le d, \tag{4.7}$$

where $H_n = G_n^{-1}\Gamma(G^T_{n-1})^{-1}$. It turns out that (4.7) implies that H_n is of the special form $H_n = (h_{\alpha+\beta})_{|\alpha|=n,|\beta|=n-1}$, where the order of elements in $\{\alpha : |\alpha| = n\}$ and $\{\beta : |\beta| = n-1\}$ is taken to be lexicographical. The proof of this fact is rather technical; we delay it to Lemma 4.2 below. Using this fact we can now complete the proof of this theorem. Indeed, this special form of H_n allows us to introduce a linear functional, $\mathcal{L}^*$, defined on Π^d such that $\mathcal{L}^*(x^\alpha) = h_\alpha$, and write $H_n = \mathcal{L}^*(x^n(x^{n-1})^T)$. From this we have, by (2.9),

$$L_{n,i}H_n = \mathcal{L}^*(x_i x^{n-1}(x^{n-1})^T),$$

from which follows that $L_{n,i}H_n$ is symmetric. By the definition of H_n, this means

$$L_{n,i}G_n^{-1}\Gamma(G_{n-1}^{-1})^T = G_{n-1}^{-1}\Gamma^T(G_n^{-1})^T L_{n-1,i}^T,$$

which, by (2.10), implies that $A_{n-1,i}\Gamma$ is symmetric. □

LEMMA 4.2. *A matrix H_n satisfies the equation* (4.7) *if and only if it is of the form $H_n = (h_{\alpha+\beta})_{|\alpha|=n,|\beta|=n-1}$, where the rows and columns are indexed in lexicographical order of $\{\alpha : |\alpha| = n\}$ and $\{\beta : |\beta| = n-1\}$, respectively.*

Proof. By using the $\mathcal{L}^*$ as in the above, we see that the sufficient part follows readily. We now prove the necessary part. Let $H_n = (h_{\alpha,\beta})_{|\alpha|=n,|\beta|=n-1}$. We shall call (α,β) and (α',β') *related* if $|\alpha| = |\alpha'| = n$, $|\beta| = |\beta'| = n-1$, and $\alpha+\beta = \alpha'+\beta'$. We need to prove that

$$h_{\alpha,\beta} = h_{\alpha',\beta'} \tag{4.8}$$

for related (α,β) and (α',β').

First we note that we only need to prove (4.8) for the special case that α and α' differ at only two components—say, $\alpha_i \neq \alpha_i'$ and $\alpha_j \neq \alpha_j'$—this implies that β and β' differ at only two components, $\beta_i \neq \beta_i'$ and $\beta_j \neq \beta_j'$, since $\alpha+\beta = \alpha'+\beta'$. Indeed, if (α,β) and (α',β') are related, then we can always relate these two pairs by finite many intermediate pairs $(\alpha^{(k)},\beta^{(k)})$, where two consecutive pairs, say, $(\alpha^{(k)},\beta^{(k)})$ and $(\alpha^{(k+1)},\beta^{(k+1)})$, are related and differ only at two components. Next, let

$$\mathcal{N}_n = \{\alpha \in \mathbb{N}_0^d : |\alpha| = n\}, \qquad \mathcal{N}_{n,i} = \{\alpha \in \mathcal{N}_n : \alpha_i > 0\}.$$

We show that if (α,β) and (α',β') are related and differ only at their ith and jth components, then either

$$\alpha \in \mathcal{N}_{n,i}, \quad \beta \in \mathcal{N}_{n-1,j}, \quad \alpha' \in \mathcal{N}_{n,j}, \quad \beta' \in \mathcal{N}_{n-1,i}, \tag{4.9}$$

or (4.9) with i and j interchanged. Indeed, if $\alpha_i = 0$ then $\alpha_j \neq 0$, and from $|\alpha| = |\alpha'|$ or $\alpha_i + \alpha_j = \alpha_i' + \alpha_j'$, we have $\alpha_i' \neq 0$. Also, from $\alpha_i + \beta_i = \alpha_i' + \beta_i'$ we have $\beta_i \neq 0$. It then follows from $\beta_i + \beta_j = \beta_i' + \beta_j'$ that

$$\beta_j' = \beta_i + \beta_j - \beta_i' = \alpha_i' + \beta_j \neq 0.$$

The same proof works if $\alpha_i' \neq 0$. The case that α_i, α_j, α_i', and α_j' are all nonzero is easier. Thus, (4.9) is proved.

Suppose that (α,β) and (α',β') are related and differ only at their ith and jth components, and that (4.9) holds. We only need to prove (4.8) for these two pairs. We note that $L_{n,i}$ maps $\mathcal{N}_n$ to $\mathcal{N}_{n,i}$. Therefore, using (4.9), the equation (4.7) implies

$$(h_{\alpha,\beta})_{\alpha\in\mathcal{N}_{n,i},\beta\in\mathcal{N}_{n-1,j}} = (h_{\alpha,\beta})_{\alpha\in\mathcal{N}_{n,j},\beta\in\mathcal{N}_{n-1,i}}.$$

Let $(\alpha_i) = (\alpha_1 \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_d) \in \mathbb{N}_0^d$. Then the above equation is equivalent to

$$(h_{(\alpha_i),(\beta_j)})_{\alpha\in\mathcal{N}_{n-1},\beta\in\mathcal{N}_{n-2}} = (h_{(\alpha_j),(\beta_i)})_{\alpha\in\mathcal{N}_{n-1},\beta\in\mathcal{N}_{n-2}},$$

in particular, $h_{(\alpha_i),(\beta_j)} = h_{(\alpha_j),(\beta_i)}$. Since (α_i) and (α_j) differ at only two components, α_j and β_j can be taken as depending on α_i and β_i, respectively. If we let $\alpha_i + \beta_i = k$, then we have

$$h_{\alpha,\beta}|_{\alpha_i=1,\beta_i=k-1} = h_{\alpha,\beta}|_{\alpha_i=2,\beta_i=k-2} = \cdots = h_{\alpha,\beta}|_{\alpha_i=k-1,\beta_i=1},$$

where $h_{\alpha,\beta}|_{\alpha_i=a,\beta_i=b}$ means that α_j and β_j are assigned accordingly and other elements of α and β are unchanged in these equations. In particular, we have proved (4.8) for the pairs (α,β) and (α',β'). □

COROLLARY 4.3. *If $\mathbb{Q}_n$ has $\dim \Pi_{n-1}$ distinct zeros, then the associated matrices $S_{n,i}$ are symmetric.*

For $d = 2$ Theorem 4.1 was proved in [12] and [15], where the monic orthogonal polynomials $\mathbb{P}_n^* = G_n^{-1}\mathbb{P}_n = x^n + \cdots$ were used. The equation (4.7) was stated as that H_n is a Hankel matrix. For $d > 2$ the Hankel property is replaced by the form $H_n = (h_{\alpha+\beta})$. For $\Gamma = 0$, this theorem reduces to Mysovskikh's condition [10], see [22]. Our condition (4.1) is essential in the proof of Theorem 4.4 below. It is also essential in extending the result of [21] on Lagrange interpolation to $\mathbb{R}^d$, $d > 2$; see §5.

If Λ is a zero of $\mathbb{Q}_n$ and at least one partial derivative of $\mathbb{Q}_n$ at Λ is not zero, then we say that Λ is a simple zero of $\mathbb{Q}_n$.

THEOREM 4.4. *If $\mathbb{Q}_n$ has $\dim \Pi_{n-1}$ zeros, then all zeros are real, distinct, and simple. Moreover, polynomials $\mathbb{Q}_n$ and $\mathbb{P}_{n-1}$ do not have common zeros.*

Proof. By Corollary 4.3, all eigenvalues of $S_{n,i}$ are real, thus, all zeros of $\mathbb{Q}_n$ are real. From the Christoffel–Darboux formula, we can write

$$\begin{aligned}(x_i - y_i)K_n(x,y) = \mathbb{Q}_n^T(x)A_{n-1,i}^T[\mathbb{P}_{n-1}(y) - \mathbb{P}_{n-1}^T(x)] \\ -\mathbb{P}_{n-1}^T(x)A_{n-1,i}^T[\mathbb{Q}_n(y) - \mathbb{Q}_n(x)],\end{aligned}$$

where we have used (4.1). From this identity we have by dividing $x_i - y_i$ and letting $y_i \to x_i$,

$$K_n(x,x) = \mathbb{P}_{n-1}^T(x)A_{n-1,i}\partial_i\mathbb{Q}_n(x) - \mathbb{Q}_n^T(x)A_{n-1,i}^T\partial_i\mathbb{P}_{n-1}(x),$$

where $\partial_i = \partial/\partial x_i$ denotes the partial derivative with respect to x_i. Therefore, if Λ is a zero of $\mathbb{Q}_n$, then we have

$$\mathbb{P}_{n-1}^T(\Lambda)A_{n-1,i}\partial_i\mathbb{Q}_n(\Lambda) = \sum_{k=0}^{n-1}\mathbb{P}_k^T(\Lambda)\mathbb{P}_k(\Lambda) > 0.$$

Thus, $\partial_i\mathbb{Q}_n(\Lambda) \neq 0$, and $\mathbb{P}_{n-1}(\Lambda) \neq 0$. □

For $d = 2$ this theorem is proved in [21] using the same argument. We choose to give the full proof here since the formulation in [21] is in terms of monic orthogonal polynomials, thus, somewhat different. Part of this theorem has been proved in [13] using the algebraic ideal theory.

5. Applications. As we mentioned in the Introduction, Gaussian cubature of degree $2n-2$ exists if and only if there is a matrix Γ such that the corresponding nth quasi-orthogonal polynomial has $\dim \Pi_{n-1}$ zeros [9], [12], [13]. From Theorem 4.1, we immediately have the following.

THEOREM 5.1. *For a given weight function W, a Gaussian cubature of degree $2n-2$ exists if and only if there is a matrix Γ that satisfies* (4.1) *and* (4.2).

Two classes of these cubature formulae are found recently in [16], including even cubatures of degree $2n-1$. The support set for these integrals is the image of a simplex under the transformation $x_k \to u_k$, where $u_k = u_k(x_1,\ldots,x_d)$ are the elementary symmetric polynomials of $x_1,\ldots,x_d$. The significance of Gaussian cubature of degree $2n-2$ perhaps lies in the fact that it may exist for the classical integrals, while

Gaussian cubature of degree $2n-1$ does not exist. For $d=2$ and the weight function $W(x,y)=(1-x^2)^{1/2}(1-y^2)^{1/2}$ on $[-1,1]^2$ – product Chebyshev weight of the second kind, a cubature formula of degree $2n-2$ was found in [9] that has nodes based on quasi-orthogonal polynomials with

$$\Gamma = \begin{bmatrix} \bigcirc & & 1 \\ & \ddots & \\ 1 & & \bigcirc \\ 0 & \dots & 0 \end{bmatrix}.$$

Later, all Γ that lead to cubature of degree $2n-2$ for this weight function were found in [14]. However, for $d>2$ even the product of Chebyshev polynomials of the second kind turns out to be difficult. In the following we verify the first nontrivial case, $d=3$, $n=2$.

We recall that the Chebyshev polynomials of the second kind is given by $U_n(x) = \sin(n+1)\theta/\sin\theta$, where $x=\cos\theta$. These polynomials are orthonormal with respect to weight function $w(x)=\sqrt{\frac{2}{\pi}}(1-x^2)^{1/2}$, and they satisfy the three-term relation

$$xU_n(x) = \tfrac{1}{2}U_{n+1}(x) + \tfrac{1}{2}U_{n-1}(x).$$

Let $W(x,y,z) = w(x)w(y)w(z)$. The multivariate orthonormal polynomials with respect to W are given by

$$P^n_{j,k}(x,y,z) = U_i(x)U_j(y)U_k(z), \qquad i+j+k=n.$$

Using the vector notation $\mathbb{P}_n$, it is easy to verify that $\{\mathbb{P}_n\}$ satisfies the three-term relation

$$x_i\mathbb{P}_n(x) = \tfrac{1}{2}L_{n,i}\mathbb{P}_{n+1}(x) + \tfrac{1}{2}L^T_{n-1,i}\mathbb{P}_{n-1}(x),$$

where $x=(x_1,x_2,x_3)=(x,y,z)$. The first nontrivial case is $n=2$, where the quasi-orthogonal polynomial is $\mathbb{Q}_2 = \mathbb{P}_2 + \Gamma\mathbb{P}_1$. There are six components in $\mathbb{Q}_2$. We choose our Γ to be

$$\Gamma^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Since we have

$$L_{2,1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad L_{2,2} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$L_{2,3} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

it is easy to verify that all conditions in Theorem 4.1 are satisfied. Therefore, there are $4(=\dim \Pi^3_1)$ zeros for $\mathbb{Q}_2$. They lead to a Gaussian cubature of degree 2 on $\mathbb{R}^3$. Actually, we can write the components of $\mathbb{Q}_2$ out explicitly. They are $U_2(x)$,

$U_1(x)U_1(y)+U_1(z)$, $U_1(x)U_1(z)+U_1(y)$, $U_2(y)$, $U_1(y)U_1(z)+U_1(x)$, and $U_2(z)$. Since $U_1(x) = 2x$ and $U_2(x) = 4x^2-1$, it is easy to verify that the zeros are $(1/2, 1/2, -1/2)$, $(1/2, -1/2, 1/2)$, $(-1/2, 1/2, 1/2)$, and $(-1/2, -1/2, -1/2)$.

However, we have not been able to find a proper Γ in the case $d = 3$ and $n = 3$, not to say the general case for this weight function. From our initial investigation, we believe that the result as simple as that in [9] for $d = 2$ is not possible for $d > 2$, and perhaps the existence of Γ that solves (4.1) and (4.2) is mostly negative, even for the product Chebyshev weight.

Finally we mention another application of Theorem 4.1, concerning Lagrange interpolation based on the zeros of $\mathbb{Q}_n$. Suppose $\mathbb{Q}_n$ has $N := \dim \Pi_{n-1}$ zeros, denoted by $\{x_k\}_{k=1}^N$. The question is to find a polynomial P in $\mathbb{P}_{n-1}$ such that

$$P(x_k) = f(x_k), \qquad 1 \le k \le N$$

for every function f. Let λ_k be weights in cubature formula (1.2). Then we have the final theorem.

THEOREM 5.2. *Suppose that Γ satisfies the condition* (4.1) *and* (4.2). *Then the polynomial $L_n(x) = L_n(f; x)$, defined by*

$$L_n(x) = \sum_{k=1}^{N} f(x_k)\lambda_k K_n(x, x_k),$$

is the unique solution of the Lagrange interpolation based on the zeros of $\mathbb{Q}_n$.

This theorem is proved in [21] for $d = 2$. The proof for the case $d > 2$ is the same in spirit, but (4.1) is vital. We note that monic orthogonal polynomials are used in [21], thus some formulas need to be modified accordingly in order to carry the proof to $d > 2$. The modification is mostly straightforward; we will not elaborate. The L^2 convergence of $L_n(f)$ and other expressions of fundamental polynomials, $\ell_{kn} = \lambda_k K_n(\cdot, x_k)$, are also discussed in [21]. The extension to $d > 2$ poses no difficulty.

REFERENCES

[1] H. BERENS AND H. SCHMID, *On the number of nodes of odd degree cubature formulae for integrals with Jacobi weights on a simplex*, in Numerical Integration, T.O. Espelid and A. Genz, eds., Kluwer Academic Publications, Derdrecht, 1992, pp. 37–44.

[2] H. BERENS, H. SCHMID, AND Y. XU, *On two dimensional definite orthogonal systems and on a lower bound for the number of nodes of associated cubature formulae*, SIAM J. Math. Anal., to appear.

[3] T.S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Mathematics and its Applications, Vol. 13, Gordon and Breach, New York, 1978.

[4] P.J. DAVIS AND P. ROBINOWITZ, *Methods of Numerical Integration*, Academic Press, New York, 1975.

[5] R.A. HORN AND C.R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[6] M.A. KOWALSKI, *The recursion formulas for orthogonal polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 309–315.

[7] ———, *Orthogonality and recursion formulas for polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 316–323.

[8] H. MÖLLER, *Kubaturformeln mit minimaler knotenzahl*, Numer. Math., 25 (1976), pp. 185–200.

[9] C.R. MORROW AND T.N.L. PATTERSON, *Construction of algebraic cubature rules using polynomial idea theory*, SIAM J. Numer. Anal., 15 (1978), pp. 953–976.

[10] I.P. MYSOVSKIKH, *Numerical characteristics of orthogonal polynomials in two variables*, Vestnik Leningrad Univ. Math., 3 (1976), pp. 323–332.

[11] I. P. MYSOVSKIKH, *Interpolation cubature formulas*, Nauka, Moscow, 1981. (In Russian.)
[12] H. SCHMID, *On cubature formulae with a minimal number of knots*, Numer. Math., 31 (1978), pp. 282–297.
[13] ———, *Interpolatorische Kubaturformeln*, Diss. Math., (220) 1983, pp. 1–122.
[14] ———, *On minimal cubature formulae of even degree*, Internat. Ser. Numer. Math. Vol. 85, Birkhäuser, Basel, 1988, pp. 216–225.
[15] ———, *Minimal cubature formulae and matrix equation*, preprint.
[16] H. SCHMID AND Y. XU, *On bivariate Gaussian cubature formulae*, Proc. Amer. Math. Soc., to appear.
[17] A. STROUD, Approximate Calculation of Multiple Integrals, Prentice Hall, Englewood Cliffs, NJ, 1971.
[18] G. SZEGÖ, Orthogonal Polynomials, Amer. Math. Soc. Colloq. Publ. Vol. 23, Providence, RI, 4th ed., 1975.
[19] Y. XU, *On multivariate orthogonal polynomials*, SIAM J. Math. Anal., 24 (1993), pp. 783–794.
[20] ———, *Multivariate orthogonal polynomials and operator theory*, Trans. Amer. Math. Soc., to appear.
[21] ———, *Gaussian cubature and bivariate polynomial interpolation*, Math. Comp., 59 (1992), pp. 547–555.
[22] ———, *Block Jacobi matrices and zeros of multivariate orthogonal polynomials*, Trans. Amer. Math. Soc., to appear.
[23] ———, *Recurrence formulas for multivariate orthogonal polynomials*, Math. Comp., to appear.

SIAM J. MATH. ANAL.
Vol. 25, No. 3, pp. 1002–1022, May 1994

012

BARNES AND RAMANUJAN-TYPE INTEGRALS ON THE q-LINEAR LATTICE*

MIZAN RAHMAN† AND SERGEI K. SUSLOV‡

This paper is dedicated to Frank Olver and Richard Askey.

Abstract. Let C be a contour in the complex s-plane and $\rho(s)$ be the solution of a Pearson-type first-order difference equation with coefficient functions $\sigma(s)$ and $\tau(s)$, on the q-linear lattice $x(s) = q^{-s}$, $0 < q < 1$. For the cases in which (i) $\sigma(s), \tau(s)$ are polynomials of degrees 2 and 1, respectively, and (ii) $\sigma(s)$ is a polynomial of degree 2 but $\tau(s)$ has a simple pole, the integral $\int_C \rho(s) q^{-s}\, ds$ is considered. When C is the whole real line, q-analogues of some formulas due to Ramanujan are obtained, and some of the questions raised in a previous paper are resolved. When C is along the imaginary axis, the iteration technique in the parameters of $\rho(s)$ works, permitting alternative proofs of a formula due to Askey and Roy, as well as its extension by Gasper, to be given.

Key words. nonuniform lattices, difference equations, Pearson-type equation, basic hypergeometric series, Askey–Roy integral

AMS subject classifications. primary 33A15; secondary 33A10

1. Introduction.

The purpose of this paper is to resolve some of the questions raised in the authors' recent article [25] on the beta integrals, particularly those concerned with the sums and integrals on the q-linear lattice.

A difference analogue of the classical hypergeometric differential equation is

$$\tilde{\sigma}(x(s)) \frac{\nabla}{\nabla x_1(s)} \left[\frac{\triangle y(s)}{\triangle x(s)} \right] + \frac{\tilde{\rho}(x(s))}{2} \left[\frac{\triangle y(s)}{\triangle x(s)} + \frac{\nabla y(s)}{\nabla x(s)} \right] + \lambda y(s) = 0, \tag{1.1}$$

where $x(s)$ is generally a nonuniform lattice, $s \in \mathbb{C}$, $x_k(s) = x(s + \frac{k}{2})$, $k \in \mathbb{C}$, $\triangle f(s) = f(s+1) - f(s)$, $\nabla f(s) = \triangle f(s-1)$, λ is a constant, and $\tilde{\sigma}(x(s))$, $\tilde{\tau}(x(s))$ are polynominals in $x(s)$ of degrees at most 2 and 1, respectively. Difference equations on nonuniform lattices are a familiar topic in the Russian mathematical literature (see, for example, the standard Russian textbook [28]). However, a comprehensive analysis of the solutions of (1.1), particularly the polynomial solutions, corresponding to linear and quadratic lattices as well as their q-analogues, has only recently appeared in the works of the Russian mathematicians; see Nikiforov and Suslov [19], Nikiforov and Uvarov [20]–[22], Nikiforov, Suslov, and Uvarov [23], [24], Suslov [30], and Atakishiyev and Suslov [9].

By denoting

$$v_1(s) = \frac{\triangle y(s)}{\triangle x(s)}, \qquad v_2(s) = \frac{\triangle v_1(s)}{\triangle x_1(s)}, \tag{1.2}$$

equation (1.1) can be written in the form

$$\sigma(s) v_2(s-1) + \tau(s) v_1(s) + \lambda y(s) = 0, \tag{1.3}$$

* Received by the editors July 7, 1992; accepted for publication February 3, 1993. This work, supported in part by the Natural Sciences and Engineering Research Council of Canada grant A6197, was completed while the second author was visiting Carleton University in January–May 1992. This paper was originally submitted for the special issue dedicated to Frank Olver and Richard Askey (*SIAM J. Math. Anal.*, vol 25, no. 2).

† Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6 Canada.

‡ Kurchatov Institute of Atomic Energy, Moscow 123182, Russia.

where

$$\begin{aligned} \tau(s) &= \tilde{\tau}(x(s)), \\ \sigma(s) &= \tilde{\sigma}(x(s)) - \tfrac{1}{2}\tilde{\tau}(x(s))\nabla x_1(s). \end{aligned} \tag{1.4}$$

If $\rho(s)$ satisfies the Pearson-type equation

$$\triangle[\rho(s)\sigma(s)] = \rho(s)\tau(s)\nabla x_1(s), \tag{1.5}$$

then (1.3) can be expressed in a self-adjoint form

$$\nabla[\rho(s+1)\sigma(s+1)v_1(s)] + \lambda y(s)\nabla x_1(s) = 0. \tag{1.6}$$

It was shown in [10] that the hypergeometric property of (1.3), namely, that the successive difference derivatives of $y(s)$, $v_n(s) = \triangle v_{n-1}(s)/\triangle x_{n-1}(s)$ satisfy equations of the same kind, is maintained if and only if the lattice $x(s)$ is of the form

$$x(s) = \begin{cases} C_1 q^{-s} + C_2 q^s & \text{if } q \neq 1, \\ C_1 s^2 + C_2 s & \text{if } q = 1, \end{cases} \tag{1.7}$$

where C_1 and C_2 are arbitrary constants not both zero. When $q = 1$, the lattice is linear if $x(s) = s$ and quadratic if $x(s) = C_1 s^2 + C_2 s$, $C_1 \neq 0$. When $q \neq 1$ and C_1 and C_2 are both nonzero, the lattice is called q-quadratic; if one of C_1, C_2 is zero, then the lattice is q-linear. For the purposes of this paper we shall assume that $0 < q < 1$ when $q \neq 1$ and that the q-linear lattice is defined by $x(s) = q^{-s}$.

In [25] we were concerned with solutions of (1.5), their sums and integrals in the complex plane, over linear and q-quadratic lattices. We introduced a greater degree of generality by allowing $\sigma(s)$ and $\tau(s)$ to have simple poles in addition to the zeros just mentioned (see also [26]). In the concluding section of [25] we gave a summary of results and left some question marks on the formulas corresponding to the q-linear lattice. This paper is devoted entirely to this particular case.

For the domain of s in (1.5) or (1.6) there are two possible situations: (i) s is discrete variable varying in unit steps from $s = a$ to $s = b - 1$ (it is permissible for a to be $-\infty$ and/or b to be ∞); (ii) s varies continuously on a smooth curve in some domain of the complex plane.

In case (i) we get the formula

$$\sum_{s=a}^{b-1} \rho(s)\tau(s)\nabla x_1(s) = \rho(b)\sigma(b) - \rho(a)\sigma(a). \tag{1.8}$$

If a and b are both finite, then usually they are both zeros of $\rho(s)\sigma(s)$, so that the right side vanishes and we get

$$\sum_{s=a}^{b-1} \rho(s)\tau(s)\nabla x_1(s) = 0. \tag{1.9}$$

This itself is not a summation formula, but usually this leads to a two-term recurrence formula for the sum $\sum_{s=a}^{b-1} \rho(s)\nabla x_1(s)$ in terms of one of the parameters of $\rho(s)$ and hence can be iterated. This procedure ultimately leads to a summation formula.

In the continuous case (ii) we may find a suitable contour C in the complex s-plane that does not go through any singularities of the integrands, so that (1.5) gives

$$\int_C \triangle [\rho(s)\sigma(s)]\, ds = \int_C \rho(s)\tau(s)\nabla x_1(s)\, ds. \tag{1.10}$$

If C has the property that

$$\int_C \rho(s+1)\sigma(s+1)\, ds = \int_{C'} \rho(s'+1)\sigma(s'+1) ds', \tag{1.11}$$

where C' is the contour obtained from C by the shift $s' = s - 1$, then (1.10) gives

$$\int_C \rho(s)\sigma(s)\nabla x_1(s)\, ds = 0. \tag{1.12}$$

Note that, by Cauchy's theorem, (1.11) implies that there are no singularities of $\rho(s+1)\sigma(s+1)$ between C and C'. Similar to (1.9), (1.12) is not an integration formula but leads to an evaluation of the integral $\int_C \rho(s)\nabla x_1(s)\, ds$ by iteration in terms of the parameters of $\rho(s)$.

The classical beta integral of Euler, namely,

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \qquad \mathrm{Re}(\alpha,\beta) > 0, \tag{1.13}$$

has been extended in many different ways in the last 200 years, notably by Cauchy [13], Barnes [11], [12], Ramanujan [27], and, more recently, Andrews and Askey [1], Askey [2]–[6], Askey and Roy [7], and Askey and Wilson [8].

Broadly speaking, a beta integral can be classified into two distinct types. The first is the Barnes-type integral, typical of which is Barnes' first lemma [11]:

$$\begin{aligned} &\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \Gamma(a+s)\Gamma(b+s)\Gamma(c-s)\Gamma(d-s)\, ds \\ &\qquad = \frac{\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)}{\Gamma(a+b+c+d)}. \end{aligned} \tag{1.14}$$

Its q-analogue is due to Watson [31]:

$$\begin{aligned} &\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \Gamma_q(a+s)\Gamma_q(b+s)\Gamma_q(c-s)\Gamma_q(d-s)w(s)q^s\, ds \\ &\qquad = q^c \frac{\Gamma(c-d)\Gamma(1+d-c)}{\Gamma_q(c-d)\Gamma_q(1+d-c)} \frac{\Gamma_q(a+c)\Gamma_q(a+d)\Gamma_q(b+c)\Gamma_q(b+d)}{\Gamma_q(a+b+c+d)}, \end{aligned} \tag{1.15}$$

where

$$w(s) = \frac{\Gamma(c-s)\Gamma(1-c+s)\Gamma(d-s)\Gamma(1-d+s)}{\Gamma_q(c-s)\Gamma_q(1-c+s)\Gamma_q(d-s)\Gamma_q(1-d+s)} \tag{1.16}$$

and $\Gamma_q(x)$ is the q-gamma function defined by

$$\Gamma_q(x) = \frac{(q;q)_\infty}{(q^x;q)_\infty}(1-q)^{1-x}, \qquad 0 < q < 1,\ x \neq 0, -1, -2, \ldots, \tag{1.17}$$

with $\lim_{q\to 1^-} \Gamma_q(x) = \Gamma(x)$,

$$(a;q)_\infty = \prod_{n=0}^{\infty} (1-aq^n); \tag{1.18}$$

see Gasper and Rahman [16].

The second is the Ramanujan-type integral, typical of which is Ramanujan's integral:

$$\begin{aligned}\int_{-\infty}^{\infty} &\frac{w(x)\,dx}{\Gamma(a+x)\Gamma(b+x)\Gamma(c-x)\Gamma(d-x)} \\ &= \frac{\Gamma(a+b+c+d-3)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)} \int_0^1 w(x)\,dx,\end{aligned} \tag{1.19}$$

where $w(x\pm 1) = w(x)$ and $\operatorname{Re}(a+b+c+d) > 3$. To our knowledge, no one has found a q-analogue of this formula, and so we will give one in §5. That there may be some problems with q-extensions of Ramanujan-type beta integrals could be anticipated from the following facts. There is a sum version of (1.19) known as Dougall's sum [14]:

$$\begin{aligned}\sum_{n=-\infty}^{\infty} &\frac{1}{\Gamma(a+n)\Gamma(b+n)\Gamma(c-n)\Gamma(d-n)} \\ &= \frac{\Gamma(a+b+c+d-3)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)},\end{aligned} \tag{1.20}$$

$\operatorname{Re}(a+b+c+d) > 3$. This may be written in the standard bilateral series notation

$$ {}_2H_2\left[\begin{matrix}1-c, & 1-d\\ a, & b\end{matrix}\right] = \frac{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)\Gamma(a+b+c+d-3)}{\Gamma(a+c-1)\Gamma(a+d-1)\Gamma(b+c-1)\Gamma(b+d-1)}. \tag{1.21}$$

However, there is no such simple summation formula for the basic bilateral series ${}_2\psi_2$. The closest analogue we have is

$$\begin{aligned}{}_2\psi_2&\left[\begin{matrix}q/c, & q/d\\ a, & b\end{matrix};q,abcd/q^3\right] \\ &= \frac{(\alpha,q/\alpha,ab/\alpha q,\alpha q^2/ab;q)_\infty}{(a/\alpha,\alpha q/a,b/\alpha,\alpha q/b;q)_\infty}\frac{(q,ac/q,ad/q,bc/q,bd/q;q)_\infty}{(a,b,c,d,abcd/q^3;q)_\infty} \\ &\quad + \frac{\alpha^2}{q^2}\frac{(q/a,q/b,\alpha c/q,\alpha d/q;q)_\infty}{(c,d,\alpha/a,\alpha/b;q)_\infty}\,{}_2\psi_2\left[\begin{matrix}q^2/\alpha c, & q^2/\alpha d\\ aq/\alpha, & bq/\alpha\end{matrix};q,abcd/q^3\right],\end{aligned} \tag{1.22}$$

where $|abcd/q^3| < 1$, α is an arbitrary parameter such that $\alpha \neq q^{\pm n}$, $n = 0,1,2,\ldots$, and no zero factors appear on the denominators of the two expressions on the right side. For the definition of ${}_2\psi_2$ and other basic hypergeometric series that we shall use in this paper, see [16]. Formula (1.22) is not listed in this particular form in the literature (see, however, [25] for an equivalent form), but it can be easily deduced from the general transformation formula [16, eq. (5.4.3)] by specializing the parameters. It

is not immediately obvious how (1.22) can be regarded as a q-analogue of (1.21) in the sense that (1.22) leads to (1.21) in the limit $q \to 1^-$. The limit is easier to take if one chooses, for example, $\alpha = -1$ and uses the q-gamma functions. One can show by replacing a, b, c, d by q^a, q^b, q^c, q^d that the second term on the right side of (1.22) approaches 0 as $q \to 1^-$ if $\mathrm{Re}(a+b+c+d) > 3$.

The technique of iterating with respect to the parameters of $\rho(s)$ to evaluate its integrals works very well for Barnes-type contours because of the vanishing of the left side of (1.10) in most cases of interest, enabling one to reduce the evaluation problem to two-term recurrences. This was the technique used by some of the authors mentioned previously and by the authors in [25]. See also the interesting papers by Kalnins and Miller [17] and [18].

However, this technique is not very efficient when one deals with a Ramanujan-type integral where C is the whole real line or part of it. Generally, one gets a nonzero contribution from the left side of (1.10), as we saw in [25], resulting in a nonhomogeneous two-term recurrence calling for a special treatment of the inhomogeneous term. The problem is not too bad when one has symmetries with respect to all the parameters, as is the case for quadratic and q-quadratic lattices (see [25]), but it becomes very messy in the q-linear case, where the symmetries are broken. It is in this sense that the statements in the last paragraph of [25] about the difficulties with the q-linear case can be defended.

However, for integrals on the real line one can resort to a simpler approach. Suppose $f(x)$ is continuous on $[a, \infty)$, it has no singularities, and its integral on $[a, \infty)$ exists. Suppose also that $\sum_{k=0}^{\infty} f(x+k)$ converges uniformly for all $x \in [a, a+1]$. Then it can be shown that

$$\int_a^{\infty} f(x)\,dx = \int_a^{a+1} \sum_{k=0}^{\infty} f(x+k)\,dx. \tag{1.23}$$

For integrals on the whole real line the corresponding formula is

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 \sum_{k=-\infty}^{\infty} f(x+k)\,dx, \tag{1.24}$$

provided $f(x)$ has no singularities on $\mathbb{R}$, $\sum_{k=-\infty}^{\infty} f(x+k)$ converges uniformly for $x \in [0,1]$, and the integral on the left side exists. These are the formulas that will enable us to resolve the questions that were raised in [25] regarding the difficulties with the q-linear case. We really do not need a Ramanujan to come to our rescue or even a helping hand from Richard Askey, as we wished in [25]. The price is that we now need to rely heavily on the summation and transformation formulas of basic hypergeometric series, which is not the case for Barnes-type integrals.

The paper is organized in the following way. In §2 we shall consider the solutions of the Pearson-type equation (1.5) for the q-linear lattice with different sets of choices of the functions $\sigma(s)$ and $\sigma(s) + \tau(s)\nabla x_1(s)$. In §3 we shall give an integral version of the q-Gauss sum

$$\sum_{s=0}^{\infty} \frac{(a,b;q)_s}{(q,c;q)_s} (c/ab)^s = \frac{(c/a, c/b; q)_{\infty}}{(c, c/ab; q)_{\infty}}, \tag{1.25}$$

$|c/ab| < 1$; see [16]. An integral analogue of the nonterminating balanced ${}_3\phi_2$ series will be dealt with in §4 (see [16] for the definitions). In §5 we shall show that a particular q-analogue of Ramanujan's integral (1.19) is

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \left(aq^s, bq^s, cq^{-s}, dq^{-s}; q\right)_\infty q^{s^2-2s} (ab)^s w(s)\, ds \\
&\quad = \frac{q^2}{ab} \frac{\left(q, ac/q, ad/q, bc/q, bd/q; q\right)_\infty}{\left(-a, -q/a, -b, -q/b, abcd/q^3; q\right)_\infty} \\
&\quad\quad \cdot \int_0^1 \left(-q^s, -q^{1-s}, -abq^{s-2}, -q^{3-s}/ab; q\right)_\infty q^{s^2-3s} (ab)^s w(s)\, ds \\
&\quad\quad - q^{-1} \frac{\left(-c/q, -d/q; q\right)_\infty}{\left(-1/a, -1/b; q\right)_\infty} {}_2\psi_2 \left[\begin{matrix} -q^2/c, & -q^2/d \\ -qa, & -qb \end{matrix} ; q, abcd/q^3 \right] \\
&\quad\quad \cdot \int_0^1 \left(aq^s, q^{1-s}/a, bq^s, q^{1-s}/b; q\right)_\infty q^{s^2-s} (ab)^s w(s)\, ds,
\end{aligned} \tag{1.26}
$$

where $|abcd/q^3| < 1$ and $w(s \pm 1) = w(s)$. Note the integrands in both integrals on the right side are unit-periodic functions of s. By replacing a, b, c, d by q^a, q^b, q^c, q^d, respectively, and taking the limit $q \to 1^-$ it is not hard to see that the limit of this formula gives (1.19). In §6 we shall give an extension of (1.26). In §7 we shall give an alternative proof of Askey and Roy's formula

$$
\begin{aligned}
&\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\left(\alpha e^{i\theta}/d, qde^{-i\theta}/\alpha, qe^{i\theta}/\alpha c, \alpha ce^{-i\theta}; q\right)_\infty}{\left(ae^{i\theta}, be^{i\theta}, ce^{-i\theta}, de^{-i\theta}; q\right)_\infty} d\theta \\
&\quad = \frac{\left(\alpha, q/\alpha, \alpha c/d, dq/\alpha c, abcd; q\right)_\infty}{\left(ac, ad, bc, bd, q; q\right)_\infty},
\end{aligned} \tag{1.27}
$$

where α is an arbitrary parameter with $\alpha cd \neq 0$ and $\max(|a|, |b|, |c|, |d|) < 1$; see [7] or [16]. In §8 we shall give a proof of Gasper's extension of (1.27) given in [15]. These proofs will be based on two-term recurrences that we mentioned earlier.

2. The q-linear lattice and the solution of the Pearson equation. For the q-linear lattice $x(s) = q^{-s}$, $0 < q < 1$, we shall consider both polynomial- and rational-function-type expressions for the functions $\sigma(s)$ and $\sigma(s) + \tau(s)\nabla x_1(s)$. For the polynomial case we shall take

$$
\begin{aligned}
&\sigma(s) = s_3 s_4 \left(1 - s_1 q^{-s}\right)\left(1 - s_2 q^{-s}\right), \\
&\sigma(s) + \tau(s)\nabla x_1(s) = \left(q^{-s} - s_3\right)\left(q^{-s} - s_4\right).
\end{aligned} \tag{2.1}
$$

In the rational function case we choose

$$
\begin{aligned}
&\sigma(s) = \left(1 - s_1 q^{-s}\right)\left(1 - s_2 q^{-s}\right)/s_1 s_2, \\
&\sigma(s) + \tau(s)\nabla x_1(s) = q^{-2s} \frac{\left(1 - s_3 q^s\right)\left(1 - s_4 q^s\right)\left(1 - s_5 q^s\right)}{1 - q^{s+1}/s_6},
\end{aligned} \tag{2.2}
$$

with $s_1 s_2 s_3 s_4 s_5 s_6 = q$. From (2.1) we get

$$
\frac{\rho(s+1)}{\rho(s)} = \frac{q^{2s+2}}{s_1 s_2} \frac{\left(1 - q^{-s}/s_3\right)\left(1 - q^{-s}/s_4\right)}{\left(1 - q^{s+1}/s_1\right)\left(1 - q^{s+1}/s_2\right)} \tag{2.3}
$$

$$
= \frac{\left(1 - q^{-s}/s_3\right)\left(1 - q^{-s}/s_4\right)}{\left(1 - s_1 q^{-s-1}\right)\left(1 - s_2 q^{-s-1}\right)} \tag{2.4}
$$

$$= \frac{q^2}{s_1 s_2 s_3 s_4} \frac{(1 - s_3 q^s)(1 - s_4 q^s)}{(1 - q^{s+1}/s_1)(1 - q^{s+1}/s_2)} \tag{2.5}$$

$$= \frac{q^{-2s}}{s_3 s_4} \frac{(1 - s_3 q^s)(1 - s_4 q^s)}{(1 - s_1 q^{-s-1})(1 - s_2 q^{-s-1})}. \tag{2.6}$$

For Ramanujan-type integrals on $(-\infty, \infty)$, the appropriate form is (2.3). For Gauss-type integrals on $(-\infty, -b)$, (2.4) leads to the right asymptotics, while (2.5) is the right one on (a, ∞), where a and b are finite. For the Barnes-type integrals on the imaginary axis, however, the appropriate form is (2.6). The different appropriate forms in the rational-function case are, likewise,

$$\frac{\rho(s+1)}{\rho(s)} = \frac{q^{2s+2}}{s_1 s_2} \frac{(1 - q^{-s}/s_3)(1 - q^{-s}/s_4)(1 - q^{-s}/s_5)}{(1 - q^{s+1}/s_1)(1 - q^{s+1}/s_2)(1 - s_6 q^{-s-1})} \tag{2.7}$$

$$= \frac{(1 - q^{-s}/s_3)(1 - q^{-s}/s_4)(1 - q^{-s}/s_5)}{(1 - s_1 q^{-s-1})(1 - s_2 q^{-s-1})(1 - s_6 q^{-s-1})} \tag{2.8}$$

$$= q^2 \frac{(1 - s_3 q^s)(1 - s_4 q^s)(1 - s_5 q^s)}{(1 - q^{s+1}/s_1)(1 - q^{s+1}/s_2)(1 - q^{s+1}/s_6)} \tag{2.9}$$

$$= \frac{q^{-2s}}{s_3 s_4} \frac{(1 - s_3 q^s)(1 - s_4 q^s)(1 - q^{-s}/s_5)}{(1 - s_1 q^{-s-1})(1 - s_2 q^{-s-1})(1 - s_6 q^{-s-1})}. \tag{2.10}$$

The general solution of (2.3) is

$$\begin{aligned} \rho(s) &= p(s; s_1, s_2, s_3, s_4) \\ &= p(s) \left(q^{s+1}/s_1, q^{s+1}/s_2, q^{1-s}/s_3, q^{1-s}/s_4; q\right)_\infty, \end{aligned} \tag{2.11}$$

where

$$\frac{p(s+1)}{p(s)} = \frac{q^{2s+2}}{s_1 s_2}. \tag{2.12}$$

We may take as the general solution of (2.12)

$$p(s) = q^{s^2} \left(\frac{q}{s_1 s_2}\right)^s w(s), \tag{2.13}$$

or

$$p(s) = \frac{w(s)}{\left(q^{s+1}/\alpha s_1, \alpha s_1 q^{-s}, \alpha q^{s+1}/s_2, s_2 q^{-s}/\alpha; q\right)_\infty}, \tag{2.14}$$

where $w(s \pm 1) = w(s)$ and $\alpha s_1 s_2 \neq 0$, α arbitrary. The general solution of (2.4) and (2.5) can similarly be written in the forms

$$\rho(s) = \frac{\left(q^{1-s}/s_3, q^{1-s}/s_4; q\right)_\infty}{\left(s_1 q^{-s}, s_2 q^{-s}; q\right)_\infty} w(s) \tag{2.15}$$

and

$$\rho(s) = \frac{\left(q^{s+1}/s_1, q^{s+1}/s_2, ; q\right)_\infty}{\left(s_3 q^s, s_4 q^s; q\right)_\infty} \left(\frac{q^2}{s_1 s_2 s_3 s_4}\right)^s w(s), \tag{2.16}$$

respectively, with $w(s\pm 1) = w(s)$. Note that the rational-function extension of (2.15) is

$$\rho(s) = \frac{\left(q^{1-s}/s_3, q^{1-s}/s_4, q^{1-s}/s_5; q\right)_\infty}{\left(s_1 q^{-s}, s_2 q^{-s}, s_6 q^{-s}; q\right)_\infty} w(s) \tag{2.17}$$

and that of (2.16) is

$$\rho(s) = q^{2s} \frac{\left(q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6; q\right)_\infty}{\left(s_3 q^s, s_4 q^s, s_5 q^s; q\right)_\infty} w(s). \tag{2.18}$$

Finally, the general solution of (2.6) is

$$\rho(s) = \frac{p(s)}{\left(s_3 q^s, s_4 q^s, s_1 q^{-s}, s_2 q^{-s}; q\right)_\infty}, \tag{2.19}$$

with

$$\frac{p(s+1)}{p(s)} = \frac{q^{-2s}}{s_3 s_4}. \tag{2.20}$$

Since

$$\frac{\left(1 - q^{-s}/\alpha s_3\right)\left(1 - \alpha q^{-s-1}/s_4\right)}{\left(1 - \alpha s_3 q^s\right)\left(1 - s_4 q^{s+1}/\alpha\right)} q = \frac{q^{-2s}}{s_3 s_4}, \qquad \alpha s_3 s_4 \neq 0, \tag{2.21}$$

the general solution of (2.20) can be written in the form

$$p(s) = \left(\alpha s_3 q^s, q^{1-s}/\alpha s_3, s_4 q^{s+1}/\alpha, \alpha q^{-s}/s_4; q\right) q^s w(s). \tag{2.22}$$

The corresponding solution in the rational-function case is

$$\begin{aligned} \rho(s) &= \rho_\alpha(s; s_1, s_2, s_3, s_4, s_5) \\ &= \frac{\left(\alpha q^s/s_2, s_1 q^{1-s}/\alpha, q^{s+1}/\alpha s_2, \alpha s_2 q^{-s}, q^{s+1}/s_6; q\right)_\infty}{\left(s_3 q^s, s_4 q^s, s_5 q^s, s_1 q^{-s}, s_2 q^{-s}; q\right)_\infty} q^s w(s), \end{aligned} \tag{2.23}$$

where, once again, $w(s)$ is a unit-periodic function of s.

Since $\nabla x_1(s) = q^{-1/2}(1-q)q^{-s}$, the integral that we shall need to consider throughout this paper is

$$\int_C \rho(s) \nabla x_1(s)\, ds = q^{-1/2}(1-q) \int_C \rho(s) q^{-s}\, ds. \tag{2.24}$$

3. An integral analogue of the q-Gauss sum. Let us consider the integral

$$I(s_1, s_2, s_3, s_4) = \int_a^\infty \frac{\left(q^{s+1}/s_1, q^{s+1}/s_2; q\right)_\infty}{\left(s_3 q^s, s_4 q^s; q\right)_\infty} w(s) \left(\frac{q}{s_1 s_2 s_3 s_4}\right)^s ds, \tag{3.1}$$

corresponding to (2.16), where $|q^a| < \min(|s_1|, |s_2|)$ and

$$\left|\frac{q}{s_1 s_2 s_3 s_4}\right| < 1. \tag{3.2}$$

If we assume that $|q^a s_3|$ and $|q^a s_4|$ are not of the form q^{-k}, $k = 0, 1, 2, \ldots$, then the integrand in (3.1) has no singularities and behaves like $\left(q/s_1 s_2 s_3 s_4\right)^s$ for large

s, and so the integral converges provided that the periodic factor $w(s)$ has, itself, no singularities in $[a, a+1]$, which we shall assume to be the case. Then by (1.23) we have

$$I(s_1,s_2,s_3,s_4)=\int_a^{a+1} ds\frac{\left(q^{s+1}/s_1,q^{s+1}/s_2;q\right)_\infty}{\left(s_3q^s,s_4q^s;q\right)_\infty}w(s)\left(\frac{q}{s_1s_2s_3s_4}\right)^s$$

(3.3)
$$\cdot{}_3\phi_2\left[\begin{matrix}q, & s_3q^s, & s_4q^s\\ & & \\ & q^{s+1}/s_1, & q^{s+1}/s_2\end{matrix};q,\frac{q}{s_1s_2s_3s_4}\right].$$

Applying the transformation formulas [16, eq. (3.2.10)], [16, eq. II.24], and [16, eq. (3.2.7)], in that order, we find that

$$\begin{aligned}&{}_3\phi_2\left[\begin{matrix}s_3q^s, & q, & s_4q^s\\ & & \\ & q^{s+1}/s_1, & q^{s+1}/s_2\end{matrix};q,\frac{q}{s_1s_2s_3s_4}\right]\\ &=\frac{\left(q,q/s_1s_4,q/s_2s_4,s_3q^s,q^{s+1}/s_1s_2s_3,s_1s_2s_3q^{-s};q\right)_\infty}{\left(s_1s_3,s_2s_3,q/s_1s_2s_3s_4,q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_4;q\right)_\infty}\\ &\quad+\frac{\left(q/s_1s_4,q^s/s_1,q^s/s_2,s_1s_2s_3q^{1-s};q\right)_\infty}{\left(s_2s_3,q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_4;q\right)_\infty}\frac{s_1s_2s_3q^{-s}}{1-s_1s_3}\\ &\quad\cdot{}_3\phi_2\left[\begin{matrix}s_1s_3, & s_1s_2s_3s_4, & s_1q^{1-s}\\ & & \\ & s_1s_3q, & s_1s_2s_3q^{1-s}\end{matrix};q,q/s_1s_4\right]\end{aligned}\tag{3.4}$$

if $|q/s_1s_4|<1$. Hence

$$\begin{aligned}&\int_a^\infty\frac{\left(q^{s+1}/s_1,q^{s+1}/s_2;q\right)_\infty}{\left(s_3q^s,s_4q^s;q\right)_\infty}w(s)\left(\frac{q}{s_1s_2s_3s_4}\right)^s ds\\ &=\frac{\left(q,q/s_1s_4,q/s_2s_4;q\right)_\infty}{\left(s_1s_3,s_2s_3,q/s_1s_2s_3s_4;q\right)_\infty}\int_a^{a+1}\frac{\left(s_1s_2s_3q^{-s},q^{s+1}/s_1s_2s_3;q\right)_\infty}{\left(s_4q^s,q^{1-s}/s_4;q\right)_\infty}\\ &\quad\cdot w(s)\left(\frac{q}{s_1s_2s_3s_4}\right)^s ds\\ &\quad+\frac{\left(q/s_1s_4;q\right)_\infty s_1s_2s_3}{\left(s_2s_3;q\right)_\infty\left(1-s_1s_3\right)}\int_a^{a+1}ds\frac{\left(q^s/s_1,q^s/s_2,s_1s_2s_3q^{1-s};q\right)_\infty}{\left(s_3q^s,s_4q^s,q^{1-s}/s_4;q\right)_\infty}\\ &\quad\cdot w(s)\left(\frac{1}{s_1s_2s_3s_4}\right)^s{}_3\phi_2\left[\begin{matrix}s_1s_3, & s_1s_2s_3s_4, & s_1q^{1-s}\\ & & \\ & s_1s_3q, & s_1s_2s_3q^{1-s}\end{matrix};q,q/s_1s_4\right].\end{aligned}\tag{3.5}$$

In the special case $s_1s_2s_3s_4 = 1$, this reduces to

$$\begin{aligned}\int_a^\infty &\frac{\left(q^{s+1}/s_1, q^{s+1}/s_2; q\right)_\infty}{\left(s_3q^s, q^s/s_1s_2s_3; q\right)_\infty} w(s)\, q^s\, ds \\ &= -\frac{s_1s_2s_3}{(1-s_1s_3)(1-s_2s_3)} \int_0^1 w(s)\, ds \\ &\quad + \frac{s_1s_2s_3}{(1-s_1s_3)(1-s_2s_3)} \int_a^{a+1} \frac{\left(q^s/s_1, q^s/s_2; q\right)_\infty}{\left(s_3q^s, q^s/s_1s_2s_3; q\right)_\infty} w(s)\, ds.\end{aligned} \tag{3.6}$$

This is a q-analogue of Ramanujan's formula [27, eq. (10.1)] in which there appears to be an error in that there is no term corresponding to the first term on the right side of (3.6).

Observe that a direct application of [16, eq. (3.2.7)] on the ${}_3\phi_2$ series in (3.3) gives another form of (3.5):

$$\begin{aligned}\int_a^\infty &\frac{\left(q^{s+1}/s_1, q^{s+1}/s_2; q\right)_\infty}{\left(s_3q^s, s_4q^s; q\right)_\infty} w(s) \left(\frac{q}{s_1s_2s_3s_4}\right)^s ds \\ &= \frac{1}{1-\frac{q}{s_1s_2s_3s_4}} \int_a^{a+1} ds\, \frac{\left(q^s/s_1, q^{s+1}/s_2; q\right)_\infty}{\left(s_3q^s, s_4q^s; q\right)_\infty} w(s) \left(\frac{q}{s_1s_2s_3s_4}\right)^s \\ &\qquad \cdot {}_3\phi_2\left[\begin{matrix} q, & q/s_2s_3, & q/s_2s_4 \\ & q^2/s_1s_2s_3s_4, & q^{s+1}, \end{matrix}; q, q^s/s_1\right],\end{aligned} \tag{3.7}$$

where the assumption $|q^a| < |s_1|$ is needed to ensure the convergence of the ${}_3\phi_2$ series.

4. An integral analogue of the balanced ${}_3\phi_2$ series. Using (2.18), we shall now consider the integral

$$J(s_1, s_2, s_3, s_4, s_5) = \int_a^\infty \frac{\left(q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6; q\right)_\infty}{\left(s_3q^s, s_4q^s, s_5q^s; q\right)_\infty} w(s)\, q^s\, ds, \tag{4.1}$$

where it is assumed that the denominator of the integrand has no zeros on $[a, \infty)$ and that $|q^a| < \min(|s_1|, |s_2|, |s_6|)$, with s_6 satisfying the balance condition $s_1s_2s_3s_4s_5s_6 = q$. Proceeding as before, we have

$$\begin{aligned}J(s_1, s_2, s_3, s_4, s_5) &= \int_a^{a+1} ds\, \frac{\left(q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6; q\right)_\infty}{\left(s_3q^s, s_4q^s, s_5q^s; q\right)_\infty} w(s)\, q^s \\ &\quad \cdot {}_4\phi_3\left[\begin{matrix} q, & s_3q^s, & s_4q^s, & s_5q^s \\ & q^{s+1}/s_1, & q^{s+1}/s_2, & q^{s+1}/s_6 \end{matrix}; q, q\right].\end{aligned} \tag{4.2}$$

The ${}_4\phi_3$ series is balanced, so that there are some transformation formulas that can be applied on it. Thus, use of [16, eq. (2.10.10)] gives

(4.3)

$$\frac{\left(q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6; q\right)_\infty}{\left(s_3q^s, s_4q^s, s_5q^s; q\right)_\infty} {}_4\phi_3\left[\begin{matrix} q, & s_3q^s, & s_4q^s, & s_5q^s \\ & q^{s+1}/s_1, & q^{s+1}/s_2, & q^{s+1}/s_6 \end{matrix}; q, q\right]$$

$$= s_6q^{-s}\frac{\left(q, qs_6/s_1, qs_6/s_2; q\right)_\infty}{\left(s_3s_6, s_4s_6, s_5s_6; q\right)_\infty} {}_3\phi_2\left[\begin{matrix} s_3s_6, & s_4s_6, & s_5s_6 \\ & qs_6/s_1, & qs_6/s_2 \end{matrix}; q, q\right]$$

$$+ \frac{\left(q^{s+1}/s_1s_2s_3, q^{s+1}/s_1s_2s_4, q^{s+1}/s_1s_2s_5, q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6, s_6q^{-s}; q\right)_\infty}{\left(s_3s_6, s_4s_6, s_5s_6, s_3q^s, s_4q^s, s_5q^s, q^{2s+1}/s_1s_2; q\right)_\infty}$$

$$\cdot {}_8W_7\left(q^{2s}/s_1s_2; q^s/s_1, q^s/s_2, s_3q^s, s_4q^s, s_5q^s; q, s_6q^{1-s}\right),$$

where

(4.4)

$${}_8W_7\left(a; b, c, d, e, f; q, a^2q^2/bcdef\right)$$

$$:= {}_8\phi_7\left[\begin{matrix} a, & q\sqrt{a}, & -q\sqrt{a}, & b, & c, & d, & e, & f \\ & \sqrt{a}, & -\sqrt{a}, & aq/b, & aq/c, & aq/d, & aq/e, & aq/f \end{matrix}; q, \frac{a^2q^2}{bcdef}\right]$$

is the very well-poised ${}_8\phi_7$ series; see [16]. Using the transformation formula [16, eq. (2.10.1)]

(4.5)

$${}_8W_7\left(a; b, c, d, e, f; q, \lambda q/ef\right)$$

$$= \frac{\left(aq, aq/ef, \lambda q/e, \lambda q/f; q\right)_\infty}{\left(\lambda q, \lambda q/ef, aq/e, aq/f; q\right)_\infty} {}_8W_7\left(\lambda; \lambda b/a, \lambda c/a, \lambda d/a, e, f; q, aq/ef\right),$$

where $\lambda = qa^2/bcd$, we can transform the ${}_8W_7$ series in (4.3) to a more convenient form. The result of using (4.3) and (4.5) in (4.2) is

(4.6)

$$\int_a^\infty \frac{\left(q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6; q\right)_\infty}{\left(s_3q^s, s_4q^s, s_5q^s; q\right)_\infty} w(s)\, q^s\, ds$$

$$= s_6\frac{\left(q, qs_6/s_1, qs_6/s_2; q\right)_\infty}{\left(s_3s_6, s_4s_6, s_5s_6; q\right)_\infty}$$

$$\cdot\left\{ {}_3\phi_2\left[\begin{matrix} s_3s_6, & s_4s_6, & s_5s_6 \\ & qs_6/s_1, & qs_6/s_2 \end{matrix}; q, q\right]\int_0^1 w(s)\, ds\right.$$

$$- \int_a^{a+1} ds\, w(s)\frac{\left(q^{s+1}/s_1s_2s_3, q^{s+1}/s_1s_2s_4, q^{s+1}s_1s_2s_5; q\right)_\infty}{\left(s_3q^s, s_4q^s, s_5q^s; q\right)_\infty}$$

$$\left.\cdot\frac{\left(q^s/s_6; q\right)_\infty}{\left(s_6q^{s+1}/s_1s_2; q\right)_\infty} {}_8W_7\left(s_6q^s/s_1s_2; q^s/s_1, q^s/s_2, s_3s_6, s_4s_6, s_5s_6; q, q\right)\right\}.$$

5. A q-analogue of Ramanujan's integral. As we saw in §2, the appropriate solution of the Pearson equation in this case is given by (2.11), with $p(s)$ as in (2.13) or in (2.14). We will see that it does not matter which form of $p(s)$ we choose since an appropriate adjustment of $w(s)$ leads the result of one choice to the other. For the sake of definiteness let us pick (2.13), so that the integral to consider is

$$I(s_1,s_2,s_3,s_4)=\int_{-\infty}^{\infty}\left(q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_3,q^{1-s}/s_4;q\right)_\infty \cdot w(s)\,q^{s^2}\,(s_1s_2)^{-s}\,ds. \tag{5.1}$$

Use of (1.24) then gives

(5.2)
$$I(s_1,s_2,s_3,s_4)=\int_0^1 ds\,\left(q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_3,q^{1-s}/s_4;q\right)_\infty w(s) \cdot q^{s^2}\,(s_1s_2)^{-s}\,{}_2\psi_2\left[\begin{matrix} s_3q^s, & s_4q^s, \\ q^{s+1}/s_1, & q^{s+1}/s_2 \end{matrix};q,\frac{q}{s_1s_2s_3s_4}\right].$$

By (1.22) and appropriate choice of the arbitrary parameter α we have

$$\begin{aligned} &{}_2\psi_2\left[\begin{matrix} s_3q^s, & s_4q^s, \\ q^{s+1}/s_1, & q^{s+1}/s_2 \end{matrix};q,\frac{q}{s_1s_2s_3s_4}\right] \\ &\quad=\frac{q^{-s}}{\alpha}\frac{\left(q,q/s_1s_3,q/s_1s_4,q/s_2s_3,q/s_2s_4;q\right)_\infty}{\left(q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_3,q^{1-s}/s_4,q/s_1s_2s_3s_4;q\right)_\infty} \\ &\qquad\cdot\frac{\left(-\alpha q^s,-q^{1-s}/\alpha,-q^s/\alpha s_1s_2,-\alpha s_1s_2q^{1-s};q\right)_\infty}{\left(-q\alpha s_1,-q\alpha s_2,-1/\alpha s_1,-1/\alpha s_2;q\right)_\infty} \\ &\qquad-\alpha q^{s-1}\frac{\left(-\alpha/s_3,-\alpha/s_4,s_1q^{-s},s_2q^{-s};q\right)_\infty}{\left(-\alpha s_1/q,-\alpha s_2/q,q^{1-s}/s_3,q^{1-s}/s_4;q\right)_\infty} \\ &\qquad\cdot{}_2\psi_2\left[\begin{matrix} -qs_3/\alpha, & -qs_4/\alpha, \\ -q^2/\alpha s_1, & -q^2/\alpha s_2 \end{matrix};q,\frac{q}{s_1s_2s_3s_4}\right]. \end{aligned} \tag{5.3}$$

This gives the formula

$$\int_{-\infty}^{\infty}\left(q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_3,q^{1-s}/s_4;q\right)_\infty w(s)\,q^{s^2}\,(s_1s_2)^{-s}\,ds \tag{5.4}$$
$$=\frac{1}{\alpha}\frac{\left(q,q/s_1s_3,q/s_1s_4,q/s_2s_3,q/s_2s_4;q\right)_\infty}{\left(-\alpha qs_1,-\alpha qs_2,-1/\alpha s_1,-1/\alpha s_2,q/s_1s_2s_3s_4;q\right)_\infty}$$

$$\cdot\int_0^1 \left(-\alpha q^s,-q^{1-s}/\alpha,-q^s/\alpha s_1s_2,-\alpha s_1s_2q^{1-s};q\right)_\infty q^{s^2}\left(s_1s_2q\right)^{-s}w(s)\,ds$$

$$-\frac{\alpha}{q}\frac{\left(-\alpha/s_3,-\alpha/s_4;q\right)_\infty}{\left(-\alpha s_1/q,-\alpha s_2/q;q\right)_\infty}{}_2\psi_2\left[\begin{matrix}-qs_3/\alpha, & -qs_4/\alpha,\\ -q^2/\alpha s_1, & -q^2/\alpha s_2\end{matrix};q,\frac{q}{s_1s_2s_3s_4}\right]$$

$$\cdot\int_0^1 \left(s_1q^{-s},q^{s+1}/s_1,s_2q^{-s},q^{s+1}/s_2;q\right)_\infty q^{s^2}\left(\frac{q}{s_1s_2}\right)^s w(s)\,ds.$$

Replacing s_1,s_2,s_3,s_4 by $q/a,q/b,q/c,q/d$, respectively, and setting $\alpha=1$, we obtain (1.26).

If we replace $q^{s^2}\left(q/s_1s_2\right)^s w(s)$ in (5.4) by the expression in (2.14) with a different α, then we will obtain the formula corresponding to this second choice. So the two are really equivalent.

6. An extension of (1.26). Let us now consider the rational-function extension of the integrand in (5.1) and take

$$\rho(s)=\frac{\left(q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_3,q^{1-s}/s_4,q^{1-s}/s_5;q\right)_\infty}{\left(q^{s+1}/\alpha s_1,\alpha s_1q^{-s},\alpha q^{s+1}/s_2,s_2q^{-s}/\alpha,s_6q^{-s};q\right)_\infty}w(s), \tag{6.1}$$

where $s_1s_2s_3s_4s_5s_6=q$, $w(s\pm1)=w(s)$, $\alpha s_1s_2\neq0$, and assume that there are no real poles. Then the integral that extends the one in (5.1) is

$$J(s_1,s_2,s_3,s_4,s_5)$$

$$=\int_{-\infty}^{\infty}\frac{\left(q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_3,q^{1-s}/s_4,q^{1-s}/s_5;q\right)_\infty}{\left(q^{s+1}/\alpha s_1,\alpha s_1q^{-s},\alpha q^{s+1}/s_2,s_2q^{-s}/\alpha,s_6q^{-s};q\right)_\infty}q^{-s}w(s)\,ds. \tag{6.2}$$

By (1.24) we get

$$J(s_1,s_2,s_3,s_4,s_5)$$

$$=\int_0^1 ds\,\frac{\left(q^{s+1}/s_1,q^{s+1}/s_2,q^{1-s}/s_3,q^{1-s}/s_4,q^{1-s}/s_5;q\right)_\infty}{\left(q^{s+1}/\alpha s_1,\alpha s_1q^{-s},\alpha q^{s+1}s_2,s_2q^{-s}/\alpha,s_6q^{-s};q\right)_\infty}$$

$$\cdot q^{-s}w(s)\,{}_3\psi_3\left[\begin{matrix}s_3q^s, & s_4q^s, & s_5q^s\\ q^{s+1}/s_1, & q^{s+1}/s_2, & q^{s+1}/s_6\end{matrix};q,q\right]. \tag{6.3}$$

Using the transformation formula [25, eq. 15_R] that can be deduced from [16, eq. (5.4.3)], we find that

$${}_3\psi_3\left[\begin{matrix}s_3q^s, & s_4q^s, & s_5q^s\\ q^{s+1}/s_1, & q^{s+1}/s_2, & q^{s+1}/s_6\end{matrix};q,q\right]$$

$$=\frac{\beta\left(s_1q^{-s},s_2q^{-s},s_6q^{-s},\beta q^{1-s}/s_3,\beta q^{1-s}/s_4,\beta q^{1-s}/s_5;q\right)_\infty}{\left(\beta s_1q^{-s},\beta s_2q^{-s},\beta s_6q^{-s},q^{1-s}/s_3,q^{1-s}/s_4,q^{1-s}/s_5;q\right)_\infty} \tag{6.4}$$

$$\cdot {}_3\psi_3\left[\begin{matrix} s_3q^s/\beta, & s_4q^s/\beta, & s_5q^s/\beta \\ q^{s+1}/\beta s_1, & q^{s+1}/\beta s_2, & q^{s+1}/\beta s_6 \end{matrix}; q, q\right]$$

$$+ \frac{(\beta, q/\beta, q, q/s_1s_3, q/s_1s_4, q/s_1s_5, q/s_2s_3, q/s_2s_4, q/s_2s_5; q)_\infty}{(s_3s_6, s_4s_6, s_5s_6, q^{s+1}/s_1, q^{s+1}/s_2, q^{1-s}/s_3, q^{1-s}/s_4, q^{1-s}/s_5; q)_\infty}$$

$$\cdot \frac{(s_6q^{-s}, \beta q^{-2s}s_1s_2, q^{2s+1}/\beta s_1s_2; q)_\infty}{(\beta s_1q^{-s}, q^{s+1}/\beta s_1, \beta s_2q^{-s}, q^{s+1}/\beta s_2; q)_\infty}$$

$$+ \frac{(s_3q^s, s_4q^s, s_5q^s; q)_\infty}{(q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6; q)_\infty}$$

$$\cdot \left\{ 1 - \beta^2 \frac{(s_3q^s/\beta, \beta q^{1-s}/s_3, s_4q^s/\beta, \beta q^{1-s}/s_4, s_5q^s/\beta, \beta q^{1-s}/s_5; q)_\infty}{(q^{s+1}/\beta s_1, \beta s_1q^{-s}, q^{s+1}/\beta s_2, \beta s_2q^{-s}, q^{s+1}/\beta s_6, \beta s_6q^{-s}; q)_\infty} \right.$$

$$\left. \cdot \frac{(q^{s+1}/s_1, q^{s+1}/s_2, q^{s+1}/s_6, s_1q^{-s}, s_2q^{-s}, s_6q^{-s}; q)_\infty}{(s_3q^s, q^{1-s}/s_3, s_4q^s, q^{1-s}/s_4, s_5q^s, q^{1-s}/s_5; q)_\infty} \right\}$$

$$\cdot \frac{s_6q^{-s}}{(1 - s_4s_6)(1 - s_5s_6)} {}_3\phi_2\left[\begin{matrix} q, & q/s_1s_3, & q/s_2s_3, \\ & s_4s_6q, & s_5s_6q \end{matrix}; q, s_3s_6\right],$$

where β is arbitrary, with $\beta s_1s_2s_6 \neq 0$. Replacing β by $-\beta q^s$ and substituting into (6.4), we obtain the formula

(6.5)

$$\int_{-\infty}^{\infty} \frac{(q^{s+1}/s_1, q^{s+1}/s_2, q^{1-s}/s_3, q^{1-s}/s_4, q^{1-s}/s_5; q)_\infty}{(q^{s+1}/\alpha s_1, \alpha s_1q^{-s}, \alpha q^{s+1}/s_2, s_2q^{-s}/\alpha, s_6q^{-s}; q)_\infty} w(s)\, q^{-s}\, ds$$

$$= \frac{(q, q/s_1s_3, q/s_1s_4, q/s_1s_5, q/s_2s_3, q/s_2s_4, q/s_2s_5; q)_\infty}{(-\beta s_1, -\beta s_2, -q/\beta s_1, -q/\beta s_2, s_3s_6, s_4s_6, s_5s_6; q)_\infty}$$

$$\cdot \int_0^1 \frac{(-\beta q^s, -q^{1-s}/\beta, -q^{s+1}/\beta s_1s_2, -\beta s_1s_2q^{-s}; q)_\infty}{(q^{s+1}/\alpha s_1, \alpha s_1q^{-s}, \alpha q^{s+1}/s_2, s_2q^{-s}/\alpha; q)_\infty} w(s)\, ds$$

$$- \beta \frac{(-\beta q/s_3, -\beta q/s_4, -\beta q/s_5; q)_\infty}{(-\beta s_1, -\beta s_2, -\beta s_6; q)_\infty} {}_3\psi_3\left[\begin{matrix} -s_3/\beta, & -s_4/\beta, & -s_5/\beta \\ -q/\beta s_1, & -q/\beta s_2, & -q/\beta s_6 \end{matrix}; q, q\right]$$

$$\cdot \int_0^1 \frac{(q^{s+1}/s_1, s_1q^{-s}, q^{s+1}/s_2, s_2q^{-s}; q)_\infty}{(q^{s+1}/\alpha s_1, \alpha s_1q^{-s}, \alpha q^{s+1}/s_2, s_2q^{-s}/\alpha; q)_\infty} w(s)\, ds$$

$$+ \frac{s_6}{(1 - s_4s_6)(1 - s_5s_6)} {}_3\phi_2\left[\begin{matrix} q, & q/s_1s_3, & q/s_2s_3 \\ & s_4s_6q, & s_5s_6q, \end{matrix}; q, s_3s_6\right]$$

$$\cdot \int_0^1 \left\{ \frac{(s_3q^s, q^{1-s}/s_3, s_4q^s, q^{1-s}/s_4, s_5q^s, q^{1-s}/s_5; q)_\infty}{(q^{s+1}/\alpha s_1, \alpha s_1q^{-s}, \alpha q^{s+1}/s_2, s_2q^{-s}/\alpha, q^{s+1}/s_6, s_6q^{-s}; q)_\infty} w(s)\, q^{-2s} \right.$$

$$-\beta^2\frac{(-s_3/\beta,-\beta q/s_3,-s_4/\beta,-\beta q/s_4,-s_5/\beta,-\beta q/s_5;q)_\infty}{(-\beta s_1,-q/\beta s_1,-\beta s_2,-q/\beta s_2,-\beta s_6,-q/\beta s_6;q)_\infty}$$

$$\cdot\frac{(q^{s+1}/s_1,s_1q^{-s},q^{s+1}/s_2,s_2q^{-s};q)_\infty}{(q^{s+1}/\alpha s_1,\alpha s_1q^{-s},\alpha q^{s+1}/s_2,s_2q^{-s}/\alpha;q)_\infty}w(s)\Bigg\}ds.$$

Since α and β are arbitrary other than the requirement that there are no poles in any of the preceding expressions, we can choose them to simplify this formula as much as possible. For example, if we choose $\alpha=-q/\beta s_1$ and then replace s_1,s_2,s_3,s_4,s_5, and s_6 by $q/a,q/b,q/c,q/d,q/e$, and q/f, respectively, so that $abcdef=q^5$, then (6.5) can be rewritten in the form

(6.6)

$$\int_{-\infty}^{\infty}\frac{(aq^s,bq^s,cq^{-s},dq^{-s},eq^{-s};q)_\infty}{(-\beta q^s,-q^{1-s}/\beta,-abq^{s-1}/\beta,-\beta q^{2-s}/ab,q^{1-s}/f;q)_\infty}w(s)\,ds$$

$$=\frac{(q,ac/q,ad/q,ae/q,bc/q,bd/q,be/q;q)_\infty}{(-\beta q/a,-a/\beta,-\beta q/b,-b/\beta,q^2/cf,q^2/df,q^2/ef;q)_\infty}\int_0^1 w(s)\,ds$$

$$-\frac{\beta(-\beta c,-\beta d,-\beta e;q)_\infty}{(-\beta q/a,-\beta q/b,-\beta q/f;q)_\infty}{}_3\psi_3\left[\begin{matrix}-q/\beta c, & -q/\beta d, & -q/\beta e\\ -a/\beta, & -b/\beta, & -f/\beta,\end{matrix};q,q\right]$$

$$\cdot\int_0^1\frac{(aq^s,q^{1-s}/a,bq^s,q^{1-s}/b;q)_\infty}{(-\beta q^s,-q^{1-s}/\beta,-abq^{s-1}/\beta,-\beta q^{2-s}/ab;q)_\infty}w(s)\,q^{-s}\,ds$$

$$+\frac{\frac{q}{f}(q,aq/f,bq/f;q)_\infty}{(q^2/cf,q^2/df,q^2/ef;q)_\infty}{}_3\phi_2\left[\begin{matrix}q^2/cf, & q^2/df, & q^2/ef\\ & aq/f, & bq/f\end{matrix};q,q\right]$$

$$\cdot\int_0^1\Bigg\{\frac{(q^{s+1}/c,cq^{-s},q^{s+1}/d,dq^{-s},q^{s+1}/e,eq^{-s};q)_\infty}{(-\beta q^s,-q^{1-s}/\beta,-abq^{s-1}/\beta,-\beta q^{2-s}/ab,fq^s,q^{1-s}/f;q)_\infty}w(s)$$

$$-\beta^2\frac{(-c\beta,-q/c\beta,-d\beta,-q/d\beta,-e\beta,-q/e\beta;q)_\infty}{(-a/\beta,-\beta q/a,-b/\beta,-\beta q/b,-f/\beta,-q\beta/f;q)_\infty}$$

$$\cdot\frac{(aq^s,q^{1-s}/a,bq^s,q^{1-s}/b;q)_\infty\,w(s)\,q^s}{(-\beta q^s,-q^{1-s}/\beta,-abq^{s-1}/\beta,-\beta q^{2-s}/ab;q)_\infty}\Bigg\}ds.$$

Note that we have used the transformation formula [16, eq. (3.2.10)] in rewriting the ${}_3\phi_2$ in series in (6.5) in the preceding balanced form. Also note that this formula is valid provided that

$$\text{(6.7)}\quad \begin{matrix}\text{(i)}\ \ abcdef=q^5, & \text{(ii)}\ \ \operatorname{Im} f\neq 0, & \text{(iii)}\ \ |q^2/f|<\min(|c|,|d|,|e|),\\ \text{(iv)}\ \ \beta ab\neq 0, & \text{(v)}\ \ 0<\arg\beta<\pi.\end{matrix}$$

Formula (6.6) may be regarded as a q-extension of [25, eq. (5.23)]. Observe that all the integrands on the right side of (6.6) are unit-periodic functions of s that are continuous on [0,1] whenever $w(s)$ is.

7. A Barnes-type integral: Askey–Roy formula. Let us now consider the integral

$$I_\alpha(s_1,s_2,s_3,s_4)=\frac{1}{2\pi i}\int_C \frac{(\alpha s_3 q^s, q^{1-s}/\alpha s_3, s_4 q^{s+1}/\alpha, \alpha q^{-s}/s_4; q)_\infty}{(s_3 q^s, s_4 q^s, s_1 q^{-s}, s_2 q^{-s}; q)_\infty}\,ds, \tag{7.1}$$

which corresponds to (2.19) and (2.22) with $w(s)=1$. Here C is the part of the imaginary axis from $-iT$ to iT, where $T=\pi/\log q^{-1}$, that defines the principal strip of the complex plane beyond which the integrand repeats its value by periodicity since $q^{s\pm 2ikT}=q^s$, $k=0,1,2,\dots$. We will assume that $\alpha s_3 s_4\neq 0$ and that $\max(|s_1|,|s_2|,|s_3|,|s_4|)<1$. Under these assumptions one can show by an analysis similar to the one given by Slater [29, Chap. 5] that the integral in (7.1) converges.

From (2.1) it follows that

$$\begin{aligned}\tau(s)\nabla x_1(s) &= \left(q^{-s}-s_3\right)\left(q^{-s}-s_4\right)-s_3s_4\left(1-s_1q^{-s}\right)\left(1-s_2q^{-s}\right)\\ &= \frac{(1-s_1s_3)(1-s_1s_4)}{s_1}q^{-s}-\frac{(1-s_1s_2s_3s_4)}{s_1}\left(1-s_1q^{-s}\right)q^{-s} \qquad (7.2)\\ &= -s_4(1-s_1s_3)(1-s_2s_3)q^{-s}+(1-s_1s_2s_3s_4)(1-s_3q^s)q^{-2s}. \qquad (7.3)\end{aligned}$$

Hence by (7.2)

$$\begin{aligned}\frac{1}{2\pi i}\int_C \rho_\alpha(s)\tau(s)\nabla x_1(s)\,ds &= \frac{(1-s_1s_3)(1-s_1s_4)}{s_1}I_\alpha(s_1,s_2,s_3,s_4)\\ &\quad -\frac{(1-s_1s_2s_3s_4)}{s_1}I_\alpha(s_1q,s_2,s_3,s_4),\end{aligned} \tag{7.4}$$

where

$$\rho_\alpha(s)=\frac{(\alpha s_3 q^s, q^{1-s}/\alpha s_3, s_4 q^{s+1}/\alpha, \alpha q^{-s}/s_4; q)_\infty}{(s_3 q^s, s_4 q^s, s_1 q^{-s}, s_2 q^{-s}; q)_\infty}q^s. \tag{7.5}$$

Since

$$\rho(s+1)\sigma(s+1)=s_3s_4\frac{(\alpha s_3 q^{s+1}, q^{-s}/\alpha s_3, s_4 q^{s+2}/\alpha, \alpha q^{-s-1}/s_4; q)}{(s_3 q^{s+1}, s_4 q^{s+1}, s_1 q^{-s}, s_2 q^{-s}; q)_\infty}q^{s+1}$$

has no poles between C and C', the line one unit to the left of C, because of the restrictions on the parameters, the integral $\int_C \triangle[\rho(s)\sigma(s)]\,ds$ vanishes, and so by (1.5) the integral on the left side of (7.4) also vanishes. This leads to the recurrence formula

$$I_\alpha(s_1,s_2,s_3,s_4)=\frac{1-s_1s_2s_3s_4}{(1-s_1s_3)(1-s_1s_4)}I_\alpha(s_1q,s_2,s_3,s_4). \tag{7.6}$$

By symmetry we also have

$$I_\alpha(s_1,s_2,s_3,s_4)=\frac{1-s_1s_2s_3s_4}{(1-s_2s_3)(1-s_2s_4)}I_\alpha(s_1,s_2q,s_3,s_4). \tag{7.7}$$

Use of (7.3) gives

$$I_\alpha(s_1,s_2,s_3,s_4)=\frac{1-s_1s_2s_3s_4}{(1-s_1s_3)(1-s_2s_3)}\left(-\frac{\alpha s_3}{s_4}\right)\cdot I_\alpha(s_1,s_2,s_3q,s_4). \tag{7.8}$$

Similarly,

$$I_\alpha(s_1,s_2,s_3,s_4)=\frac{1-s_1s_2s_3s_4}{(1-s_1s_4)(1-s_2s_4)}\left(-\frac{qs_4}{\alpha s_3}\right)\cdot I_\alpha(s_1,s_2,s_3,s_4q). \tag{7.9}$$

Finally, it follows directly from (7.1) that

$$I_\alpha(s_1,s_2,s_3,s_4)=\frac{\alpha^2 s_3}{s_4}I_{\alpha q}(s_1,s_2,s_3,s_4). \tag{7.10}$$

It follows from (7.6)–(7.10) that

$$\begin{aligned}I_\alpha(s_1,s_2,s_3,s_4)&=\frac{(\alpha,q/\alpha,\alpha s_3/s_4,qs_4/\alpha s_3,s_1s_2s_3s_4;q)_\infty}{(s_1s_3,s_1s_4,s_2s_3,s_2s_4;q)_\infty}\\&\quad\cdot M_\alpha(s_1,s_2,s_3,s_4),\end{aligned} \tag{7.11}$$

where M_α is symmetric in s_1,s_2 and satisfies the periodicity property

$$\begin{aligned}M_\alpha(s_1,s_2,s_3,s_4)&=M_\alpha(s_1q,s_2,s_3,s_4)\\&=M_\alpha(s_1,s_2q,s_3,s_4)=M_\alpha(s_1,s_2,s_3q,s_4)\\&=M_\alpha(s_1,s_2,s_3,s_4q)\\&=M_{\alpha q}(s_1,s_2,s_3,s_4).\end{aligned} \tag{7.12}$$

We shall first prove that M_α is actually independent of s_1 and s_2. Let k be a positive integer, and let us replace s_1 by s_1q^k in (7.11). Then by (7.12) we have

$$\begin{aligned}&\frac{(q^ks_1s_3,q^ks_1s_4;q)_\infty}{(q^ks_1s_2s_3s_4;q)}\frac{1}{2\pi i}\int_C\frac{(\alpha s_3q^s,q^{1-s}/\alpha s_3,s_4q^{s+1}/\alpha,\alpha q^{-s}/s_4;q)_\infty}{(s_1q^{k-s},s_2q^{-s},s_3q^s,s_4q^s;q)_\infty}\,ds\\&\quad=\frac{(\alpha,q/\alpha,\alpha s_3/s_4,qs_4/\alpha s_3;q)_\infty}{(s_2s_3,s_2s_4;q)_\infty}M_\alpha(s_1,s_2,s_3,s_4).\end{aligned} \tag{7.13}$$

Since the expression on the right side is independent of k, we conclude by use of the Lebesgue dominated convergence theorem that

$$\begin{aligned}&\frac{1}{2\pi i}\int_C\frac{(\alpha s_3q^s,q^{1-s}/\alpha s_3,s_4q^{s+1}/\alpha,\alpha q^{-s}/s_4;q)_\infty}{(s_2q^{-s},s_3q^s,s_4q^s;q)_\infty}\,ds\\&\quad=\frac{(\alpha,q/\alpha,\alpha s_3/s_4,qs_4/\alpha s_3;q)_\infty}{(s_2s_3,s_2s_4;q)_\infty}M_\alpha(s_1,s_2,s_3,s_4).\end{aligned} \tag{7.14}$$

It follows immediately that M_α is independent of s_1, and so by symmetry it is independent of s_2 as well.

Let us now consider the dependence of M_α on α. Replacing α by αq^k, k a positive integer, and observing that

$$\begin{aligned}&\frac{(\alpha s_3q^{s+k},q^{1-s-k}/\alpha s_3,s_4q^{s+1-k}/\alpha,\alpha q^{-s+k}/s_4;q)_\infty}{(\alpha q^k,q^{1-k}/\alpha,\alpha s_3q^k/s_4,s_4q^{1-k}/\alpha s_3;q)_\infty}\\&\quad=\frac{(\alpha s_3q^s,q^{1-s}/\alpha s_3,s_4q^{s+1}/\alpha,\alpha q^{-s}/s_4;q)_\infty}{(\alpha,q/\alpha,\alpha s_3/s_4,s_4q/\alpha s_3;q)_\infty},\end{aligned} \tag{7.15}$$

we conclude that M_α is also independent of α. It follows that

$$M_\alpha(s_1,s_2,s_3,s_4)$$

$$= \frac{(s_1s_3,s_1s_4,s_2s_3,s_2s_4;q)_\infty}{(\alpha,q/\alpha,\alpha s_3/s_4,qs_4/\alpha s_3;q)_\infty\,(s_1s_2s_3s_4;q)_\infty} \tag{7.16}$$

$$\cdot\frac{1}{2\pi i}\int_C \frac{(\alpha s_3q^s,q^{1-s}/\alpha s_3,s_4q^{s+1}/\alpha,\alpha q^{-s}/s_4;q)_\infty}{(s_3q^s,s_4q^s,s_1q^{-s},s_2q^{-s};q)_\infty}\,ds$$

is actually independent of s_1, s_2 and α. So we may set, for example, $s_1 = \alpha/s_4$ and $s_2 = q/\alpha s_3$ in (7.16) to cancel out some terms in the integrand, and we get

$$(q;q)_\infty M_\alpha = \frac{1}{2\pi i}\int_C \frac{(\alpha s_3q^s,s_4q^{s+1}/\alpha;q)_\infty}{(s_3q^s,s_4q^s;q)_\infty}\,ds. \tag{7.17}$$

Let us now replace s_3, s_4 by s_3q^k, s_4q^k, respectively, k a positive integer. Recalling that M_α is periodic in s_3 and s_4, we find that

$$\begin{aligned}(q;q)_\infty M_\alpha &= \frac{1}{2\pi i}\lim_{k\to\infty}\int_C \frac{(\alpha s_3q^{s+k},s_4q^{s+1+k}/\alpha;q)_\infty}{(s_3q^{s+k},s_4q^{s+k};q)_\infty}\,ds\\ &= \frac{1}{2\pi i}\int_C ds\\ &= \frac{1}{2\pi}\int_{-\pi/\log q^{-1}}^{\pi/\log q^{-1}} dy\\ &= \frac{1}{\log q^{-1}},\end{aligned} \tag{7.18}$$

where the integration is done over the interval of length $T = 2\pi/\log q^{-1}$. Thus we have the formula

$$\begin{aligned}&\frac{1}{2\pi i}\int_C \frac{(\alpha s_3q^s,q^{1-s}/\alpha s_3,s_4q^{s+1}/\alpha,\alpha q^{-s}/s_4;q)_\infty}{(s_3q,s_4q,s_1q^{-s},s_2q^{-s};q)_\infty}\,ds\\ &\quad= \frac{(\alpha,q/\alpha,\alpha s_3/s_4,qs_4/\alpha s_3,s_1s_2s_3s_4;q)_\infty}{(q,s_1s_3,s_1s_4,s_2s_3,s_2s_4;q)_\infty\log q^{-1}}.\end{aligned} \tag{7.19}$$

With a change of variable $s = i\theta/\log q^{-1}$ and a relabeling of the parameters one can easily deduce (1.27) from (7.19).

8. An extension of the Askey–Roy integral: Gasper's formula. We now consider the integral

$$J_\alpha(s_1,s_2,s_3,s_4,s_5)$$

$$= \frac{1}{2\pi i}\int_C \frac{(\alpha q^s/s_1,s_1q^{1-s}/\alpha,q^{s+1}/\alpha s_2,\alpha s_2q^{-s},q^{s+1}/s_6;q)_\infty}{(s_3q^s,s_4q^s,s_5q^s,s_1q^{-s},s_2q^{-s};q)_\infty}\,ds, \tag{8.1}$$

where C is the same as in §7, $s_1s_2s_3s_4s_5s_6 = q$. From (2.2) one finds that

$$\tau(s)\nabla x_1(s) = -\frac{(1-s_1s_5)(1-s_2s_5)}{s_1s_2s_5}q^{-s} + \frac{(1-s_1s_2s_3s_5)(1-s_1s_2s_4s_5)}{s_1s_2s_5}\frac{1-s_5q^s}{1-q^{s+1}/s_6}q^{-s}, \tag{8.2}$$

so that we have the recurrence formula

$$J_\alpha(s_1,s_2,s_3,s_4,s_5) = \frac{(1-s_1s_2s_3s_5)(1-s_1s_2s_4s_5)}{(1-s_1s_5)(1-s_2s_5)}J_\alpha(s_1,s_2,s_3,s_4,qs_5). \tag{8.3}$$

Iterating this n times and then taking the limit $n\to\infty$, we find that

$$J_\alpha(s_1,s_2,s_3,s_4,s_5) = \frac{(s_1s_2s_3s_5,s_1s_2s_4s_5;q)_\infty}{(s_1s_5,s_2s_5;q)_\infty}\lim_{n\to\infty}J_\alpha(s_1,s_2,s_3,s_4,s_5q^n). \tag{8.4}$$

Since $\lim_{n\to\infty}\left(q^{s+1+n}/s_6;q\right)_\infty/\left(s_5q^{n+s};q\right)_\infty = 1$ for $s\in C$, we have

$$\begin{aligned}\lim_{n\to\infty}J_\alpha(s_1,s_2,s_3,s_4,s_5q^n) &= \frac{1}{2\pi i}\int_C \frac{\left(\alpha q^s/s_1, s_1q^{1-s}/\alpha, \alpha s_2q^{-s}, q^{s+1}/\alpha s_2;q\right)_\infty}{\left(s_1q^{-s},s_2q^{-s},s_3q^s,s_4q^s;q\right)_\infty}ds\\ &= \frac{\left(\alpha,q/\alpha,\alpha s_2/s_1,s_1q/\alpha s_2,s_1s_2s_3s_4;q\right)_\infty}{\left(q,s_1s_3,s_1s_4,s_2s_3,s_2s_4;q\right)_\infty\log q^{-1}}\end{aligned} \tag{8.5}$$

by (7.19). Thus we have the formula

$$\frac{1}{2\pi i}\int_C \frac{\left(\alpha q^s/s_1, s_1q^{1-s}/\alpha, q^{s+1}/\alpha s_2, \alpha s_2q^{-s}, q^ss_1s_2s_3s_4s_5;q\right)_\infty}{\left(s_1q^{-s},s_2q^{-s},s_3q^s,s_4q^s,s_5q^s;q\right)_\infty}ds = \frac{\left(\alpha,q/\alpha,\alpha s_2/s_1,s_1q/\alpha s_2,s_1s_2s_3s_4,s_1s_2s_3s_5,s_1s_2s_4s_5;q\right)_\infty}{\log q^{-1}\left(q,s_1s_3,s_1s_4,s_1s_5,s_2s_3,s_2s_4,s_2s_5;q\right)_\infty}. \tag{8.6}$$

Changing the variable by setting $s = i\theta/\log q^{-1}$, we get Gasper's formula [15]:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left(\alpha e^{-i\theta}/s_1, qs_1e^{i\theta}/\alpha, \alpha s_2e^{i\theta}, qe^{-i\theta}/\alpha s_2, s_1s_2s_3s_4s_5e^{-i\theta};q\right)_\infty}{\left(s_1e^{i\theta},s_2e^{i\theta},s_3e^{-i\theta},s_4e^{-i\theta},s_5e^{-i\theta};q\right)_\infty}d\theta = \frac{\left(\alpha,q/\alpha,\alpha s_2/s_1,qs_1/\alpha s_2,s_1s_2s_3s_4,s_1s_2s_3s_5,s_1s_2s_4s_5;q\right)_\infty}{\left(q,s_1s_3,s_1s_4,s_1s_5,s_2s_3,s_2s_4,s_2s_5;q\right)_\infty}. \tag{8.7}$$

REFERENCES

[1] G. E. ANDREWS AND R. ASKEY (1981), *Another q-extension of the beta function*, Proc. Amer. Math. Soc., 81, pp. 97–100.

[2] R. ASKEY (1981), *A q-extension of Cauchy's form of the beta integral*, Quart. J. Math. Oxford (Ser. 2), 32, pp. 255–266.

[3] ——— (1985), *Continuous Hahn polynomials*, J. Phys. A, 19, pp. L 1017–L 1019.

[4] ——— (1988), *Beta integrals in Ramanujan's papers, his unpublished work and further examples*, in Ramanujan Revisited, G. E. Andrews et. al., eds., Academic Press, New York, pp. 561–590.

[5] ——— (1988), *Beta integrals and q-extensions*, J. Ramanujan Math. Soc., Special issue: Proc. Ramanujan Centennial International Conference, Annamalainagar, India, Dec. 15–17, 1987, pp. 85–102.

[6] ——— (1989), *Beta integrals and the associated orthogonal polynomials*, in *Number Theory*, K. Alladi, ed., Lecture Notes in Mathematics, Vol. 1395, Springer-Verlag, New York, pp. 84–121.

[7] A. ASKEY AND R. ROY (1986), *More q-beta integrals*, Rocky Mountain J. Math., 16, pp. 365–372.

[8] A. ASKEY AND J. A. WILSON (1985), *Some basic hypergeometric polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc, 319, pp. 1–55.

[9] N. M. ATAKISHIYEV AND S. K. SUSLOV (1992), *On the Askey–Wilson polynomials*, Const. Approx., 8, pp. 363–369.

[10] N. M. ATAKISHIYEV, M. RAHMAN AND S. K. SUSLOV (1992), *A definition and a classification of the classical orthogonal polynomials*, submitted.

[11] E. W. BARNES (1908), *A new development of the theory of hypergeometric functions*, Proc. London Math. Soc. 2, 6, pp. 141–177.

[12] ——— (1910), *A transformation of generalized hypergeometric series*, Quart. J. Math., 14, pp. 136–140.

[13] A. -L. CAUCHY (1825), *Sur les intégrales définies prises entre des limites imaginaires*, Bulletin de Ferussoc, T. III, pp. 214–221; Oeuvres de Cauchy, 2^e série, T. II, Gauthier–Villars, Paris, 1958, pp. 57–66.

[14] J. DOUGALL (1907), *On Vandermonde's theorem and some more general expansions*, Proc. Edinburgh Math. Soc., 25, pp. 114–132.

[15] G. GASPER (1989), *q-Extensions of Barnes', Cauchy's, and Euler's beta integrals*, in Topics in Mathematical Analysis, T. M. Rassias, ed., World Scientific, Singapore, pp. 294–314.

[16] G. GASPER AND M. RAHMAN (1990), *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, UK.

[17] E. G. KALNINS AND W. MILLER (1988), *q-Series and orthogonal polynomials associated with Barnes' first lemma*, SIAM J. Math. Anal., 19, pp. 1216–1231.

[18] ——— (1989), *Symmetry techniques for q-series: Askey–Wilson polynomials*, Rocky Mountain J. Math., 19, pp. 1–8.

[19] A. F. NIKIFOROV AND S. K. SUSLOV (1985), *Systems of Classical Orthogonal Polynomials of a Discrete Variable on Nonuniform Lattices*, preprint 8, Keldysh Institute of Applied Mathematics, Moscow. (In Russian.)

[20] A. F. NIKIFOROV AND V. B. UVAROV (1983), *Classical Orthogonal Polynomials of a Discrete Variable on Nonuniform Lattices*, preprint 17, Keldysh Institute of Applied Mathematics, Moscow. (In Russian.)

[21] ——— (1987), *Fundamentals of the Theory of Classical Orthogonal Polynomials of a Discrete Variable*, preprint 56, Keldysh Institute Applied Mathematics, Moscow. (In Russian.)

[22] ——— (1987), *Construction of q-Analogues of Classical Orthogonal Polynomials of a Discrete Variable on Nonuniform Lattices*, preprint 179, Keldysh Institute Applied Mathematics, Moscow. (In Russian.)

[23] A. F. NIKIFOROV, S. K. SUSLOV AND V. B. UVAROV (1982), *Racah Polynomials and Dual Hahn Polynomials as a Generalization of Classical Orthogonal Polynomials of a Discrete Variable*, preprint 165, Keldysh Institute of Applied Mathematics, Moscow. (In Russian.)

[24] ——— (1985), *Classical Orthogonal Polynomials of a Discrete Variable*, Nauka, Moscow (in Russian); English translation, Springer-Verlag, Berlin, 1991.

[25] M. RAHMAN AND S. K. SUSLOV (1993), *The Pearson equation and the beta integrals*, SIAM J. Math. Anal., 25, pp. 000–000.

[26] ——— (1993), *Classical biorthogonal rational functions*, in Methods of Approximation Theory in Complex Analysis and Mathematical Physics, A. A. Gonchar and E. B. Saff, eds., Lecture Notes in Math., Vol. 1550, Springer-Verlag, Berlin, pp. 131–146.

[27] S. RAMANUJAN (1920), *A class of definite integrals*, Quart. J. Math., 48, pp. 294–310; reprinted in *Collected Papers of Srinivasa Ramanujan*, G. H. Hardy, P. V. Seshu Aiyar, and B. M. Wilson, eds., Cambridge University Press, Cambridge, UK, 1927; reprinted by Chelsea, New York, 1962.

[28] A. A. SAMARSKII (1977), *Teoriya Raznostnykh Skhem* (*Theory of Difference Schemes*), Nauka, Moscow. (In Russian.)

[29] L. J. SLATER (1966), *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, UK.

[30] S. K. SUSLOV (1989), *The theory of difference analogues of special functions of hypergeometric type*, Russian Math. Surveys, 44 (2), pp. 227–278.

[31] G. N. WATSON (1910), *The continuation of functions defined by generalized hypergeometric series*, Trans. Cambridge Philos. Soc., 21, pp. 281–299.

ACKNOWLEDGMENT

The March issue of the *SIAM Journal on Mathematical Analysis* was a collection of contributions in special functions dedicated to Profs. R. Askey and F. Olver. The papers in this special issue were collected and edited by

G. Gasper, Editor-in-Chief, Northwestern University,
G. Andrews, Pennsylvania State University,
M. Ismail, University of Southern Florida,
P. Nevai, Ohio State University.

We at SIAM are all grateful to these editors for their hard work, dedication, and high professional standards, which made the Askey–Olver Special Issue such a success.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1023–1068, July 1994

VALIDITY OF THE QUASIGEOSTROPHIC MODEL FOR LARGE-SCALE FLOW IN THE ATMOSPHERE AND OCEAN*

ALFRED J. BOURGEOIS†‡ AND J. THOMAS BEALE†

Abstract. The well-known quasigeostrophic system (QGS) for zero Rossby number flow has been used extensively in oceanography and meteorology for modeling and forecasting mid-latitude oceanic and atmospheric circulation. Formulation of QGS requires a (singular) perturbation expansion of a set of primitive equations at small Rossby number, and the quasigeostrophic equation expresses conservation of the zero-order potential vorticity of the flow. The formal expansion is justified by investigating the behavior of solutions of a set of primitive equations (PE) with a particular scaling, in the limit of zero Rossby number. This primitive model represents adiabatic, inviscid, incompressible flow with variable density and Coriolis force. Difficulties arise because PE, scaled for small Rossby number, contains unwanted solutions varying on a fast time scale with frequencies inversely proportional to the Rossby number. Without restrictions on the initial conditions, solutions of the scaled PE model do not necessarily converge to solutions of QGS in the singular limit. It is proven that, provided certain simple restrictions on the initial data are satisfied, solutions of QGS are valid approximations of solutions of the scaled PE model, with error on the order of the Rossby number. Going further in the PE expansion, the first correction to the QGS solution is obtained and it is shown that the improved approximation is second order accurate. The essential part of the analysis is to obtain energy estimates for the ageostrophic part of the solutions which allow suppression of the rapid growth. A new proof of the existence of solutions of QGS is also given.

Key words. geophysical fluid flow, quasigeostrophic equations, Rossby number, beta plane, bounded derivative method

AMS subject classifications. 86A05, 35B25, 76U05, 76V05, 86A10

1. Introduction.

The well-known quasigeostrophic system (QGS) has been used extensively in oceanography and meteorology for modeling and forecasting mid-latitude oceanic and atmospheric circulation. In this paper we prove that solutions of a quasigeostrophic model approximate solutions to a model for large-scale flows with order ε accuracy as the Rossby number ε goes to zero.

By large-scale (or global) flow we mean solutions of a primitive model for oceanic or atmospheric flow in which the equations of motion have been scaled for small Rossby number motion. The traditional primitive equations for planetary circulations [19] represent adiabatic, inviscid, incompressible flow with variable density and Coriolis force, with additional assumptions that the flow is Boussinesq and hydrostatic. The Boussinesq approximation, which says that density (or potential temperature, in the atmospheric case) variation is small compared to some reference value, allows the horizontal momentum equations to be written without the density (potential temperature) variable. Instead of the usual hydrostatic assumption, in our model we retain the vertical acceleration term in the third momentum equation. Keeping this term preserves the symmetry of the equations and defines a better-posed problem. We later make a scaling choice resulting in a near hydrostatic model. So our primitive equations (PE) differ from the typical primitive equations in this respect, and are defined by (2.1)–(2.5), where we have chosen the oceanic case for definiteness. In the atmospheric case, a modified pressure function is used instead of vertical height.

* Received by the editors July 27, 1992; accepted for publication (in revised form) March 15, 1993.

† Mathematics Department, Duke University, Durham, North Carolina, 27708. This research was supported by National Science Foundation grant DMS-9102782.

‡ Present address, Lawrence Livermore National Laboratory, University of California, Livermore, California 94551.

Informally, large-scale motion refers to the major current systems of the ocean (e.g., the Gulf Stream) and atmosphere (e.g., the Jet Stream), which vary on slow time scales relative to one rotation period of the earth. Prediction of such global phenomena is a fundamental problem in oceanography and meteorology. This type of motion is primarily horizontal (measured relative to the earth's surface) because on large-scale, the fluid is confined to a relatively thin spherical shell. Also, the density stratification resulting from the near hydrostatic equilibrium discourages vertical motion. The essential scales for describing large-scale motion are horizontal velocity and horizontal length. Since relative accelerations become small over large length scales, the horizontal acceleration is largely Coriolis, which is independent of length scale. Thus global flow can be thought of as flow in which the Coriolis force is significant, and is defined formally as small Rossby number flow, where the Rossby number is the ratio of the relative horizontal acceleration to the Coriolis acceleration. This ratio is small for motion with large time scale compared to the earth's rotation period.

Formulation of QGS [20] requires a perturbation expansion of PE at small Rossby number. The resulting zero-order equations express geostrophic and hydrostatic balance, in which the zero-order horizontal velocity is perpendicular to the zero-order horizontal pressure gradient. These equations are not sufficient to determine flow dynamics, and the first-order equations in ε are necessary to describe the time evolution of the zero-order flow. Elimination of the first-order terms in the first-order equations leads to a nonlinear evolution equation in zero-order pressure. This is the quasigeostrophic equation, which is a conservation law for the zero-order potential vorticity of the flow. See, e.g., [12] for the use of this equation to study the baroclinic instability mechanism.

The quasigeostrophic equation, with initial data and boundary conditions specifying no normal flow and determining a mathematically well-posed system that determines the time evolution of the zero-order flow. This system is simpler than PE because it filters out fast time scale solutions. Previous existence and uniqueness proofs for quasigeostophic systems have been given. Dutton [13] demonstrated existence of weak solutions on a bounded domain, and showed solutions are continuously dependent on the initial data. Bennett and Kloeden [3], [5] established existence of strong solutions.

For our system, we choose a horizontally periodic domain, as in [3], and a simplified boundary condition that expresses constant density on the top and bottom horizontal boundaries. We refer to this system as QGS. The object of this paper is to justify the formal expansion mentioned above, by investigating the behavior of solutions of PE in the limit of zero Rossby number ε, and establishing conditions in which these solutions converge to the QGS solution. Difficulties arise because PE scaled for small Rossby number contains solutions varying on a fast time scale t/ε and a slow time scale t. Without restrictions on the initial conditions, solutions do not converge in the singular limit $\varepsilon = 0$. We show that if solutions and their first time derivatives are initially bounded independently of ε, then solutions of PE exist on an arbitrarily long time interval for small enough ε, and these solutions converge to QGS solutions with $O(\varepsilon)$ accuracy. To carry the expansion further, we derive equations for the time evolution of the first-order correction and show that the QGS solutions with this correction give $O(\varepsilon^2)$ accurate solutions of PE.

Analogous results have been proven by Schochet [23] for the quasigeostrophic approximation of the shallow water model with no density stratification and no dependence on the vertical variable, which is a hyperbolic system. Browning, Kreiss,

and Kasahara [8], [9] have devised a general initialization procedure, referred to as the bounded derivative method, for suppressing fast time solutions of linear hyperbolic systems of partial differential equations containing multiple time scales. This method has been extended to certain nonlinear symmetric hyperbolic systems [1]. However, PE is nonlinear and nonhyperbolic. Klainerman and Majda [16], [17] have given a rigorous treatment of a variety of related singular limits with rapid scales. In recent work [7], Browning et al. have viewed the system we call PE as a reduced system for slowing down gravity for compressible flow. Camassa and Holm [10] use a similar set of starting equations for their new model of barotropic mesoscale ocean dynamics, which incorporates dispersive effects due to weak hydrostatic imbalance in the presence of topography and stratification.

1.1. Outline. In §2 we present the standard formulation of the quasigeostrophic equation for an ocean model. We begin with the governing primitive equations (PE), which are given in (2.1)–(2.5). Then we discuss how PE are scaled for small Rossby number flow, and the resulting scaled primitive equations (SPE) are listed in (2.6)–(2.10). Finally, we outline the standard derivation of QGS from SPE.

In §3 we prove an existence theorem (Theorem 3.4) for QGS. This theorem is customized for use in our main theorem comparing QGS and SPE solutions. The proof is new, and is patterned after a proof by Kato and Ponce for well-posedness of the Euler equations [15]. It entails the use of a special L^∞ estimate on the second derivatives of pressure in terms of vorticity [2], [15].

Our main result, described above, is given in §4 as Theorem 4.5. It essentially says that SPE solutions converge with order ε accuracy to QGS solutions in the limit of zero Rossby number, if SPE is initialized so that fast solutions are suppressed. The bulk of the proof of Theorem 4.5 is in obtaining an energy estimate for the ageostrophic velocity and density that allows growth only on the long time scale.

The Rossby number for global circulations can be large enough that the QGS soluti on has significant error, and a correction term may be useful in such settings. In §5 we derive an equation (5.10) for the first-order correction to the QGS solution. This equation plays the same role at the first order as the vorticity equation in the QGS model. Finally, in §6, we verify the accuracy of the formal approximation to first order, obtained by adding this first correction term to the QGS solution. We prove that the improved approximation is within order ε^2 of the SPE solution, again with assumptions on the initial data.

A fairly high degree of regularity is assumed in order to estimate nonlinear terms. It is likely that this degree could be reduced with further effort. Of course, fairly general initial data can be approximated by smooth data.

1.2. Preliminaries. We use B for the open rectangular box in R^3 described by

$$B = \left\{ (x, y, z) : -\tfrac{1}{2} < x < \tfrac{1}{2},\ -\tfrac{1}{2} < y < \tfrac{1}{2},\ 0 < z < h \right\}.$$

The region B physically represents the principal flow region for our ocean models, and the surfaces $z = h$ and $z = 0$ represent the top and bottom boundaries of the ocean, respectively.

The standard multi-index notation $D^\alpha = \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \partial_{x_3}^{\alpha_3}$ denotes higher-order derivatives, where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and $|\alpha| = \sum_{j=1}^3 \alpha_j$.

We use $L^2(B)$ for the Lebesgue space of square-integrable functions over B, with

the corresponding norm

$$|f|_0 = \left\{ \int_B f^2(\mathbf{x})\, d\mathbf{x} \right\}^{1/2},$$

and inner product

$$\Big(f, g\Big)_{L_2} = \int_B f(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x}\,.$$

We write $H^s(B)$ for the Sobolev space of horizontally periodic functions in B, with generalized derivatives up to order s belonging to $L^2(B)$. We will also have need of the Sobolev space $H^{s,per}(B_p)$ of fully periodic functions in the periodic box

$$B_p = \left\{ (x, y, z) : -\tfrac{1}{2} < x < \tfrac{1}{2}, -\tfrac{1}{2} < y < \tfrac{1}{2}, -h < z < h \right\}.$$

The spaces $H^{s,per}_{even}(B_p)$ and $H^{s,per}_{odd}(B_p)$ are $H^{s,per}(B_p)$ functions that are even and odd, respectively, in z. The following theorem describes the correspondence between $H^s(B)$ and the even and odd $H^{s,per}(B_p)$ spaces.

THEOREM 1.1. *For integers s, a function f in $H^s(B)$ can be extended to $H^{s,per}_{even}(B_p)$ if and only if all odd z derivatives of f with index less than s are zero on the boundaries $z = 0$ and $z = h$ of B. A function f in $H^s(B)$ can be extended to $H^{s,per}_{odd}(B_p)$ if and only if all even z derivatives of f with index less than s (including f itself) are zero on the boundaries $z = 0$ and $z = h$.*

We use

$$|f|_s = \sum_{|\alpha| \le s} |D^\alpha f|_0$$

for the $H^s(B)$ norm. However, $|f|_\infty$ will denote the norm in $L^\infty(B)$, that is, $\sup_{\mathbf{x}\in B} |f(\mathbf{x})|$. We write $C([0,T]; H^s(B))$, with corresponding norm

$$|f|_{s,T} = \sup_{0 \le t \le T} |f|_s$$

for the space of functions $f(\mathbf{x}, t)$ on $B \times [0,T]$, continuous in time with values in $H^s(B)$.

In our estimates, C is used as a generic constant and may change from line to line.

2. The quasigeostrophic equation for an ocean model. In this section we briefly describe the standard formulation of the quasigeostrophic equation for the motion of a stratified fluid on a rotating sphere. A comprehensive discussion can be found in the text of Pedlosky [20], where the quasigeostrophic equation is derived in spherical coordinates. A systematic derivation of the quasigeostrophic equation in Cartesian coordinates is presented in [22]. Here we derive the quasigeostrophic equation from an ocean model in a similar way. Our governing primitive equations (PE) of motion for oceanic flow are

$$\frac{Du}{Dt} - fv = -\varphi_x\,, \tag{2.1}$$

$$\frac{Dv}{Dt} + fu = -\varphi_y\,, \tag{2.2}$$

$$\frac{Dw}{Dt} + \frac{g}{\varrho_0}\varrho = -\varphi_z \,, \tag{2.3}$$

$$\nabla\cdot\mathbf{u} = 0 \,, \tag{2.4}$$

$$\frac{D\varrho}{Dt} = 0 \,. \tag{2.5}$$

In these equations t is time, the Cartesian coordinates x,y, and z are directed eastward, northward, and upward, respectively, and $\mathbf{u} = (u, v, w)$ are the corresponding velocities. $\nabla = (\partial_x, \partial_y, \partial_z)$ is the gradient operator and D/Dt is the material derivative $\partial_t + \mathbf{u}\cdot\nabla$. The Coriolis parameter f is $2\Omega\sin\theta$, where θ is latitude and $\Omega = 7.3 \times 10^{-5}\text{sec}^{-1}$ is the angular velocity of the earth. The Boussinesq approximation has been made, and the potential φ equals p/ϱ_0, where p is pressure and ϱ_0 is a reference density. We define $\bar{\varrho} = \bar{\varrho}(z)$ to be the (known) background density profile, which we assume to satisfy $\bar{\varrho}_z < 0\,, 0 \le z \le h$, and $\rho = \rho(x, y, z, t)$ to be the departure from $\bar{\varrho}$. Similarly, we let ϕ represent the departure from the background potential $\bar{\varphi}(z)$. Therefore we have

$$\varrho(x, y, z, t) = \bar{\varrho}(z) + \rho(x, y, z, t)\,, \qquad \varphi(x, y, z, t) = \bar{\varphi}(z) + \phi(x, y, z, t)\,.$$

2.1. The scaled primitive equations (SPE). The primitive equations are made nondimensional with the following characteristic scales:

$$x = Lx' \,, \quad y = Ly' \,, \quad z = Hz' \,, \quad t = \frac{L}{U}t' \,,$$

$$u = Uu' \,, \quad v = Uv' \,, \quad w = \frac{UH}{L}w' \,,$$

$$\bar{\varrho} = P\bar{\varrho}' \,, \quad \rho = \frac{\varrho_0 f_0 UL}{gH}\rho' \,, \quad \phi = f_0 UL\phi' \,.$$

The Rossby number $\varepsilon = U/f_0 L$ is the fundamental ordering parameter in the forthcoming asymptotic expansion. A secondary ordering parameter is the scale ratio of ρ to $\bar{\varrho}$. This ratio is assumed to be ε, that is, we assume

$$\frac{\varrho_0 f_0 UL}{gH} = P\varepsilon \,.$$

Thus the density expressed in the nondimensional quantities $\bar{\varrho}'$ and ρ' is

$$\varrho = P(\bar{\varrho}'(z) + \varepsilon\rho') \,.$$

We make the usual β-plane approximation $f = f_0 + \hat{\beta}_0 y$, where f_0 is the Coriolis parameter at a central latitude θ_0 of the region, $y = (\theta - \theta_0)r_0$ is the northward Cartesian coordinate, r_0 is the earth's radius, and $\hat{\beta}_0 = 2\Omega\cos\theta_0/r_0$ is the northward gradient of f at θ_0. In accordance with this approximation for f, we assume that L/r_0 is $O(\varepsilon)$. Then defining

$$\beta_0 = \frac{\cot\theta_0}{\varepsilon}\frac{L}{r_0},$$

we write

$$f = f_0\left(1 + \frac{\hat{\beta}_0 L}{f_0}y'\right) = f_0(1 + \varepsilon\beta_0 y') \,.$$

Substituting the above scales into (PE) and leaving off primes, we have

$$\varepsilon \frac{Du}{Dt} - (1 + \varepsilon\beta_0 y)\, v = -\phi_x \,, \tag{2.6}$$

$$\varepsilon \frac{Dv}{Dt} + (1 + \varepsilon\beta_0 y)\, u = -\phi_y \,, \tag{2.7}$$

$$\varepsilon\delta^2 \frac{Dw}{Dt} + \rho = -\phi_z \,, \tag{2.8}$$

$$\nabla\cdot\mathbf{u} = 0 \,, \tag{2.9}$$

$$\varepsilon \frac{D\rho}{Dt} + w\bar{\varrho}_z = 0 \,. \tag{2.10}$$

In (2.8), the parameter δ represents the scale ratio H/L. We consider the case in which the ratio δ approaches a fixed positive value in the limit as ε approaches zero, so without loss of generality we assume $\delta = 1$. We will refer to the nondimensional equations (2.6)–(2.10) as SPE (scaled PE). In §4 we prove an existence theorem for these equations in a rectangular box with rigid upper and lower boundaries.

2.2. The quasigeostrophic system (QGS). QGS is derived by performing a singular perturbation expansion of SPE over ε. The region of flow is defined by $0 < z < h$, with the rigid boundary condition $w = 0$ on the upper and lower surfaces $z = h$ and $z = 0$. Assuming solutions $\mathbf{U}(\varepsilon) \equiv (\mathbf{u}(\varepsilon), \rho(\varepsilon))$, and substituting the formal expansion $\mathbf{U}(\varepsilon) = \mathbf{U}^{(0)} + \varepsilon\mathbf{U}^{(1)} + \varepsilon^2\tilde{\mathbf{U}}(\varepsilon)$ into SPE, the zero-order equations in ε are

$$v^{(0)} = \phi_x^{(0)} \,, \quad u^{(0)} = -\phi_y^{(0)} \,, \quad \rho^{(0)} = -\phi_z^{(0)} \,, \quad w^{(0)} = 0 \,. \tag{2.11}$$

Incompressibility and the fact that the zero-order horizontal velocity is divergence free imply $w_z^{(0)} = 0$. The last equation in (2.11) then follows from the rigid boundary condition. Equations (2.11) are referred to as the geostrophic equations.

The first-order equations in ε are needed to determine the evolution of $\phi^{(0)}$. We will use the notation

$$d_g \equiv \partial_t + \mathbf{u}\cdot\nabla = \partial_t + u^{(0)}\partial_x + v^{(0)}\partial_y$$

for the zero-order (geostrophic) material derivative. Then the first-order equations in ε are

$$d_g u^{(0)} - \beta_0 y v^{(0)} - v^{(1)} = -\phi_x^{(1)} \,, \tag{2.12}$$

$$d_g v^{(0)} + \beta_0 y u^{(0)} + u^{(1)} = -\phi_y^{(1)} \,, \tag{2.13}$$

$$\rho^{(1)} = -\phi_z^{(1)} \,, \tag{2.14}$$

$$\nabla\cdot\mathbf{u}^{(1)} = 0 \,, \tag{2.15}$$

$$d_g\rho^{(0)} + w^{(1)}\bar{\varrho}_z = 0\,. \tag{2.16}$$

The idea now is to eliminate $\phi^{(1)}$ between (2.12) and (2.13) by cross differentiation, then to eliminate the resulting first-order horizontal divergence $u_x^{(1)} + v_y^{(1)}$ with (2.15) and (2.16). Taking the two-dimensional curl of the first-order horizontal momentum equations (2.12) and (2.13), and using $u_x^{(0)} + v_y^{(0)} = 0$, as well as (2.15), we find

$$d_g\Big(v_x^{(0)} - u_y^{(0)}\Big) - w_z^{(1)} + \beta_0 v^{(0)} = 0. \tag{2.17}$$

Defining $\lambda(z) \equiv -1/\bar{\varrho}_z$, and assuming $\lambda(z)$ is bounded away from zero on $[0,h]$, we have from (2.16),

$$w_z^{(1)} = \Big(\lambda d_g\rho^{(0)}\Big)_z = \Big(d_g(\lambda\rho^{(0)})\Big)_z\,, \tag{2.18}$$

since $d_g\lambda = 0$. Expanding the last term in (2.18), we have

$$\Big(d_g(\lambda\rho^{(0)})\Big)_z = d_g(\lambda\rho^{(0)})_z + u_z^{(0)}(\lambda\rho^{(0)})_x + v_z^{(0)}(\lambda\rho^{(0)})_y\,. \tag{2.19}$$

Using the zero-order equations (2.11), we can rewrite the last two terms on the right-hand side of (2.19) as $-\lambda u_z^{(0)} v_z^{(0)}$ and $\lambda v_z^{(0)} u_z^{(0)}$, respectively. So (2.19) simplifies to

$$\Big(d_g(\lambda\rho^{(0)})\Big)_z = d_g(\lambda\rho^{(0)})_z\,, \tag{2.20}$$

and (2.18) becomes

$$w_z^{(1)} = d_g(\lambda\rho^{(0)})_z\,. \tag{2.21}$$

Plugging (2.21) into (2.17) results in

$$d_g\Big(v_x^{(0)} - u_y^{(0)} - (\lambda\rho^{(0)})_z + \beta_0 y\Big) = 0\,. \tag{2.22}$$

Equation (2.22) represents conservation of potential vorticity

$$v_x^{(0)} - u_y^{(0)} - (\lambda\rho^{(0)})_z + \beta_0 y$$

along material paths. Expressing everything in terms of $\phi^{(0)}$, we have

$$\Big(\partial_t - \phi_y^{(0)}\partial_x + \phi_x^{(0)}\partial_y\Big)\Big(\phi_{xx}^{(0)} + \phi_{yy}^{(0)} + (\lambda\phi_z^{(0)})_z + \beta_0 y\Big) = 0\,. \tag{2.23}$$

Equation (2.23) is the well-known quasigeostrophic equation. QGS consists of (2.23) with initial and boundary conditions on $\phi^{(0)}$. From (2.16) and the third equation in (2.11), the condition on $\phi^{(0)}$ for rigid boundaries is $d_g\phi_z^{(0)} = 0$, since this forces $w^{(1)} = 0$. Short-time existence of solutions that are periodic in x and y and satisfy the rigid horizontal boundary condition $d_g\phi_z^{(0)} = 0$ on the upper and lower boundaries was shown by Bennett and Kloeden in [3]. The quasigeostrophic model with periodic horizontal boundary conditions has been used, for example, by Charney [11] for analytic studies of geostrophic turbulence, and by Bretherton [6] for numerical mid-ocean modeling. We will use periodic horizontal boundary conditions with the simpler, rigid, upper

and lower boundary condition $\phi_z^{(0)}=0$. This simpler (Neumann) boundary condition has been used in [13] and [5] to demonstrate long-time existence of solutions for a reduced system called the simplified quasigeostrophic equations. This simpler boundary condition is equivalent to the more general boundary condition $d_g\phi_z^{(0)}=0$, if we assume $\phi_z^{(0)}$ is initially zero on the boundary. In §3 we provide a new proof for the existence of QGS solutions satisfying this simpler boundary condition.

3. An existence proof for QGS. In this section we will prove an existence theorem for the quasigeostrophic equation in a rectangular box with periodic horizontal boundary conditions and rigid horizontal boundaries at top and bottom. We define the box-shaped region $B=\Sigma\times(0,h)$, where $\Sigma=(-\frac{1}{2},\frac{1}{2})\times(-\frac{1}{2},\frac{1}{2})$, and Σ_z will be used to denote the horizontal cross-section $\Sigma\times\{z\}$. All functions and partial derivatives considered here are assumed periodic with period 1 in both horizontal directions for each fixed z and t. For each time t a constant can be added to the pressure potential $\phi^{(0)}$ without affecting the system, so we will assume that $\int_B\phi^{(0)}=0$. We saw in §2 that rigid horizontal boundaries Σ_0 and Σ_h imply that the zero-order vertical velocity $w^{(0)}$ in region B is identically zero. From (2.16), rigid boundary conditions on Σ_0 and Σ_h imply $d_g\rho^{(0)}=0$ on these boundaries. We impose the stronger (Neumann) boundary condition $\rho^{(0)}=0$. Omitting superscripts, we write the periodic QGS as

$$v=\phi_x\,, \tag{3.1}$$

$$u=-\phi_y\,, \tag{3.2}$$

$$\rho=-\phi_z\,, \tag{3.3}$$

$$v_x-u_y-(\lambda\rho)_z=\omega \quad \text{in } B\,, \tag{3.4}$$

$$\omega_t+\mathbf{u}\cdot\nabla\omega=-\beta_0 v \quad \text{in } B\times[0,T]\,, \tag{3.5}$$

$$\rho=0 \quad \text{on } \Sigma_0\times[0,T] \quad \text{and} \quad \Sigma_h\times[0,T]\,, \tag{3.6}$$

$$\omega(\mathbf{x},t=0)=\omega_0(\mathbf{x}) \quad \text{in } B \text{ at } t=0\,, \tag{3.7}$$

where $\mathbf{x}=(x,y$ and $[0,T]$ is the time interval $0\le t\le T$. We will prove a global existence theorem for the periodic QGS problem (3.1)–(3.7). All functions in this section are zero-order terms from the asymptotic expansion of SPE discussed in §2. In this section, for example, $\mathbf{u}$ represents the zero-order velocity $(u^{(0)},v^{(0)},0)$. We leave off the superscripts here because superscripts are reserved for sequence iterations in the upcoming existence proof.

3.1. Local existence of solutions. We begin with the following short-time existence theorem.

THEOREM 3.1 (QGS short-time existence). *If the initial vorticity ω_0 is in $H^s(B)$ for some $s\ge 3$, with $|\omega_0|_s\le M$, then there exists a time $T^*>0$ and a solution ω in*

$C([0,T^*];H^s(B))$ *to* QGS, *where* T^* *is defined by* (3.18) *below and depends only on* M, B, λ, *and* β_0. *The vorticity* ω *satisfies the estimate* $\|\omega\|_{s,T^*} \le 2M$.

Proof. From (3.1)–(3.4) and (3.6), notice that ϕ is determined by ω at each time t through the boundary value problem

$$\begin{aligned} &\phi_{xx} + \phi_{yy} + (\lambda\phi_z)_z = \omega \quad \text{in } B\,, \\ &\phi_z = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h\,, \\ &\int_B \phi = 0\,. \end{aligned} \tag{3.8}$$

Let ϕ_0 be the initial pressure corresponding to ω_0. Then consider the iteration for $k = 0, 1, 2, \ldots,$

$$\begin{cases} \phi^k_{xx} + \phi^k_{yy} + (\lambda\phi^k_z)_z = \xi^k \quad \text{in } B\,, \\ \phi^k_z = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h\,, \\ \displaystyle\int_B \phi^k = 0\,, \end{cases} \tag{3.9}$$

$$u^k = -\phi^k_y\,, \qquad v^k = \phi^k_x\,, \tag{3.10}$$

$$\xi^{k+1}_t + \mathbf{u}^k \cdot \nabla\xi^{k+1} = -\beta_0 v^k, \qquad \xi^{k+1}_0 \equiv \omega_0\,, \tag{3.11}$$

where $\xi^0(t) \equiv \omega_0$ and $\phi^0(t) \equiv \phi_0$. With each iteration k, the kth vorticity iterate ξ^k is used to solve the Neumann problem (3.9). The kth velocity iterates are determined from ϕ^k by (3.10), and these velocities are used in (3.11) to get the $(k+1)$th vorticity iterate ξ^{k+1}. We assume that the initial vorticity ω_0 satisfies the compatibility condition $\int_B \omega_0 = 0$ for the initial Neumann problem. It can be checked that this condition persists in time under (3.11), i.e., $\int_B \xi^{k+1} dx$ has time derivative zero, and therefore remains zero. This is necessary so that ϕ^k can be determined from (3.9).

We will need the following lemma.

LEMMA 3.2. *If the initial vorticity* ω_0 *is in* $H^s(B)$ *for some* $s \ge 3$, *and* $|\omega_0|_s \le M$, *there exists a time* $T' > 0$ *defined below by* (3.17) *and depending only on* M, B, λ, *and* β_0, *such that for each nonnegative integer* k, *the* k*th vorticity iterate* ξ^k *is in* $C([0,T'];H^s(B))$ *and satisfies* $|\xi^k(t)|_s \le 2M$ *for* $0 < t \le T'$.

Proof. Since $|\xi^0(t)|_s = |\omega_0|_s \le M$, the statement in the lemma is true for $k = 0$. Arguing by induction on k, we assume $|\xi^k|_s \le 2M$ for some $k > 0$. From elliptic theory and (3.9), we have the estimate

$$|\phi^k|_{s+2} \le C_0 |\xi^k|_s\,,$$

where the positive constant C_0 depends only on B and λ. Now from (3.1), (3.2), and our induction hypothesis, we can write

$$|\mathbf{u}^k|_{s+1} \le |\phi^k|_{s+2} \le 2C_0 M\,. \tag{3.12}$$

The linear vorticity equation

$$\xi^{k+1}_t + \mathbf{u}^k \cdot \nabla\xi^{k+1} = -\beta_0 v^k \tag{3.13}$$

has the unique solution ξ^{k+1} in $C([0,T'];H^s(B))$; this degree of regularity is evident from the following discussion. We differentiate (3.13) by D^α, $|\alpha| \le s$, and write ξ^k_α for $D^\alpha \xi^k$ to get

$$\frac{\partial \xi^{k+1}_\alpha}{\partial t} + D^\alpha(\mathbf{u}^k \cdot \nabla \xi^{k+1}) = -\beta_0 v^k_\alpha ,$$

which we may rewrite as

$$\frac{\partial \xi^{k+1}_\alpha}{\partial t} + \mathbf{u}^k \cdot D^\alpha \nabla \xi^{k+1} = -F_\alpha - \beta_0 v^k_\alpha , \tag{3.14}$$

where

$$F_\alpha = D^\alpha(\mathbf{u}^k \cdot \nabla \xi^{k+1}) - \mathbf{u}^k \cdot D^\alpha \nabla \xi^{k+1} .$$

Since $\nabla \cdot \mathbf{u} = 0$, this is equivalent to

$$F_\alpha = D^\alpha \nabla \cdot (\mathbf{u}^k \xi^{k+1}) - \mathbf{u}^k \cdot D^\alpha \nabla \xi^{k+1} .$$

We will estimate F_α with the calculus inequality

$$|D^\alpha(fg) - fD^\alpha g|_0 \le C\Big(|f|_s \, |g|_\infty + |\nabla f|_\infty \, |g|_{s-1} \Big) . \tag{3.15}$$

This inequality is well known for $f \in H^s \cap C^1$ and $g \in H^s \cap C^0$. We need it here with $g \in H^{s-1} \cap C^0$. The proof of inequality (3.15) can be found in [16] and is based on the Gagliardo–Nirenberg inequalities. It can be modified with a passage to the limit to show that it holds for $g \in H^{s-1} \cap C^0$. (Notice that the individual terms on the left side of (3.15) may not be in L^2.) Letting $f = \mathbf{u}^k$, $g = \xi^{k+1}$, and replacing D^α with $D^\alpha \nabla$ and s with $s+1$, we require $\mathbf{u}^k \in H^{s+1}(B) \cap C^1(B)$ and $\xi^{k+1} \in H^s(B) \cap C^0(B)$. By Sobolev's lemma, this holds provided $s > 3/2$; therefore, from (3.15) we get

$$|F_\alpha|_0 \le C\Big(|\mathbf{u}^k|_{s+1} \, |\xi^{k+1}|_\infty + |\nabla \mathbf{u}^k|_\infty \, |\xi^{k+1}|_s \Big) .$$

Again by Sobolev's lemma, since $s \ge 3$, there is a positive constant C_1 such that

$$|\xi^{k+1}|_\infty \le C_1 |\xi^{k+1}|_s \quad \text{and} \quad |\nabla \mathbf{u}^k|_\infty \le C_1 |\mathbf{u}^k|_{s+1} .$$

Combining these estimates and (3.12) results in

$$|F_\alpha|_0 \le C_2 |\xi^{k+1}|_s , \tag{3.16}$$

where we have defined $C_2 \equiv 4MCC_0C_1$. Multiplying (3.14) by ξ^{k+1}_α and integrating over B gives us

$$\frac{1}{2}\frac{d}{dt}|\xi^{k+1}_\alpha|^2_0 + \left(\mathbf{u}^k \cdot \nabla \xi^{k+1}_\alpha , \xi^{k+1}_\alpha\right)_{L^2} = \left(-F_\alpha , \xi^{k+1}_\alpha\right)_{L^2} - \left(\beta_0 v^k_\alpha , \xi^{k+1}_\alpha\right)_{L^2} .$$

The second term is zero by the divergence theorem and the periodic boundary conditions. Now taking absolute values, applying the Schwarz inequality, summing over α, $|\alpha| \le s$, and using estimates (3.12) and (3.16), we have

$$\frac{d}{dt}|\xi^{k+1}|_s \le C_2 \, |\xi^{k+1}|_s + C_3 ,$$

where $C_3 \equiv 2\beta_0 C_0 M$. Recalling $\xi_0^{k+1} \equiv \omega_0$, we solve this differential inequality to get

$$|\xi^{k+1}|_s \le \Big(M + \frac{C_3}{C_2}\Big)e^{C_2 t} - \frac{C_3}{C_2}\,.$$

We want to choose $T'(M)$ so that if $0 \le t \le T'$, this remains less than $2M$. By defining

$$T' = \frac{1}{C_2}\ln\Big(1 + \frac{M}{M + (C_3/C_2)}\Big)\,, \tag{3.17}$$

we see that $|\xi^{k+1}(t)|_s \le 2M$ when $0 \le t \le T'$, and Lemma 3.2 is proved. □

Now we find a time interval for which the sequence of vorticity iterates $\{\xi^k(t)\}$ converges. From (3.13), the linear vorticity equation for the difference $\xi^{k+1} - \xi^k$ is

$$\frac{\partial}{\partial t}(\xi^{k+1} - \xi^k) + \mathbf{u}^k\cdot\nabla(\xi^{k+1} - \xi^k) = -(\mathbf{u}^k - \mathbf{u}^{k-1})\cdot\nabla\xi^k - \beta_0(v^{k+1} - v^k)\,.$$

Multiplying by $\xi^{k+1} - \xi^k$, integrating over B, and applying the divergence theorem results in

$$\frac{1}{2}\frac{d}{dt}|\xi^{k+1} - \xi^k|_0^2 = -\Big((\mathbf{u}^k - \mathbf{u}^{k-1})\cdot\nabla\xi^k - \beta_0(v^{k+1} - v^k)\,,\ \xi^{k+1} - \xi^k\Big)_{L^2};$$

therefore, we have

$$\frac{d}{dt}|\xi^{k+1} - \xi^k|_0 \le (\beta_0 + |\nabla\xi^k|_\infty)\,|\mathbf{u}^k - \mathbf{u}^{k-1}|_0 \le (\beta_0 + |\nabla\xi^k|_\infty)\,|\xi^{k+1} - \xi^k|_0\,.$$

By Sobolev's lemma, since $s \ge 3$ and $\xi^k \in H^s(B)$, there is a positive constant C_4 such that $|\nabla\xi^k|_\infty \le C_4|\xi^k|_s$; therefore, we can write

$$\frac{d}{dt}|\xi^{k+1} - \xi^k|_0 \le C_4\,(\beta_0 + |\xi^k|_s)\,|\xi^k - \xi^{k-1}|_0\,.$$

By Lemma 3.2, $|\xi^k|_s \le 2M$ for $t \le T'$ and we have

$$\frac{d}{dt}|\xi^{k+1} - \xi^k|_0 \le (2M + \beta_0)C_4\,|\xi^k - \xi^{k-1}|_0,$$

provided that $t \le T'$. Finally, we solve this differential inequality to get

$$\|\xi^{k+1} - \xi^k\|_{0,T} \le (2M + \beta_0)C_4 T\,\|\xi^k - \xi^{k-1}\|_{0,T}$$

for any time T such that $0 < T \le T'$. Choosing $T^* > 0$ to satisfy

$$T^* < \min\Big\{T'\,,\ [(2M + \beta_0)C_4]^{-1}\Big\}\,, \tag{3.18}$$

we have

$$\|\xi^{k+1} - \xi^k\|_{0,T^*} \le \gamma\,\|\xi^k - \xi^{k-1}\|_{0,T^*}, \qquad \gamma < 1,$$

which implies that $\{\xi^k\}$ is a Cauchy sequence in $C([0,T^*];H^0(B))$. Let $\omega \in C([0,T^*];H^0(B))$ be the limit of $\{\xi^k\}$. We need to show that ω is in $C([0,T^*];H^s(B))$.

For $0 \le s' < s$, we have the Sobolev inequality

$$|\xi^k - \xi^l|_{s'} \le C\,|\xi^k - \xi^l|_0^{1-s'/s}\,|\xi^k - \xi^l|_s^{s'/s} \le C(4M)^{s'/s}|\xi^k - \xi^l|_0^{1-s'/s},$$

where we have again used Lemma 3.2. Since $\{\xi^k\}$ is Cauchy in $C([0,T^*];H^0(B))$, the above inequality shows that $\{\xi^k\}$ is Cauchy in $C([0,T^*];H^{s'}(B))$, for all s' with $0 \le s' < s$. Thus ω is in $C([0,T^*];H^{s-1}(B))$. It follows from the convergence of the ξ^k and elliptic theory that ϕ^k converges in $C([0,T^*];H^{s+1}(B))$ to a limit ϕ, and $\mathbf{u}^k$ converges in $C([0,T^*];H^s(B))$ to a limit $\mathbf{u}$. It is now a simple matter to pass to the limit in (3.13) and conclude that (3.5) is satisfied, along with (3.1)–(3.4). Thus ω is a solution of the QGS initial value problem.

For each t such that $0 < t \le T^*$, $\{\xi^k(t)\}$ is bounded in $H^s(B)$ and therefore contains a weakly convergent subsequence $\{\xi^{k_n}(t)\}$, which must converge to $\omega(t)$. Then we necessarily have

$$|\omega(t)|_s \le \liminf |\xi^{k_n}(t)|_s \le 2M \,,$$

and can conclude that ω is in $L^\infty([0,T^*];H^s(B))$. It remains to show the continuity of $|\omega(t)|_s$ in time. We will use vorticity equation (3.5) and the method of characteristics.

Differentiating (3.5) by D^α, $|\alpha| \le s$, we get

$$\frac{\partial \omega_\alpha}{\partial t} + \mathbf{u}\cdot\nabla\omega_\alpha = -F_\alpha \,, \tag{3.19}$$

where

$$F_\alpha = D^\alpha\nabla\cdot(\mathbf{u}\omega) - \mathbf{u}\cdot D^\alpha\nabla\omega + \beta_0 v_\alpha \,.$$

As before, since $\mathbf{u}$ is in $H^{s+1}(B)\bigcap C^1(B)$ and ω is in $H^s(B)\bigcap C^0(B)$, the calculus inequality (3.15) shows that $|F_\alpha|_{0,T^*}$ is bounded. Now by the method of characteristics and the Duhamel principle, the solution of (3.19) is

$$\begin{aligned}\omega_\alpha(\mathbf{x},t) = {} & \omega_{0\alpha}\Big(X(0;\mathbf{x},t)\,,\, Y(0;\mathbf{x},t)\,,\, z\Big) \\ & - \int_0^t F_\alpha\Big(X(\tau;\mathbf{x},t)\,,\, Y(\tau;\mathbf{x},t)\,,\, z\,,\, \tau\Big)\, d\tau \,,\end{aligned}$$

where $(X(\tau;\mathbf{x},t)\,,\, Y(\tau;\mathbf{x},t)\,,\, z)$ is the position at past time τ of a particle with position $\mathbf{x}$ at time t. That is, for a fixed time $\tau < T^*$ and position $\mathbf{x}$, the curve $(X(\tau;\mathbf{x},t), Y(\tau;\mathbf{x},t), z)$ is uniquely determined by the two-dimensional system of ordinary differential equations

$$\frac{dX}{dt} = u(X,Y,z,t)\,, \qquad \frac{dY}{dt} = v(X,Y,z,t)\,,$$

with initial condition

$$\Big(X(\tau;\mathbf{x},\tau), Y(\tau;\mathbf{x},\tau), z\Big) = \mathbf{x}\,.$$

Therefore, for any two times t and t' less than T^*, we have

$$\begin{aligned}|\omega_\alpha(t) - \omega_\alpha(t')|_0 \;\le\; & |t-t'|\,\sup_{0\le t\le T^*}|F_\alpha(t)|_0 \\ & + \Big|\omega_{0\alpha}\Big(X(0;\mathbf{x},t)\,,\, Y(0;\mathbf{x},t)\,,\, z\Big) \\ & - \omega_{0\alpha}\Big(X(0;\mathbf{x},t')\,,\, Y(0;\mathbf{x},t')\,,\, z\Big)\Big|_0 \,.\end{aligned} \tag{3.20}$$

We showed above that $|F_\alpha|_{0,T^*}$ is bounded, so the first term after the inequality in (3.20) goes to zero as t' approaches t. For the second term after the inequality, we use the fact that the set of continuous functions on the closure of B is dense in $L^2(B)$. Since the distance between the two points is bounded by $|t-t'|\,\|\mathbf{u}\|_{\infty,T^*}$, we see that if $\omega_{0\alpha}$ is a continuous function on the closure of B, then the second term after the inequality approaches zero as t' approaches t. Since arbitrary $\omega_{0\alpha}$ can be approximated as closely as desired in $L^2(B)$ by such a function, it follows that the same is true in general. Hence ω_α is in $C([0,T^*];H^0(B))$ for $|\alpha|\le s$, and therefore ω is in $C([0,T^*];H^s(B))$. Since Lemma 3.2 guarantees $\|\xi^{k_n}\|_{s,T^*}\le 2M$ for each n, then ω must also satisfy $\|\omega\|_{s,T^*}\le 2M$. This completes the proof of Theorem 3.1. □

3.2. Global existence of solutions. To show that solutions of QGS exist up to arbitrary time T, we need a global estimate on the QGS vorticity ω. Getting such a global estimate requires the use of a special L^∞ estimate (see (3.25) below) on the second derivatives of the QGS pressure ϕ. This estimate is derived in Appendix A and results from the fact that the QGS pressure ϕ is determined by the QGS vorticity ω through the elliptic problem (3.8).

LEMMA 3.3. *There exists a continuous function $K(t)$ defined for $0\le t<\infty$ that depends only on $|\omega_0|_s$, such that if $\omega\in H^s(B)$, some $s\ge 3$, is a solution of* QGS *for $0<t\le T$, then*

$$|\omega(t)|_s \le K(t), \qquad 0\le t\le T\,. \tag{3.21}$$

Proof. We will first use the α-derivative vorticity equation (3.19) to estimate $|\omega(t)|_s$ in terms of bounds on the initial data ω_0 and density and velocity gradients. At each time t, $0\le t\le T$, the vorticity ω defined by (3.4) determines the pressure ϕ through the boundary value problem (3.8). We assume that ω_0 is given so that $\int_B \omega_0 = 0$. We noted at the beginning of the proof of Theorem 3.1 that this condition persists in time, so that the compatibility condition $\int_B \omega = 0$ for (3.8) is always satisfied. From (3.8) we get the elliptic estimate

$$|\mathbf{u}|_{s+1} + |\rho|_{s+1} \le C\,|\omega|_s\,. \tag{3.22}$$

Applying the calculus inequality (3.15) to F_α in (3.19) for $|\alpha|\le s$, we have

$$|F_\alpha|_0 \le C\left(|\mathbf{u}|_{s+1}\,|\omega|_\infty + |\nabla\mathbf{u}|_\infty\,|\omega|_s\right) + \beta_0|\mathbf{u}|_s\,.$$

Then using estimate (3.22) and the fact that $|\omega|_\infty \le |\nabla\mathbf{u}|_\infty + |\nabla\rho|_\infty$, we get

$$|F_\alpha|_0 \le C\left(\beta_0 + |\nabla\mathbf{u}|_\infty + |\nabla\rho|_\infty\right)|\omega|_s\,. \tag{3.23}$$

Multiplying (3.19) by ω_α, integrating over B, and using estimate (3.23) results in

$$\frac{d}{dt}|\omega|_s \le C\left(\beta_0 + |\nabla\mathbf{u}|_\infty + |\nabla\rho|_\infty\right)|\omega|_s\,.$$

Upon solving, we have

$$|\omega|_s \le C\,|\omega_0|_s \exp\left\{C\int_0^t \left(|\nabla\mathbf{u}(\tau)|_\infty + |\nabla\rho(\tau)|_\infty\right)d\tau\right\}, \tag{3.24}$$

where C depends on β_0 and T. To complete the argument, we will use a time-independent estimate for $|\nabla \mathbf{u}|_\infty + |\nabla \rho|_\infty$ in terms of the L^∞ norm of the initial vorticity, with a slight dependence on the H^s norm of vorticity. In Appendix A we derive the estimate

$$|\nabla \mathbf{u}|_\infty + |\nabla \rho|_\infty \leq C\,|\omega|_\infty \left(1 + \log^+ \frac{|\omega|_s}{|\omega|_\infty}\right), \tag{3.25}$$

where C is a universal constant and $\log^+ a = \log a$ if $a \geq 1$, $\log^+ a = 0$ otherwise. Estimate (3.25) follows from the boundary value problem (3.8), and its derivation is similar to the one presented in [2] for the free space problem. We use (3.25) here in the same way Kato and Ponce [15] use a similar estimate to prove existence of global solutions for the two-dimensional Euler equations. In the right side of (3.25) we want an L^∞ estimate for ω in terms of ω_0 on the finite time interval $0 \leq t \leq T$. From the vorticity equation (3.5) it can be shown that

$$|\omega(t)|_\infty \leq C|\omega_0|_\infty \leq C|\omega_0|_s, \qquad 0 \leq t \leq T\,. \tag{3.26}$$

Now $b(1 + \log^+ a/b)$ is monotone increasing in b, so we can use (3.26) in estimate (3.25) to get

$$|\nabla \mathbf{u}|_\infty + |\nabla \rho|_\infty \leq C\,|\omega_0|_s \left(1 + \log^+ \frac{|\omega|_s}{|\omega_0|_s}\right).$$

Now using estimate (3.24) to eliminate $|\omega|_s$ results in

$$|\nabla \mathbf{u}(t)|_\infty + |\nabla \rho(t)|_\infty \leq C\,|\omega_0|_s \left\{1 + \int_0^t \left(|\nabla \mathbf{u}(\tau)|_\infty + |\nabla \rho(\tau)|_\infty\right) d\tau\right\}.$$

By the Gronwall lemma, we then have

$$|\nabla \mathbf{u}(\tau)|_\infty + |\nabla \rho(\tau)|_\infty \leq C\,|\omega_0|_s\, e^{C|\omega_0|_s t}\,.$$

Substituting this in the right-hand side of (3.24), we find

$$|\omega(t)|_s \leq C\,|\omega_0|_s \exp\left\{C\,(e^{C|\omega_0|_s t} - 1)\right\} \equiv K(t)\,. \tag{3.27}$$

Notice $K(t)$ is defined for all time t, and depends only on $|\omega_0|_s$. This completes the proof of Lemma 3.3. □

We now demonstrate that long-time solutions to QGS exist.

THEOREM 3.4 (QGS global existence). *If ω_0 is in $H^s(B)$ for some $s \geq 3$, s arbitrary, then given any time $T > 0$, there exists a solution $\omega \in C([0,T]; H^s(B))$ to* QGS.

Proof. By the short-time existence Theorem 3.1, there is a time $T_1(|\omega_0|_s) > 0$ and solution $\omega \in C([0,T_1]; H^s(B))$ to QGS. If T_1 is greater than or equal to T, we have the desired solution. Otherwise, define

$$K_T = \max_{0 \leq t \leq T} K(t),$$

where $K(t)$ is defined by 3.27. Then $|\omega(T_1)|_s$ is bounded above by K_T, and we again invoke the short-time existence Theorem 3.1 to continue the solution to time, say, $T_1 + T_2$. Notice that T_2 depends only on the global bound K_T. Again, if $T_1 + T_2$ is

greater than or equal to T, we have the desired solution. Otherwise, by the global estimate (3.21), we have

$$|\omega(T_1 + T_2)|_s \leq K_T,$$

and by the short-time existence theorem we can continue the solution to time $T_1 + 2T_2$. We can repeat this argument until $T_1 + nT_2$ is greater than T. □

3.3. Initial conditions on the boundary. With restrictions on λ, QGS solutions ω have the property that if ω_z is initially zero on the boundary, it remains so. We want to establish this result for use in our comparison of QGS and SPE solutions in the next section, where we assume that $\lambda = 1$ (linear density profile) and that the QGS data satisfy the boundary condition $\omega_{0z} = 0$ on Σ_0 and Σ_h. To this end, we look at a periodic quasigeostrophic system (PQGS) consisting of equations (3.1)–(3.5) and (3.7), in the periodic domain

$$B_p = \left\{ (x, y, z) : -\tfrac{1}{2} < x < \tfrac{1}{2}, -\tfrac{1}{2} < y < \tfrac{1}{2}, -h < z < h \right\}. \tag{3.28}$$

All PQGS functions are assumed periodic in z with period $2h$, with the same horizontal periodicity as QGS functions. With only minor modification to our QGS proof, we have the following existence theorem for PQGS.

THEOREM 3.5 (PQGS global existence). *If ω_0^{per} is in $H^{s,per}_{even}(B_p)$ for some $s \geq 3$, then given any time $T > 0$, there exists a periodic solution $\omega^{per} \in C([0,T]; H^{s,per}_{even}(B_p))$ to* PQGS. *The corresponding* PQGS *pressure $\phi^{per} \in C([0,T]; H^{s+2,per}_{even}(B_p))$ is determined by ω^{per}, at each time t, $0 \leq t \leq T$, through the periodic elliptic problem*

$$\Delta\phi^{per} = \omega^{per} \quad \text{in } B_p\,, \qquad \int_{B_p} \phi^{per} = 0\,.$$

In the following discussion, we will apply Theorem 3.5 to QGS solutions for the special case $\lambda = 1$ and

$$\omega_{0z} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h. \tag{3.29}$$

We see from (3.8) that the initial pressure ϕ_0 is determined by ω_0 from

$$\begin{aligned} &\Delta\phi_0 = \omega_0 \quad \text{in } B\,, \\ &\phi_{0z} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h, \\ &\int_B \phi_0 = 0\,. \end{aligned} \tag{3.30}$$

Then the initial boundary condition (3.29) can be written as

$$\phi_{0zzz} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h. \tag{3.31}$$

Let $\phi \in C([0,T]; H^5(B))$ be the QGS solution with initial data ϕ_0. Now (3.31) and (3.6) imply that we can make the even periodic extension $\phi_0^{per} \in H^{5,per}_{even}(B_p)$ from ϕ_0 (cf. Theorem 1.1). Let $\phi^{per} \in C([0,T]; H^{5,per}_{even}(B_p))$ be the PQGS solution with initial data ϕ_0^{per}, guaranteed by Theorem 3.5. Then by the uniqueness of QGS solutions, ϕ^{per} must agree with ϕ in B. Notice from equations (3.1)–(3.3) that the PQGS velocity

and density corresponding to ϕ^{per} are in $H^{4,per}_{even}(B_p)$ and $H^{4,per}_{odd}(B_p)$, respectively. We summarize with the following corollary to Theorem 3.5.

COROLLARY 3.6. *Suppose we are given ϕ_0 in $H^5(B)$ such that*

$$\phi_{0z} = \phi_{0zzz} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h. \tag{3.32}$$

For $\lambda = 1$, let $\phi \in C([0,T]; H^5(B))$ be the QGS *solution with initial data ϕ_0. Then for each time t, $0 \le t \le T$, the* QGS *pressure ϕ satisfies*

$$\phi_z = \phi_{zzz} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h. \tag{3.33}$$

More generally, this holds for λ satisfying $\lambda_z = 0$. With this restriction on λ, a consequence of the corollary is that if ω_z is initially zero on the boundary, it remains zero there. Notice we could just as well have expressed (3.33) as $\rho = \rho_{zz} = 0$ on Σ_0 and Σ_h, due to the QGS identity $\rho \equiv \phi_z$.

4. The QGS solution as an approximation of SPE solutions. In this section we show that SPE solutions converge with $O(\varepsilon)$ accuracy to QGS solutions if fast solutions of SPE are appropriately suppressed. We will prove this for the case of a linear density profile and zero β-factor. The SPE system with initial and periodic horizontal boundary conditions is then

$$\varepsilon \frac{Du}{Dt} - v = -\phi_x \quad \text{in } B \times [0,T], \tag{4.1}$$

$$\varepsilon \frac{Dv}{Dt} + u = -\phi_y \quad \text{in } B \times [0,T], \tag{4.2}$$

$$\varepsilon \frac{Dw}{Dt} + \rho = -\phi_z \quad \text{in } B \times [0,T], \tag{4.3}$$

$$\varepsilon \frac{D\rho}{Dt} - w = 0 \quad \text{in } B \times [0,T], \tag{4.4}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } B \times [0,T], \tag{4.5}$$

$$\mathbf{u}(\mathbf{x}, t=0) = \mathbf{u}_0(\mathbf{x}) \quad \text{in } B, \tag{4.6}$$

$$\rho(\mathbf{x}, t=0) = \rho_0(\mathbf{x}) \quad \text{in } B, \tag{4.7}$$

$$w = 0 \quad \text{on } \Sigma_0 \times [0,T] \quad \text{and} \quad \Sigma_h \times [0,T]. \tag{4.8}$$

We assume $\int_B \phi = 0$ for uniqueness of solutions. Short-time solutions for this system can be shown to exist by making a slight modification (for the Coriolis terms) of a proof due to Valli in [25]. The ε factor can be treated as a time scale, and we are guaranteed short-time solutions of SPE up to some time $T^*(\varepsilon) > 0$ for each fixed value of ε. For each ε, the time T^* depends only on a bound on the H^s norm of the initial velocity and density. We state this formally in the following theorem.

THEOREM 4.1 (SPE short-time existence). *If initial data* $\mathbf{U}_0 \equiv (\mathbf{u}_0, \rho_0)$ *is in* $H^s(B)$ *for some* $s \geq 3$, *with* $|\mathbf{U}_0|_s \leq M$, *then there exists a time* $T^* > 0$ *and a solution* $\mathbf{U} \equiv (\mathbf{u}, \rho) \in C([0,T^*]; H^s(B))$ *to* SPE. *The time* T^* *depends only on* M, B, *and* ε.

The corresponding SPE pressure $\phi \in C([0,T^*]; H^{s+1}(B))$ is determined by $\mathbf{U}$ and its first derivatives, at each time t, $0 \leq t \leq T$, through the elliptic problem

$$\begin{aligned} &\Delta\phi = v_x - u_y - \rho_z - \varepsilon\left[\mathbf{u}_x \cdot \nabla u + \mathbf{u}_y \cdot \nabla v + \mathbf{u}_z \cdot \nabla w\right] \quad \text{in } B, \\ &\phi_z = -\rho \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h, \\ &\int_B \phi = 0. \end{aligned} \tag{4.9}$$

The equation for $\Delta\phi$ in (4.9) is obtained by taking the divergence of momentum equations (4.1)–(4.3), and applying the incompressibility condition (4.5). The Neumann boundary condition follows from (4.3) and (4.8). It can easily be verified that the compatibility condition for (4.9) is satisfied.

Valli's argument can also be used to prove a short-time existence theorem for a periodic SPE system (PSPE) consisting of (4.1)–(4.7) on the periodic domain B_p, defined by (3.28). This theorem is analogous to Theorem 3.5 for the periodic QGS equations.

THEOREM 4.2 (PSPE short-time existence). *If periodic initial data* $\mathbf{U}_0^{per} \equiv (\mathbf{u}_0^{per}, \rho_0^{per})$ *in* B_p *is given such that* u_0^{per}, v_0^{per} *are in* $H^{s,per}_{even}(B_p)$, *and* w_0^{per}, ρ_0^{per} *are in* $H^{s,per}_{odd}(B_p)$, *for some* $s \geq 3$, *with* $|\mathbf{U}_0^{per}|_s \leq M$, *then there exists a time* $T^* > 0$ *and a solution* $\mathbf{U}^{per} \equiv (\mathbf{u}^{per}, \rho^{per})$ *to* PSPE *such that* u^{per}, v^{per} *are in* $C([0,T^*]; H^{s,per}_{even}(B_p))$ *and* w^{per}, ρ^{per} *are in* $C([0,T^*]; H^{s,per}_{odd}(B_p))$. *The time* T^* *depends only on* M, B_p, *and* ε.

Note that the corresponding PSPE pressure $\phi^{per} \in C([0,T^*]; H^{s+1,per}_{even}(B_p))$ is determined by $\mathbf{U}^{per}$ and its first derivatives, at each time t, $0 \leq t \leq T$, through a periodic elliptic problem similar to (4.9).

4.1. Initial conditions on the boundary. If we initialize SPE such that $\mathbf{U}_0(\varepsilon)$ is in $H^4(B)$ and u_{0z}, u_{0zzz}, v_{0z}, v_{0zzz}, w_0, w_{0zz}, ρ_0, and ρ_{0zz} are zero on the boundaries Σ_0 and Σ_h, then by Theorem 1.1 we can make the appropriate periodic extensions of $\mathbf{U}_0(\varepsilon)$ to B_p. By uniqueness, the resulting SPE and PSPE solutions (guaranteed by the above existence theorems) must agree in B for each ε. Consequently, the higher boundary conditions persist for the SPE solutions. We state this formally in Corollary 4.3 below. Notice that initializing with $\rho_0 = 0$ on Σ_0 and Σ_h forces the condition $\phi_{0z} = 0$ on Σ_0 and Σ_h, due to the boundary condition $\phi_{0z} = -\rho_0$ in (4.9). Thus the initial SPE pressure is consistent with the initial PSPE pressure, which is an even periodic function of z in B_p.

COROLLARY 4.3. *Suppose we are given data* $\mathbf{U}_0(\varepsilon)$ *in* $H^4(B)$ *such that*

$$\begin{aligned} &u_{0z} = u_{0zzz} = 0, \quad v_{0z} = v_{0zzz} = 0, \\ &w_0 = w_{0zz} = 0, \quad \rho_0 = \rho_{0zz} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h. \end{aligned} \tag{4.10}$$

Let $\mathbf{U}(\varepsilon) \in C([0,T^*]; H^4(B))$ *be the* SPE *solution with initial data* $\mathbf{U}_0(\varepsilon)$, *guaranteed by Theorem* 4.1. *Then for all time* $t \in (0,T^*)$, *the solution* $\mathbf{U}$ *satisfies*

$$\begin{aligned} &u_z = u_{zzz} = 0, \quad v_z = v_{zzz} = 0, \\ &w = w_{zz} = 0, \quad \rho = \rho_{zz} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h. \end{aligned}$$

4.2. Suppression of fast-scale motion. On some time interval $[0,T]$, if we want to compare QGS solutions with SPE solutions as ε approaches zero, we must initialize SPE in a way that guarantees existence of SPE solutions up to time T for all ε near zero. In general, due to the ε factor on the time derivative terms in (4.1) and (4.2), SPE solutions will contain motions that vary on the fast time scale t/ε. QGS solutions vary on the (slow) time scale t, comparable by nature of the QGS scaling to the rotation period of the earth. There is a simple condition on the SPE initial data that is necessary and sufficient for suppression of fast-scale motion.

THEOREM 4.4 (SPE initialization). *If the initial data* $\mathbf{U}_0(\varepsilon)$ *is bounded in* $H^{s+1}(B)$ *independently of* ε, *then* SPE *solutions* $\mathbf{U}(\varepsilon)$ *have time derivatives* $\mathbf{U}_t(\varepsilon)$ *initially bounded in* $H^s(B)$ *independently of* ε *if and only if there exists* ψ *in* $H^{s+1}(B)$ *such that*

$$
\begin{aligned}
&|v_0 - \psi_x|_s = O(\varepsilon)\,, \qquad |u_0 + \psi_y|_s = O(\varepsilon)\,,\\
&|\rho_0 + \psi_z|_s = O(\varepsilon)\,, \qquad |w_0|_s = O(\varepsilon)\,.
\end{aligned}
\tag{4.11}
$$

Proof. First we want to show that if (4.11) holds, then the initial data satisfy

$$
|v_0 - \phi_{0x}|_s = O(\varepsilon)\,, \quad |u_0 + \phi_{0y}|_s = O(\varepsilon)\,, \quad |\rho_0 + \phi_{0z}|_s = O(\varepsilon)\,,
\tag{4.12}
$$

where ϕ_0 is the initial pressure determined by the initial data $\mathbf{U}_0$. From (4.9) we have

$$
\Delta\phi_0 = v_{0x} - u_{0y} - \rho_{0z} - \varepsilon\Big[\mathbf{u}_{0x}\cdot\nabla u_0 + \mathbf{u}_{0y}\cdot\nabla v_0 + \mathbf{u}_{0z}\cdot\nabla w_0\Big] \quad \text{in } B\,,
\tag{4.13}
$$

and if we substitute (4.11) into (4.13) for v_{0x}, u_{0y}, and ρ_{0z}, we get

$$
|\Delta(\psi - \phi_0)|_{s-1} = O(\varepsilon)\,.
\tag{4.14}
$$

Since we know $\mathbf{U}_0(\varepsilon)$ is bounded in $H^{s+1}(B)$ independently of ε, the right-hand side of (4.14) takes care of the ε term from (4.13). Also from (4.9) and (4.11), on the boundary of B we can apply the trace theorem for Sobolev spaces to get

$$
|\psi_z - \phi_{0z}|_{s-1/2,\partial B} = O(\varepsilon)\,.
\tag{4.15}
$$

Adjusting ψ by a constant does not affect (4.11), so we can assume $\int_B \psi = 0$. Therefore we have

$$
\int_B (\psi - \phi_0) = 0\,,
\tag{4.16}
$$

and by elliptic theory, (4.14)–(4.16) imply

$$
|\psi - \phi_0|_{s+1} = O(\varepsilon)\,.
\tag{4.17}
$$

By virtue of (4.17), initial conditions (4.12) follow from (4.11). Now using (4.12) with SPE (4.1)–(4.4), we have

$$
\begin{aligned}
&\varepsilon\left|\left(\frac{Du}{Dt}\right)_{t=0}\right|_s = |v_0 - \phi_{0x}|_s = O(\varepsilon)\,, \qquad \varepsilon\left|\left(\frac{Dv}{Dt}\right)_{t=0}\right|_s = |u_0 + \phi_{0y}|_s = O(\varepsilon)\,,\\
&\varepsilon\left|\left(\frac{Dw}{Dt}\right)_{t=0}\right|_s = |\rho_0 + \phi_{0z}|_s = O(\varepsilon)\,, \qquad \varepsilon\left|\left(\frac{D\rho}{Dt}\right)_{t=0}\right|_s = |w_0|_s = O(\varepsilon)\,.
\end{aligned}
\tag{4.18}
$$

These equations imply that $|(D\mathbf{U}/Dt)_{t=0}|_s$ is bounded independently of ε. By assumption, $\mathbf{U}_0$ is bounded in $H^{s+1}(B)$ independently of ε. Thus $|(\mathbf{U}_t)_{t=0}|_s$ is bounded independently of ε. Reversing the argument, (4.18) holds if we assume $|(\mathbf{U}_t)_{t=0}|_s$ is bounded independently of ε. □

4.3. The convergence theorem. Now we are ready to state and prove our main theorem, which tells us when QGS solutions are valid approximations of SPE solutions, with error on the order of ε. We choose the initial data for the SPE solution close to data of QGS type, in accordance with Theorem 4.4, although we do not directly use the conclusion of Theorem 4.4 in the proof of the main theorem.

Notice for the remainder of the paper, we use, for example, $\mathbf{u}^{(g)}$ instead of $\mathbf{u}^{(0)}$ for the geostrophic velocity.

THEOREM 4.5. *Assume we are given time $T > 0$ and initial* QGS *velocity and density $\mathbf{U}_0^{(g)} \equiv (\mathbf{u}_0^{(g)}, \rho_0^{(g)})$ in $H^6(B)$ of the form $\mathbf{u}_0^{(g)} = (-\phi_{0y}^{(g)}, \phi_{0x}^{(g)}, 0)$, $\rho_0^{(g)} = -\phi_{0z}^{(g)}$, for some pressure function $\phi_0^{(g)}$ satisfying* (3.32). *For $\lambda = 1$ and $\beta_0 = 0$, let $\mathbf{U}^{(g)} \equiv (\mathbf{u}^{(g)}, \rho^{(g)})$ in $C([0,T]; H^6(B))$ be the* QGS *solution with initial data $\mathbf{U}_0^{(g)}$. We consider initial* SPE *data $\mathbf{U}_0(\varepsilon) \equiv (\mathbf{u}_0(\varepsilon), \rho_0(\varepsilon))$ in $H^5(B)$ satisfying* (4.10), *such that*

$$|\mathbf{U}_0(\varepsilon) - \mathbf{U}_0^{(g)}|_4 = O(\varepsilon)\,. \tag{4.19}$$

Then there exists $\varepsilon_0 > 0$ and solutions $\mathbf{U}(\varepsilon) \equiv (\mathbf{u}(\varepsilon), \rho(\varepsilon))$ in $C([0,T]; H^5(B))$ to SPE (4.1)–(4.8) *for all $\varepsilon \le \varepsilon_0$, which converge to the* QGS *solution in $C([0,T]; H^4(B))$ with $O(\varepsilon)$ accuracy, i.e.,*

$$|\mathbf{U}(\varepsilon) - \mathbf{U}^{(g)}|_{4,T} = O(\varepsilon)\,.$$

Proof. The theorem holds for the more general case $\beta_0 \ge 0$, but for simplicity we assume $\beta_0 = 0$. We write SPE solutions $\mathbf{U}(\varepsilon)$ as the sum of geostrophic and ageostrophic parts, i.e., $\mathbf{U}(\varepsilon) \equiv \mathbf{U}^{(g)} + \varepsilon \mathbf{U}^{(a)}(\varepsilon)$, where $\mathbf{U}^{(a)} \equiv (\mathbf{u}^{(a)}, \rho^{(a)})$. The main part of the proof is to show the following.

CLAIM 4.6. *With the initialization specified above, there exist $\varepsilon_0 > 0$ and constant M such that, as long as solutions $\mathbf{U}(\varepsilon)$ to* SPE *exist in $H^5(B)$, then $|\mathbf{U}^{(a)}(\varepsilon)|_{4,T^*} \le M$ holds uniformly for $0 < \varepsilon \le \varepsilon_0$, where $T^* \le T$ is the time of existence. The bound M depends only on ε_0, norms $|\mathbf{U}_0^{(g)}|_6$ and $|\mathbf{U}_0^{(a)}(\varepsilon)|_4$ of the initial data, and the final time T.*

Once we have verified this assertion, the theorem is proved with the following argument. In the short-time existence proofs by Valli [25] and Temam [24], it is shown that as long as solutions exists, they satisfy a certain a priori estimate (see pp. 45–47 in [25] and inequality (1.9) in [24]). We can modify these arguments to get an improved estimate (see (4.20) below) for the growth of $|\mathbf{U}(\varepsilon)|_5$, in terms of the "low" norm $|\mathbf{U}(\varepsilon)|_3$. This involves use of calculus inequality (3.15) to get an estimate involving $|\nabla \mathbf{U}(\varepsilon)|_\infty$, which can be replaced by $|\mathbf{U}(\varepsilon)|_3$ using the Sobolev lemma. The reader may wish to confer with Chapter 2 of [17] for details on the use of the L^∞ norm of velocity gradients for determining the maximal interval of existence of H^s solutions for the incompressible Euler equations.

With this modification, it follows that as long as solutions $\mathbf{U}(\varepsilon)$ to SPE exist in $H^5(B)$, they satisfy the growth estimate

$$\frac{d}{dt}|\mathbf{U}(\varepsilon)|_5 \le C(\varepsilon)(1 + |\mathbf{U}(\varepsilon)|_3)|\mathbf{U}(\varepsilon)|_5\,, \tag{4.20}$$

where $C(\varepsilon)$ is $O(\varepsilon^{-1})$. Using (4.20) with Claim 4.6, as long as SPE solutions exist we have

$$|\mathbf{U}(\varepsilon)|_5 \leq K_\varepsilon, \quad \text{where } K_\varepsilon = \sup_{0\leq t\leq T} |\mathbf{U}_0(\varepsilon)|_5 \exp\left\{C(\varepsilon)(1+|\mathbf{U}^{(g)}|_3+\varepsilon M)\,t\right\}.$$

Using the global bound K_ε with the short-time existence theorem establishes the existence of large-time SPE solutions in $C([0,T];H^5(B))$ for each $\varepsilon \leq \varepsilon_0$. Now having established this, we can use Claim 4.6 to conclude that SPE solutions $\mathbf{U}(\varepsilon)=\mathbf{U}^{(g)}+\varepsilon\mathbf{U}^{(a)}(\varepsilon)$ converge in $H^4(B)$ to the QGS solution $\mathbf{U}^{(g)}$ to $O(\varepsilon)$.

Proving the claim, that $|\mathbf{U}^{(a)}(\varepsilon)|_4$ remains bounded as ε approaches zero, is difficult because of the factor of ε in the time derivative terms above. Our method for getting the desired energy estimate for $\mathbf{U}^{(a)}(\varepsilon)$ will require use of the quasigeostrophic potential vorticity conservation equation (2.23). This entails taking the two-dimensional curl of the horizontal equations of motion (4.1) and (4.2). As in the standard derivation of the quasigeostrophic equation (2.23) in §2, this process annihilates the dominant Coriolis force in the horizontal momentum balance, thus allowing us to estimate the higher-order (ageostrophic) term $\mathbf{U}^{(a)}(\varepsilon)$. Thus estimating the ageostrophic motion is brought about by considerations of vorticity dynamics, and we should expect the quasigeostrophic equation (2.23) to come into play somewhere in the argument. This will be explained in detail below.

For Claim 4.6 to hold, we must require that the initial ageostrophic motion is given such that $|\mathbf{U}_0^{(a)}(\varepsilon)|_4$ is bounded uniformly in ε. We can see that this requirement is satisfied by substituting, for example, $u_0 = -\phi_{0y}^{(g)} + \varepsilon u_0^{(a)}$ into initial condition (4.19). In the derivation of the energy estimate for $\mathbf{U}^{(a)}(\varepsilon)$ below, we encounter boundary integrals (see (4.32) and (4.33)) that must vanish in order for $\mathbf{U}^{(a)}(\varepsilon)$ to remain bounded as ε approaches zero. We need the special QGS and SPE initial boundary conditions, as specified in Corollaries 3.6 and 4.3, to conclude that these boundary integrals are zero.

We begin the energy estimate derivation for $\mathbf{U}^{(a)}(\varepsilon)$ by differentiating equations (4.1)–(4.4) by D^α, substituting $\mathbf{U}^{(g)}+\varepsilon\mathbf{U}^{(a)}(\varepsilon)$ for $\mathbf{U}$, and using the zero-order relations (3.1)–(3.3). Using the subscript notation $u_\alpha^{(a)}$ for $D^\alpha u^{(a)}$, etc., we have

$$\left(\frac{Du^{(g)}}{Dt}\right)_\alpha + \varepsilon\left(\frac{Du^{(a)}}{Dt}\right)_\alpha - v_\alpha^{(a)} = -D^\alpha\phi_x^{(a)}\,, \tag{4.21}$$

$$\left(\frac{Dv^{(g)}}{Dt}\right)_\alpha + \varepsilon\left(\frac{Dv^{(a)}}{Dt}\right)_\alpha + u_\alpha^{(a)} = -D^\alpha\phi_y^{(a)}\,, \tag{4.22}$$

$$\varepsilon\left(\frac{Dw^{(a)}}{Dt}\right)_\alpha + \rho_\alpha^{(a)} = -D^\alpha\phi_z^{(a)}\,, \tag{4.23}$$

$$\left(\frac{D\rho^{(g)}}{Dt}\right)_\alpha + \varepsilon\left(\frac{D\rho^{(a)}}{Dt}\right)_\alpha - w_\alpha^{(a)} = 0\,. \tag{4.24}$$

Multiplying (4.21)–(4.24) by $u_\alpha^{(a)}$, $v_\alpha^{(a)}$, $w_\alpha^{(a)}$, and $\rho_\alpha^{(a)}$, respectively, then summing and noting the obvious cancellation, we have

(4.25)

$$\varepsilon\left\{ u_\alpha^{(a)}\left(\frac{Du^{(a)}}{Dt}\right)_\alpha + v_\alpha^{(a)}\left(\frac{Dv^{(a)}}{Dt}\right)_\alpha + w_\alpha^{(a)}\left(\frac{Dw^{(a)}}{Dt}\right)_\alpha + \rho_\alpha^{(a)}\left(\frac{D\rho^{(a)}}{Dt}\right)_\alpha \right\}$$
$$+ u_\alpha^{(a)}\left(\frac{Du^{(g)}}{Dt}\right)_\alpha + v_\alpha^{(a)}\left(\frac{Dv^{(g)}}{Dt}\right)_\alpha + \rho_\alpha^{(a)}\left(\frac{D\rho^{(g)}}{Dt}\right)_\alpha$$
$$= -\left(u_\alpha^{(a)} D^\alpha \phi_x^{(a)} + v_\alpha^{(a)} D^\alpha \phi_y^{(a)} + w_\alpha^{(a)} D^\alpha \phi_z^{(a)} \right).$$

The term $\varepsilon u_\alpha^{(a)} \partial_t u_\alpha^{(a)}$, for example, in (4.25) becomes $\frac{1}{2}\varepsilon \frac{d}{dt} |u_\alpha^{(a)}|_{L^2}^2$ after integrating over B; therefore, if we expect to get a bounded growth estimate for $|\mathbf{U}^{(a)}(\varepsilon)|_4$ as ε approaches zero, we must show that the sum of the "large" terms (those terms with no ε) in (4.25) is actually $O(\varepsilon)$. We demonstrate shortly that the right-hand side of (4.25) vanishes upon integration. So we concentrate now on the remaining large terms

$$u_\alpha^{(a)}\left(\frac{Du^{(g)}}{Dt}\right)_\alpha + v_\alpha^{(a)}\left(\frac{Dv^{(g)}}{Dt}\right)_\alpha + \rho_\alpha^{(a)}\left(\frac{D\rho^{(g)}}{Dt}\right)_\alpha . \tag{4.26}$$

Our approach is to substitute for $v_\alpha^{(a)}$, $u_\alpha^{(a)}$, and $\rho_\alpha^{(a)}$ in (4.26) from (4.21)–(4.23), and use conservation of potential vorticity to eliminate the resulting large terms. We introduce the notation

$$d_a \equiv u^{(a)}\partial_x + v^{(a)}\partial_y + w^{(a)}\partial_z ,$$

and represent the material derivative as $D/Dt = d_g + \varepsilon d_a$, where we recall the definition of the geostrophic operator $d_g \equiv \partial_t + u^{(g)}\partial_x + v^{(g)}\partial_y$. Using this notation and making the aforementioned substitution, after the obvious cancellation (4.26),

(4.27)

$$D^\alpha \phi_x^{(a)} \left(d_g v^{(g)}\right)_\alpha - D^\alpha \phi_y^{(a)} \left(d_g u^{(g)}\right)_\alpha - D^\alpha \phi_z^{(a)} \left(d_g \rho^{(g)}\right)_\alpha$$
$$+ \varepsilon\left\{ D^\alpha \phi_x^{(a)} \left(d_a v^{(g)}\right)_\alpha - D^\alpha \phi_y^{(a)} \left(d_a u^{(g)}\right)_\alpha - D^\alpha \phi_z^{(a)} \left(d_a \rho^{(g)}\right)_\alpha \right\}$$
$$+ \varepsilon\left\{ \left(\frac{Du^{(a)}}{Dt}\right)_\alpha \left(\frac{Dv^{(g)}}{Dt}\right)_\alpha - \left(\frac{Dv^{(a)}}{Dt}\right)_\alpha \left(\frac{Du^{(g)}}{Dt}\right)_\alpha - \left(\frac{Dw^{(a)}}{Dt}\right)_\alpha \left(\frac{D\rho^{(g)}}{Dt}\right)_\alpha \right\}.$$

When we integrate by parts, we will move a derivative from the ageostrophic pressure occurring in the first three terms of (4.27), onto the geostrophic quantities $d_g v^{(g)}$, $-d_g u^{(g)}$, and $-d_g \rho^{(g)}$ to form a potential vorticity term. This term is zero by conservation of potential vorticity (i.e., the quasigeostrophic equation), and the related boundary term is zero by our periodic boundary conditions. The details are shown below.

So we have eliminated the large terms in (4.27), but at the cost of acquiring terms like

$$\left(\frac{Du^{(a)}}{Dt}\right)_\alpha \left(\frac{Dv^{(g)}}{Dt}\right)_\alpha = \left(\frac{Du^{(a)}}{Dt}\right)_\alpha \left(d_g v^{(g)}\right)_\alpha + \varepsilon \left(\frac{Du^{(a)}}{Dt}\right)_\alpha \left(d_a v^{(g)}\right)_\alpha , \tag{4.28}$$

which contain high spatial derivatives of $\mathbf{U}^{(a)}$. We can properly estimate the first term on the right-hand side of (4.28) because a derivative can be moved from $u^{(a)}$

to the (smooth) purely geostrophic part $(d_g v^{(g)})_\alpha$. This is described below in the discussion for estimating expression $A_{2,\alpha}$. The second term on the right-hand side of (4.28) will cancel with another term from (4.27) after a further substitution. In (4.27) we substitute for $D^\alpha \phi_x^{(a)}$ in the fourth term, using (4.21), obtaining

$$D^\alpha \phi_x^{(a)} \left(d_a v^{(g)}\right)_\alpha = -\left(\frac{Du^{(g)}}{Dt} + \varepsilon \frac{Du^{(a)}}{Dt} - v^{(a)}\right)_\alpha \left(d_a v^{(g)}\right)_\alpha . \tag{4.29}$$

Using (4.22) and (4.23) to do the same kind of substitution in (4.27) for

$$D^\alpha \phi_y^{(a)} \left(d_a u^{(g)}\right)_\alpha \quad \text{and} \quad D^\alpha \phi_z^{(a)} \left(d_g \rho^{(g)}\right)_\alpha ,$$

respectively, we finally rewrite (4.25) as

$$\begin{aligned}
&\varepsilon \Bigg\{ u_\alpha^{(a)} \left(\frac{Du^{(a)}}{Dt}\right)_\alpha + v_\alpha^{(a)} \left(\frac{Dv^{(a)}}{Dt}\right)_\alpha + w_\alpha^{(a)} \left(\frac{Dw^{(a)}}{Dt}\right)_\alpha + \rho_\alpha^{(a)} \left(\frac{D\rho^{(a)}}{Dt}\right)_\alpha \\
&\quad + \left(\frac{Du^{(a)}}{Dt}\right)_\alpha \left(d_g v^{(g)}\right)_\alpha - \left(\frac{Dv^{(a)}}{Dt}\right)_\alpha \left(d_g u^{(g)}\right)_\alpha - \left(\frac{Dw^{(a)}}{Dt}\right)_\alpha \left(d_g \rho^{(g)}\right)_\alpha \\
&\quad - \left(\frac{Du^{(g)}}{Dt} - v^{(a)}\right)_\alpha \left(d_a v^{(g)}\right)_\alpha + \left(\frac{Dv^{(g)}}{Dt} + u^{(a)}\right)_\alpha \left(d_a u^{(g)}\right)_\alpha \\
&\quad + \rho_\alpha^{(a)} \left(d_a \rho^{(g)}\right)_\alpha \Bigg\} \\
&= -\nabla \phi_\alpha^{(a)} \cdot D^\alpha (d_g v^{(g)}, -d_g u^{(g)}, -d_g \rho^{(g)}) \\
&\quad - \left(u_\alpha^{(a)} D^\alpha \phi_x^{(a)} + v_\alpha^{(a)} D^\alpha \phi_y^{(a)} + w_\alpha^{(a)} D^\alpha \phi_z^{(a)} \right) .
\end{aligned} \tag{4.30}$$

We will use (4.30) to get an estimate on the growth rate in time of $|\mathbf{U}^{(a)}(t)|_4$. Integrating (4.30) over B, the first term on the right-hand side of (4.30) becomes, after integrating by parts,

$$\begin{aligned}
&-\int\int\int_B \nabla \cdot \left[\phi_\alpha^{(a)} D^\alpha (d_g v^{(g)}, -d_g u^{(g)}, -d_g \rho^{(g)}) \right] d\mathbf{x} \\
&\quad + \int\int\int_B \phi_\alpha^{(a)} D^\alpha \left\{ \left[d_g v^{(g)}\right]_x - \left[d_g u^{(g)}\right]_y - \left[d_g \rho^{(g)}\right]_z \right\} d\mathbf{x} .
\end{aligned} \tag{4.31}$$

By the geostrophic relation $u_x^{(g)} = -v_y^{(g)}$, we can write

$$\left[d_g v^{(g)}\right]_x = d_g v_x^{(g)} + u_x^{(g)} v_x^{(g)} + v_x^{(g)} v_y^{(g)} = d_g v_x^{(g)} ,$$

and similarly $[d_g u^{(g)}]_y = d_g u_y^{(g)}$. Likewise, using $u_z^{(g)} = \rho_y^{(g)}$ and $v_z^{(g)} = -\rho_x^{(g)}$, we can write $[d_g \rho^{(g)}]_z = d_g \rho_z^{(g)}$. Then by the divergence theorem, (4.31) becomes

$$\begin{aligned}
&-\int\int_{\partial B} \phi_\alpha^{(a)} D^\alpha (d_g v^{(g)}, -d_g u^{(g)}, -d_g \rho^{(g)}) \cdot \mathbf{n}\, ds \\
&\quad + \int\int\int_B \phi_\alpha^{(a)} D^\alpha \left[d_g (v_x^{(g)} - u_y^{(g)} - \rho_z^{(g)}) \right] d\mathbf{x},
\end{aligned}$$

where $\mathbf{n}$ represents the unit normal to the boundary ∂B. The second integral is zero because the quasigeostrophic (2.23) holds everywhere in B. Due to the horizontal periodicity of solutions, and since $\mathbf{n} = (0,0,-1)$ on Σ_0 and $\mathbf{n} = (0,0,1)$ on Σ_h, the boundary integral reduces to

$$\int\int_{\Sigma_h} \phi_\alpha^{(a)} D^\alpha (d_g \rho^{(g)})\, dx\, dy - \int\int_{\Sigma_0} \phi_\alpha^{(a)} D^\alpha (d_g \rho^{(g)})\, dx\, dy\,. \tag{4.32}$$

We claim that the integrand in (4.32) is identically zero on Σ for $|\alpha| \le 4$. When D^α consists only of horizontal derivatives, the integrand is zero due to the QGS boundary condition (3.6). When D^α contains at most three vertical derivatives, the integrand is zero by Corollaries 3.6 and 4.3. (For the case of three vertical derivatives, we differentiate the equation for $\Delta\phi$ in (4.9) with respect to z, and apply Corollary 4.3 to get $\phi_{zzz} = 0$ on Σ.) For the remaining case of four vertical derivatives, we rewrite $\partial_z^4 (d_g \rho^{(g)}) = \partial_z^3 (d_g \rho_z^{(g)})$ (see (2.20)), and then apply the conservation of potential vorticity equation (2.22) (with $\beta_0 = 0$ and $\lambda = 1$) to get $\partial_z^4 (d_g \rho^{(g)}) = \partial_z^3 [d_g u_y^{(g)} - d_g v_x^{(g)}]$, which is zero on Σ by Corollary 3.6.

The second term on the right-hand side of (4.30) becomes, after integrating by parts,

$$-\int\int\int_B \nabla \cdot \left[\mathbf{u}_\alpha^{(a)}\, \phi_\alpha^{(a)}\right] d\mathbf{x} + \int\int\int_B \phi_\alpha^{(a)} \left(\nabla \cdot \mathbf{u}_\alpha^{(a)}\right) d\mathbf{x}\,.$$

The second integral is zero due to the incompressibility of the flow. Applying the divergence theorem, the first integral becomes

$$\int\int_{\Sigma_0} \phi_\alpha^{(a)} w_\alpha^{(a)}\, dx\, dy - \int\int_{\Sigma_h} \phi_\alpha^{(a)} w_\alpha^{(a)}\, dx\, dy\,. \tag{4.33}$$

We claim that the integrand in (4.33) is identically zero on Σ for $|\alpha| \le 4$. When D^α consists only of horizontal derivatives, the integrand is zero due to the SPE boundary condition (4.8). When D^α contains at most three vertical derivatives, the integrand is zero by Corollary 4.3 ($\phi_{zzz} = 0$ on Σ as explained above). For the remaining case of four vertical derivatives, we use $\partial_z^3 (\nabla \cdot \mathbf{u}) = 0$ and Corollary 4.3 to get that $\partial_z^4 w = 0$ on Σ.

We proceed with the energy estimate for $\mathbf{U}^{(a)}$. To integrate, for example, the term $u_\alpha^{(a)} (Du^{(a)}/Dt)_\alpha$ in (4.30), we write

$$\begin{aligned} u_\alpha^{(a)} \left(\frac{Du^{(a)}}{Dt}\right)_\alpha &= u_\alpha^{(a)} \frac{\partial u_\alpha^{(a)}}{\partial t} + u_\alpha^{(a)} \left(\mathbf{u}\cdot\nabla u^{(a)}\right)_\alpha \\ &= u_\alpha^{(a)} \frac{\partial u_\alpha^{(a)}}{\partial t} + u_\alpha^{(a)} \mathbf{u}\cdot\nabla u_\alpha^{(a)} + u_\alpha^{(a)} F_{1,\alpha}, \end{aligned}$$

where

$$F_{1,\alpha} = \left(\mathbf{u}\cdot\nabla u^{(a)}\right)_\alpha - \mathbf{u}\cdot\nabla u_\alpha^{(a)}\,. \tag{4.34}$$

Then the integral of $u_\alpha^{(a)} (Du^{(a)}/Dt)_\alpha$ becomes, after using the divergence theorem with our boundary conditions,

$$\left(u_\alpha^{(a)}\,,\, \left(\frac{Du^{(a)}}{Dt}\right)_\alpha\right)_{L_2} = \frac{1}{2}\frac{d}{dt} |u_\alpha^{(a)}|_0^2 + \left(u_\alpha^{(a)}\,,\, F_{1,\alpha}\right)_{L^2}.$$

Similarly defining $F_{2,\alpha}$, $F_{3,\alpha}$, and $F_{4,\alpha}$ for $v^{(a)}$, $w^{(a)}$, and $\rho^{(a)}$, respectively, we integrate (4.30) over B for all α, $|\alpha| \le 4$, to get

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\Big\{ |u_\alpha^{(a)}|_0^2 + |v_\alpha^{(a)}|_0^2 + |w_\alpha^{(a)}|_0^2 + |\rho_\alpha^{(a)}|_0^2 \Big\} + \left(\left(\frac{Du^{(a)}}{Dt} \right)_\alpha , \left(d_g v^{(g)} \right)_\alpha \right)_{L^2} \\
&\quad - \left(\left(\frac{Dv^{(a)}}{Dt} \right)_\alpha , \left(d_g u^{(g)} \right)_\alpha \right)_{L^2} - \left(\left(\frac{Dw^{(a)}}{Dt} \right)_\alpha , \left(d_g \rho^{(g)} \right)_\alpha \right)_{L^2} \\
&\quad + \left(\left(\frac{Du^{(g)}}{Dt} \right)_\alpha , \left(d_a v^{(g)} \right)_\alpha \right)_{L^2} + \left(\left(\frac{Dv^{(g)}}{Dt} \right)_\alpha , \left(d_a u^{(g)} \right)_\alpha \right)_{L^2} \\
&\quad + \left(v_\alpha^{(a)} , \left(d_a v^{(g)} \right)_\alpha \right)_{L^2} + \left(u_\alpha^{(a)} , \left(d_a u^{(g)} \right)_\alpha \right)_{L^2} + \left(\rho_\alpha^{(a)} , \left(d_a \rho^{(g)} \right)_\alpha \right)_{L^2} \\
&= -\left(u_\alpha^{(a)} , F_{1,\alpha} \right)_{L^2} - \left(v_\alpha^{(a)} , F_{2,\alpha} \right)_{L^2} - \left(w_\alpha^{(a)} , F_{3,\alpha} \right)_{L^2} - \left(\rho_\alpha^{(a)} , F_{4,\alpha} \right)_{L^2} .
\end{aligned} \tag{4.35}
$$

We will obtain an energy estimate for $\mathbf{U}^{(a)}(t)$ in $H^4(B)$ by integrating (4.35) with respect to time, summing over α, $|\alpha| \le 4$, and applying the Cauchy–Schwarz inequality. In preparation for the time integration, we rewrite, for example, the term

$$
\left(\left(u_t^{(a)} \right)_\alpha , \left(d_g v^{(g)} \right)_\alpha \right)_{L^2}
$$

occurring in the second term of (4.35) as

$$
\frac{d}{dt} \left(u_\alpha^{(a)} , \left(d_g v^{(g)} \right)_\alpha \right)_{L^2} - \left(u_\alpha^{(a)} , \partial_t \left(d_g v^{(g)} \right)_\alpha \right)_{L^2} .
$$

Doing the same for the third and fourth terms in (4.35), and integrating the entire equation with respect to time, we get

$$
\tfrac{1}{2}\Big\{ |u_\alpha^{(a)}|_0^2 + |v_\alpha^{(a)}|_0^2 + |w_\alpha^{(a)}|_0^2 + |\rho_\alpha^{(a)}|_0^2 \Big\}\Big|_0^t = A_{1,\alpha} + A_{2,\alpha} + A_{3,\alpha} + A_{4,\alpha}, \tag{4.36}
$$

where

$$
\begin{aligned}
A_{1,\alpha} &= \left\{ -\left(u_\alpha^{(a)}, \left(d_g v^{(g)} \right)_\alpha \right)_{L^2} + \left(v_\alpha^{(a)}, \left(d_g u^{(g)} \right)_\alpha \right)_{L^2} + \left(w_\alpha^{(a)}, \left(d_g \rho^{(g)} \right)_\alpha \right)_{L^2} \right\}\Big|_0^t \\
&\quad + \int_0^t \left\{ \left(u_\alpha^{(a)}, \partial_t \left(d_g v^{(g)} \right)_\alpha \right)_{L^2} - \left(v_\alpha^{(a)}, \partial_t \left(d_g u^{(g)} \right)_\alpha \right)_{L^2} \right. \\
&\qquad \left. - \left(w_\alpha^{(a)}, \partial_t \left(d_g \rho^{(g)} \right)_\alpha \right)_{L^2} \right\} d\tau , \\
A_{2,\alpha} &= \int_0^t \left\{ -\left(\left(\mathbf{u}\cdot\nabla u^{(a)} \right)_\alpha , \left(d_g v^{(g)} \right)_\alpha \right)_{L^2} + \left(\left(\mathbf{u}\cdot\nabla v^{(a)} \right)_\alpha , \left(d_g u^{(g)} \right)_\alpha \right)_{L^2} \right. \\
&\qquad \left. + \left(\left(\mathbf{u}\cdot\nabla w^{(a)} \right)_\alpha , \left(d_g \rho^{(g)} \right)_\alpha \right)_{L^2} \right\} d\tau ,
\end{aligned}
$$

$$A_{3,\alpha} = -\int_0^t \left\{ \left(\left(\frac{Du^{(g)}}{Dt} \right)_\alpha , \left(d_a v^{(g)} \right)_\alpha \right)_{L^2} + \left(\left(\frac{Dv^{(g)}}{Dt} \right)_\alpha , \left(d_a u^{(g)} \right)_\alpha \right)_{L^2} \right\} d\tau$$
$$- \int_0^t \left\{ \left(v_\alpha^{(a)} , \left(d_a v^{(g)} \right)_\alpha \right)_{L^2} + \left(u_\alpha^{(a)} , \left(d_a u^{(g)} \right)_\alpha \right)_{L^2} \right.$$
$$\left. + \left(\rho_\alpha^{(a)} , \left(d_a \rho^{(g)} \right)_\alpha \right)_{L^2} \right\} d\tau ,$$

$$A_{4,\alpha} = -\int_0^t \left\{ \left(u_\alpha^{(a)}, F_{1,\alpha} \right)_{L^2} + \left(v_\alpha^{(a)}, F_{2,\alpha} \right)_{L^2} + \left(w_\alpha^{(a)}, F_{3,\alpha} \right)_{L^2} + \left(\rho_\alpha^{(a)}, F_{4,\alpha} \right)_{L^2} \right\} d\tau .$$

If we define

$$Y^2 = |u^{(a)}|_4^2 + |v^{(a)}|_4^2 + |w^{(a)}|_4^2 + |\rho^{(a)}|_4^2 \tag{4.37}$$

and sum over α, $|\alpha| \le 4$, the left-hand side of (4.36) becomes $\frac{1}{2}[Y^2(t) - Y^2(0)]$. We now proceed to estimate $A_{i,\alpha}$, $i = 1, 2, 3, 4$, in terms of Y.

By Lemma 3.3, when we apply the Cauchy–Schwarz inequality to expression $A_{1,\alpha}$, the purely geostrophic terms can be bounded by a constant that depends only on the $H^6(B)$ norm of the initial QGS data; therefore, we have

$$\sum_{|\alpha| \le 4} |A_{1,\alpha}| \le C \left(Y(0) + Y(t) + \int_0^t Y(\tau)\, d\tau \right) , \tag{4.38}$$

where C depends only on $|\mathbf{U}_0^{(g)}|_6$.

In expression $A_{2,\alpha}$, terms like $(\mathbf{u} \cdot \nabla u^{(a)})_\alpha$ involve fifth-order spatial derivatives of $u^{(a)}$ when $|\alpha| = 4$, and estimation in $H^4(B)$ is tricky. We first consider the easier case where $\alpha_3 \neq 4$, i.e., at least one derivative is a horizontal one, say $\alpha_1 \ge 1$. We may then integrate by parts once in x, leaving a third derivative of $\mathbf{u} \cdot \nabla u^{(a)}$. The resulting term can be estimated by $CY + C\varepsilon Y^2$.

When $\alpha_3 = 4$, estimation is more difficult. We may again integrate by parts in z. The resulting interior term can be estimated as before, but there are now boundary terms with integrands such as $\partial_z^3 (\mathbf{u} \cdot \nabla u^{(a)})\, \partial_z^4 (d_g v^{(g)})$. The surface integrals appear to present a problem because $\partial_z^3 (\mathbf{u} \cdot \nabla u^{(a)})$ involves fourth-order derivatives of $u^{(a)}$ on the boundaries Σ_0 and Σ_h. However, these surface integrals can be estimated in terms of $|u^{(a)}|_4$ due to the boundary condition $w = 0$. The most troublesome term is

$$\mathbf{u} \cdot \nabla \partial_z^3 u^{(a)} = u \partial_z^3 u_x^{(a)} + v \partial_z^3 u_y^{(a)} + w \partial_z^3 u_z^{(a)} ,$$

with the last vanishing on Σ_0 and Σ_h. For the surface integrals corresponding to the two horizontal terms, we can again integrate by parts. After these modifications, each surface integral involves at most third-order derivatives of $u^{(a)}$. Now we know from the trace theorem for Sobolev spaces that the L^2 norm of $\partial_z^3 u^{(a)}$ on the surfaces Σ_0 and Σ_h is dominated by $|\partial_z^3 u^{(a)}|_1$, and therefore by $|u^{(a)}|_4$.

In summary, we obtain

$$\sum_{|\alpha| \le 4} |A_{2,\alpha}| \le C \int_0^t [\, Y(\tau) + \varepsilon Y^2(\tau) \,]\, d\tau . \tag{4.39}$$

We can directly estimate $A_{3,\alpha}$ to get

$$\sum_{|\alpha| \le 4} |A_{3,\alpha}| \le C \int_0^t [\, Y(\tau) + Y^2(\tau) \,]\, d\tau . \tag{4.40}$$

For $A_{4,\alpha}$, we need estimates for $F_{1,\alpha}$, $F_{2,\alpha}$, $F_{3,\alpha}$, and $F_{4,\alpha}$. From definition (4.34) for $F_{1,\alpha}$ and calculus inequality (3.15), we have

$$\begin{aligned} |F_{1,\alpha}|_0 &\le C(|\mathbf{u}|_4 |\nabla u^{(a)}|_\infty + |\nabla \mathbf{u}|_\infty |\nabla u^{(a)}|_3) \\ &\le C|\mathbf{u}|_4 |u^{(a)}|_4 \le C(\,1 + \varepsilon |\mathbf{u}^{(a)}|_4\,)|u^{(a)}|_4\,, \end{aligned}$$

and similarly for $F_{2,\alpha}$, $F_{3,\alpha}$, and $F_{4,\alpha}$. Thus we have the estimate

$$\sum_{|\alpha|\le 4} |A_{4,\alpha}| \;\le\; C \int_0^t [\, Y^2(\tau) + \varepsilon Y^3(\tau)\,]\, d\tau\,. \tag{4.41}$$

Combining (4.36) with estimates (4.38)–(4.40) and (4.41), and recalling the definition (4.37) for Y, we have

$$Y^2(t) \;\le\; C\left(Y(0) + Y^2(0) + Y(t) + \int_0^t [\, Y(\tau) + Y^2(\tau) + \varepsilon Y^3(\tau)\,]\, d\tau \right).$$

Using the inequality $ab \le (\delta a^2/2) + (b^2/2\delta)$, this reduces to

$$Y^2(t) \;\le\; C\left(1 + Y^2(0) + \int_0^t Y^2(\tau)\, d\tau + \varepsilon \int_0^t Y^3(\tau)\, d\tau \right),$$

where we allow the constant to depend on the time interval, since it is finite. Raising each term to the 3/2 power, this becomes

$$Y^3(t) \;\le\; C\left[\left(1 + Y^2(0)\right)^{3/2} + \left(\int_0^t Y^2(\tau)\, d\tau \right)^{3/2} + \varepsilon^{3/2} \left(\int_0^t Y^3(\tau)\, d\tau \right)^{3/2} \right],$$

or using the Hölder inequality,

$$Y^3(t) \;\le\; C\left[\left(1 + Y^2(0)\right)^{3/2} + \int_0^t Y^3(\tau)\, d\tau + \varepsilon^{3/2} \left(\int_0^t Y^3(\tau)\, d\tau \right)^{3/2} \right].$$

Letting $Z(t) = \int_0^t Y^3(\tau)\, d\tau$, we have $Z' = Y^3$, which results in the inequality

$$Z'(t) \;\le\; C_0\,(1 + Z + \varepsilon^{3/2} Z^{3/2})\,, \tag{4.42}$$

where the constant C_0 now also depends on $Y(0)$, i.e., on $|\mathbf{U}_0^{(a)}(\varepsilon)|_4$. Recall that initial condition (4.19) ensures $|\mathbf{U}_0^{(a)}(\varepsilon)|_4$ is bounded uniformly in ε. Finally, letting $X = (Z+1)^{1/2}$, inequality (4.42) becomes

$$X'(t) \;\le\; C_0\,(\,X + \varepsilon^{3/2} X^2\,), \qquad X \ge 1\,. \tag{4.43}$$

In general, this nonlinear growth rate will cause solutions to blow up in finite time. However, the $\varepsilon^{3/2}$ factor allows us to suppress the nonlinear growth term by choosing ε small enough so that solutions exist up to any prespecified time T. To verify this, notice that the solution $\tilde{X}(t)$ of the differential equation corresponding to (4.43) can be expressed implicitly as

$$t \;=\; \frac{1}{C_0} \ln\!\left[\left(\frac{\tilde{X}}{1 + \varepsilon^{3/2}\tilde{X}} \right) \left(\frac{1 + \varepsilon^{3/2}\tilde{X}(0)}{\tilde{X}(0)} \right) \right], \tag{4.44}$$

where $\tilde{X}(0)$ is an upper bound for $X(0)$, $0 < \varepsilon \le \varepsilon_0$. If we choose $\varepsilon_0 > 0$ small enough so that

$$T < \frac{1}{C_0} \ln \frac{1}{\varepsilon_0^{3/2} \tilde{X}(0)}, \tag{4.45}$$

then for each ε such that $0 < \varepsilon \le \varepsilon_0$, the solution $\tilde{X}(t)$ is defined up to time $T^* > T$. For each ε, solutions $X(t)$ of the differential inequality (4.43) are bounded by the solution $\tilde{X}(t)$ of the differential equation. Hence there exists a number K, depending only on C_0, $X(0)$, ε_0, and T^*, which uniformly bounds X and X' on $[0, T^*]$ for all ε such that $0 < \varepsilon \le \varepsilon_0$. From $Z' = Y^3$ and $X = (Z+1)^{1/2}$, we have $Y = (2XX')^{1/3}$, and we conclude that

$$|\mathbf{U}^{(a)}(\varepsilon)|_{4,T} \le M, \qquad 0 < \varepsilon \le \varepsilon_0, \tag{4.46}$$

where $M = 2K^{2/3}$ depends only on $|\mathbf{U}_0^{(g)}|_6$, $|\mathbf{U}_0^{(a)}(\varepsilon)|_4$, ε_0, and T.

Thus we have established the claim stated at the beginning of the proof, and therefore the existence of large-time solutions $\mathbf{U}(\varepsilon)$ in $C([0,T]; H^5(B))$ for $0 < \varepsilon \le \varepsilon_0$, which satisfy

$$|\mathbf{U}(\varepsilon) - \mathbf{U}^{(g)}|_{4,T} = \varepsilon |\mathbf{U}^{(a)}(\varepsilon)|_{4,T} \le M\varepsilon .$$

Therefore the SPE solutions converge to QGS solutions in $H^4(B)$ with $O(\varepsilon)$. This completes the proof of our main theorem. □

5. The first correction to the QGS solution. Our first goal in this section is to derive an evolution equation for the first correction in ε to the QGS approximation of SPE solutions. That is, we seek an equation that determines the first-order velocity and density $\mathbf{U}^{(1)} = (\mathbf{u}^{(1)}, \rho^{(1)})$ in the formal asymptotic expansion $\mathbf{U}(\varepsilon) = \mathbf{U}^{(g)} + \varepsilon\mathbf{U}^{(1)} + \varepsilon^2\mathbf{U}^{(2)} + \cdots$ of SPE. We find an equation (5.10) for the pressure field in the first-order correction analogous to (2.23) for the QGS solution. We then derive (5.20) for the exact ageostrophic pressure, which is defined as the difference between the SPE pressure and the QGS pressure. To assess the accuracy of the improved QGS solution, with the first correction, we need to compare the solutions of (5.10) and (5.20). We do this in §6, and prove that, if certain initial conditions analogous to those in Theorem 4.5 are satisfied, then $|\mathbf{U}^{(a)}(\varepsilon) - \mathbf{U}^{(1)}|_3$ is $O(\varepsilon)$. That is, if we initialize to appropriately suppress fast solutions, then $\mathbf{U}^{(g)} + \varepsilon\mathbf{U}^{(1)}$ provides $O(\varepsilon^2)$ accurate solutions of SPE. Recall from §4 that $\mathbf{U}(\varepsilon) = \mathbf{U}^{(g)} + \varepsilon\mathbf{U}^{(a)}(\varepsilon)$, and that $\mathbf{U}^{(g)}$ and $\mathbf{U}^{(0)}$ are the same.

5.1. The correction term $\mathbf{U}^{(1)}$. To obtain an equation for $\mathbf{U}^{(1)}$, we begin with the second-order equations in ε and imitate the steps used for formulating the quasigeostrophic equation for $\mathbf{U}^{(g)}$ described in §2. Substituting $\mathbf{U}(\varepsilon) = \mathbf{U}^{(g)} + \varepsilon\mathbf{U}^{(1)} + \varepsilon^2\mathbf{U}^{(2)} + \cdots$ into SPE (2.6)–(2.10) and equating $O(\varepsilon^2)$ terms, we get

$$d_g u^{(1)} + d_1 u^{(g)} - v^{(2)} = -\phi_x^{(2)}, \tag{5.1}$$

$$d_g v^{(1)} + d_1 v^{(g)} + u^{(2)} = -\phi_y^{(2)}, \tag{5.2}$$

$$d_g w^{(1)} + \rho^{(2)} = -\phi_z^{(2)} , \tag{5.3}$$

$$\nabla \cdot \mathbf{u}^{(2)} = 0 , \tag{5.4}$$

$$d_g \rho^{(1)} + d_1 \rho^{(g)} + w^{(2)} \bar{\varrho}_z = 0 , \tag{5.5}$$

where we have assumed $\delta = 1$, set $\beta_0 = 0$ for simplicity, and defined $d_1 \equiv \mathbf{u}^{(1)} \cdot \nabla$. Cross differentiating the horizontal momentum equations (5.1) and (5.2), we find

$$\left[d_g v^{(1)}\right]_x - \left[d_g u^{(1)}\right]_y + \left[d_1 v^{(g)}\right]_x - \left[d_1 u^{(g)}\right]_y - w_z^{(2)} = 0 ,$$

where we have used the incompressibility equation (5.4) to replace $u_x^{(2)} + v_y^{(2)}$ by $-w_z^{(2)}$. Now recalling the definition $\lambda(z) \equiv -1/\bar{\varrho}_z$, (5.5) can be written as

$$w^{(2)} = \lambda d_g \rho^{(1)} + \lambda d_1 \rho^{(g)} ,$$

so that

$$\left[d_g v^{(1)}\right]_x - \left[d_g u^{(1)}\right]_y - \left[\lambda d_g \rho^{(1)}\right]_z + \left[d_1 v^{(g)}\right]_x - \left[d_1 u^{(g)}\right]_y - \left[\lambda d_1 \rho^{(g)}\right]_z = 0 . \tag{5.6}$$

Next we use the first-order equations (2.12)–(2.16) to eliminate the first-order terms $\mathbf{U}^{(1)}$ in favor of various spatial derivatives of $\phi^{(g)}$ and $\phi^{(1)}$. Substituting the expressions

$$\begin{aligned} v^{(1)} &= \phi_x^{(1)} + d_g u^{(g)} , \\ u^{(1)} &= -\phi_y^{(1)} - d_g v^{(g)} , \\ w^{(1)} &= \lambda d_g \rho^{(g)} , \\ \rho^{(1)} &= -\phi_z^{(1)} \end{aligned} \tag{5.7}$$

into equation (5.6), and noting that

$$d_1 \equiv \mathbf{u}^{(1)} \cdot \nabla = -(\phi_y^{(1)} + d_g v^{(g)})\partial_x + (\phi_x^{(1)} + d_g u^{(g)})\partial_y + (\lambda d_g \rho^{(g)})\partial_z ,$$

we get

$$\begin{aligned} &\left[d_g(\phi_x^{(1)} + d_g u^{(g)})\right]_x + \left[d_g(\phi_y^{(1)} + d_g v^{(g)})\right]_y + \left[\lambda d_g \phi_z^{(1)}\right]_z \\ &\quad + \left[-(\phi_y^{(1)} + d_g v^{(g)}) v_x^{(g)} + (\phi_x^{(1)} + d_g u^{(g)}) v_y^{(g)} + (\lambda d_g \rho^{(g)}) v_z^{(g)}\right]_x \\ &\quad - \left[-(\phi_y^{(1)} + d_g v^{(g)}) u_x^{(g)} + (\phi_x^{(1)} + d_g u^{(g)}) u_y^{(g)} + (\lambda d_g \rho^{(g)}) u_z^{(g)}\right]_y \\ &\quad - \left[-(\phi_y^{(1)} + d_g v^{(g)}) \lambda \rho_x^{(g)} + (\phi_x^{(1)} + d_g u^{(g)}) \lambda \rho_y^{(g)} + (\lambda d_g \rho^{(g)}) \lambda \rho_z^{(g)}\right]_z = 0 . \end{aligned} \tag{5.8}$$

Expanding derivatives in (5.8), and using the zero-order relations

$$u_x^{(g)} = -v_y^{(g)} , \quad u_z^{(g)} = \rho_y^{(g)} , \quad v_z^{(g)} = -\rho_x^{(g)} ,$$

we cancel some terms and rearrange to get

$$
\begin{aligned}
&d_g\left[\phi^{(1)}_{xx} + \phi^{(1)}_{yy} + (\lambda\phi^{(1)}_z)_z\right] + \left[v^{(g)}_{xy} - u^{(g)}_{yy} - (\lambda\rho^{(g)}_y)_z\right]\phi^{(1)}_x \\
&- \left[v^{(g)}_{xx} - u^{(g)}_{xy} - (\lambda\rho^{(g)}_x)_z\right]\phi^{(1)}_y + \left[d_g(d_g u^{(g)})\right]_x + \left[d_g(d_g v^{(g)})\right]_y \\
&- \Big[(d_g v^{(g)})_x v^{(g)}_x - (d_g u^{(g)})_x v^{(g)}_y - (\lambda d_g\rho^{(g)})_x v^{(g)}_z \\
&\qquad - (d_g v^{(g)})_y u^{(g)}_x + (d_g u^{(g)})_y u^{(g)}_y + (\lambda d_g\rho^{(g)})_y u^{(g)}_z \\
&\qquad - (\lambda d_g v^{(g)})_z \rho^{(g)}_x + (\lambda d_g u^{(g)})_z \rho^{(g)}_y + (\lambda d_g\rho^{(g)})_z \lambda\rho^{(g)}_z\Big] \\
&- \Big[(d_g v^{(g)})v^{(g)}_{xx} - (d_g v^{(g)})u^{(g)}_{xy} - (\lambda d_g v^{(g)})\rho^{(g)}_{xz} \\
&\qquad - (d_g u^{(g)})v^{(g)}_{xy} + (d_g u^{(g)})u^{(g)}_{yy} + (\lambda d_g u^{(g)})\rho^{(g)}_{yz} \\
&\qquad - (\lambda d_g\rho^{(g)})v^{(g)}_{xz} + (\lambda d_g\rho^{(g)})u^{(g)}_{yz} + (\lambda d_g\rho^{(g)})(\lambda\rho^{(g)}_z)_z\Big] = 0\,.
\end{aligned}
\tag{5.9}
$$

Setting $\lambda = 1$ for the case of a linear density profile, and using (2.11), we can write (5.9) concisely as

$$
\begin{aligned}
&d_g(\Delta\phi^{(1)}) + (\Delta\phi^{(g)}_y)\phi^{(1)}_x - (\Delta\phi^{(g)}_x)\phi^{(1)}_y \\
&\quad = \nabla(d_g\phi^{(g)}_x)\cdot\nabla\phi^{(g)}_x + \nabla(d_g\phi^{(g)}_y)\cdot\nabla\phi^{(g)}_y + \nabla(d_g\phi^{(g)}_z)\cdot\nabla\phi^{(g)}_z \\
&\qquad + d_g(\nabla\phi^{(g)})\cdot\nabla(\Delta\phi^{(g)}) + \left[d_g(d_g\phi^{(g)}_y)\right]_x - \left[d_g(d_g\phi^{(g)}_x)\right]_y\,.
\end{aligned}
\tag{5.10}
$$

An equation similar to (5.10) is derived by McWilliams and Gent (see [19, eq. 4.4]), but there the pressure ϕ is not expanded in ε. We can interpret (5.10) as a linear evolution equation for $\Delta\phi^{(1)}$, with coefficients and nonhomogeneous term determined completely by the QGS solution $\mathbf{U}^{(g)}$, provided we include the conditions

$$
\phi^{(1)}_{0z} = 0 \quad \text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h, \qquad \int_B \phi^{(1)}_0 = 0\,,
\tag{5.11}
$$

to determine ϕ from $\Delta\phi$. The boundary condition in (5.11) follows from the SPE initial boundary condition $\rho_0 = 0$ and the first-order relation $\rho^{(1)} = -\phi^{(1)}_z$; the second condition is merely a normalization. Equation (5.10) thus has the form of a linear transport equation for $\Delta\phi^{(1)}$ with terms of bounded linear dependence on this unknown, as well as nonhomogeneous terms. The existence of solutions for (5.10) and (5.11) can therefore be established by standard arguments, using techniques as in §3; see Theorem 5.1 below.

Equation (5.10) plays the same role for the correction term $\mathbf{U}^{(1)}$ as the quasigeostrophic equation plays for the QGS solution $\mathbf{U}^{(g)}$, and is essential for our refined convergence theorem in the next section. There we specify $\mathbf{U}^{(1)}_0$ at time $t = 0$ in terms of $\mathbf{U}^{(g)}$, and invoke (5.10) for existence of a unique solution $\mathbf{U}^{(1)}(t)$ (see Theorem 6.2). Notice that $\mathbf{U}^{(1)}_0$ and $\mathbf{U}^{(g)}_0$ uniquely determine the initial first-order pressure $\phi^{(1)}_0$ through the elliptic problem

$$
-\Delta\phi^{(1)}_0 = -v^{(1)}_{0x} + u^{(1)}_{0y} + \rho^{(1)}_{0z} + \mathbf{u}^{(g)}_{0x}\cdot\nabla u^{(g)}_0 + \mathbf{u}^{(g)}_{0y}\cdot\nabla v^{(g)}_0 \quad \text{in } B,
\tag{5.12}
$$

with conditions (5.11), where (5.12) is the divergence of the formal first-order equations (5.7). Then the initial value problem (5.10) and (5.11) with data $\phi_0^{(1)}$ determines the first-order pressure $\phi^{(1)}(t)$ for all time, which in turn, together with $\mathbf{U}^{(g)}(t)$, determines $\mathbf{U}^{(1)}(t)$ through equations (5.7). For future referenc e, we state this as a theorem.

THEOREM 5.1. *Given time $T > 0$, a solution $\mathbf{U}^{(g)}$ of* QGS *in $C([0,T]; H^{s+1}(B))$, and initial data $\mathbf{U}_0^{(1)}$ in $H^s(B)$, for some $s \geq 3$, then there exists a unique solution $\mathbf{U}^{(1)} \equiv (\mathbf{u}^{(1)}, \rho^{(1)})$ in $C([0,T]; H^s(B))$ for the formal first-order velocity and density.*

5.2. The ageostrophic term $\mathbf{U}^{(a)}(\varepsilon)$. We would like to compare $\mathbf{U}^{(1)}$ with $\mathbf{U}^{(a)}(\varepsilon)$. To do this, we derive an equation for $\mathbf{U}^{(a)}(\varepsilon)$ similar to (5.10). Again setting $\beta_0 = 0$ for simplicity, we take the curl of the (exact) horizontal momentum equations in SPE (2.6)–(2.10) to get

$$\left(\frac{Dv}{Dt}\right)_x - \left(\frac{Du}{Dt}\right)_y + u_x + v_y = 0\,.$$

Substituting

$$u = u^{(g)} + \varepsilon u^{(a)}\,, \qquad v = v^{(g)} + \varepsilon v^{(a)}\,,$$

and using

$$u_x^{(g)} + v_y^{(g)} = 0\,, \qquad u_x^{(a)} + v_y^{(a)} = -w_z^{(a)}$$

results in

$$\left[d_g v^{(g)}\right]_x - \left[d_g u^{(g)}\right]_y + \varepsilon\left[d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right]_x - \varepsilon\left[d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right]_y = w_z^{(a)}\,. \tag{5.13}$$

From (2.10), we have

$$w^{(a)} = \lambda\frac{D\rho}{Dt} = \lambda d_g\rho^{(g)} + \varepsilon\lambda d_a\rho^{(g)} + \varepsilon\lambda\frac{D\rho^{(a)}}{Dt}\,,$$

and substituting this for $w^{(a)}$ in (5.13) gives us

$$\begin{aligned}&\left[d_g v^{(g)}\right]_x - \left[d_g u^{(g)}\right]_y - \left[\lambda d_g\rho^{(g)}\right]_z + \varepsilon\left[d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right]_x \\ &\quad - \varepsilon\left[d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right]_y - \varepsilon\left[\lambda d_a\rho^{(g)} + \lambda\frac{D\rho^{(a)}}{Dt}\right]_z = 0\,.\end{aligned} \tag{5.14}$$

We can use the relation $u_x^{(g)} = -v_y^{(g)}$ to rewrite the first two terms in (5.14) as

$$\left[d_g v^{(g)}\right]_x = d_g v_x^{(g)} + u_x^{(g)} v_x^{(g)} + v_x^{(g)} v_y^{(g)} = d_g v_x^{(g)}\,, \tag{5.15}$$

$$\left[d_g u^{(g)}\right]_y = d_g u_y^{(g)} + u_y^{(g)} u_x^{(g)} + v_y^{(g)} u_y^{(g)} = d_g u_y^{(g)}\,. \tag{5.16}$$

Also, using (2.20) and recalling that $d_g\lambda = 0$, we can rewrite the third term in (5.14) as

$$\left[\lambda d_g\rho^{(g)}\right]_z = d_g(\lambda\rho^{(g)})_z\,. \tag{5.17}$$

Now from (5.15)–(5.17) and the conservation of potential vorticity equation (2.22) with $\beta_0 = 0$, we see that (5.14) reduces to

$$\left[d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right]_x - \left[d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right]_y - \left[\lambda d_a \rho^{(g)} + \lambda \frac{D\rho^{(a)}}{Dt}\right]_z = 0\,,$$

which we rewrite as

(5.18)
$$\begin{aligned}&\left[d_g v^{(a)}\right]_x - \left[d_g u^{(a)}\right]_y - \left[\lambda d_g \rho^{(a)}\right]_z + \left[d_a v^{(g)}\right]_x - \left[d_a u^{(g)}\right]_y - \left[\lambda d_a \rho^{(g)}\right]_z \\ &\quad + \varepsilon\left\{\left[d_a v^{(a)}\right]_x - \left[d_a u^{(a)}\right]_y - \left[\lambda d_a \rho^{(a)}\right]_z\right\} = 0\,.\end{aligned}$$

Notice the similarity between (5.18) and (5.6). Mimicking the steps from our formal asymptotics derivation in the beginning of the section, we substitute into (5.18) the exact ageostrophic terms

(5.19)
$$\begin{aligned} u^{(a)} &= -\phi_y^{(a)} - d_g v^{(g)} - \varepsilon\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right), \\ v^{(a)} &= \phi_x^{(a)} + d_g u^{(g)} + \varepsilon\left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right), \\ w^{(a)} &= \lambda d_g \rho^{(g)} + \varepsilon\left(\lambda d_a \rho^{(g)} + \lambda \frac{D\rho^{(a)}}{Dt}\right), \\ \rho^{(a)} &= -\phi_z^{(a)} - \varepsilon \frac{Dw^{(a)}}{Dt} \end{aligned}$$

from SPE (2.6)–(2.10), where we recall that we are assuming $\delta = 1$. Upon substitution, the first line of (5.18) becomes

$$\begin{aligned}&\left[d_g\left(\phi_x^{(a)} + d_g u^{(g)}\right)\right]_x + \left[d_g\left(\phi_y^{(a)} + d_g v^{(g)}\right)\right]_y + \left[d_g \phi_z^{(a)}\right]_z \\ &+ \varepsilon\left[d_g\left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)\right]_x + \varepsilon\left[d_g\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)\right]_y + \varepsilon\left[d_g \frac{Dw^{(a)}}{Dt}\right]_z \\ &+ \left[-\left(\phi_y^{(1)} + d_g v^{(g)}\right)v_x^{(g)} + \left(\phi_x^{(1)} + d_g u^{(g)}\right)v_y^{(g)} + \left(\lambda d_g \rho^{(g)}\right)v_z^{(g)}\right]_x \\ &- \left[-\left(\phi_y^{(1)} + d_g v^{(g)}\right)u_x^{(g)} + \left(\phi_x^{(1)} + d_g u^{(g)}\right)u_y^{(g)} + \left(\lambda d_g \rho^{(g)}\right)u_z^{(g)}\right]_y \\ &- \left[-\left(\phi_y^{(1)} + d_g v^{(g)}\right)\lambda\rho_x^{(g)} + \left(\phi_x^{(1)} + d_g u^{(g)}\right)\lambda\rho_y^{(g)} + \left(\lambda d_g \rho^{(g)}\right)\lambda\rho_z^{(g)}\right]_z \\ &+ \varepsilon\left[-\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)v_x^{(g)} + \left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)v_y^{(g)} + \left(\lambda d_a \rho^{(g)} + \lambda \frac{D\rho^{(a)}}{Dt}\right)v_z^{(g)}\right]_x \\ &- \varepsilon\left[-\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)u_x^{(g)} + \left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)u_y^{(g)} + \left(\lambda d_a \rho^{(g)} + \lambda \frac{D\rho^{(a)}}{Dt}\right)u_z^{(g)}\right]_y \\ &- \varepsilon\left[-\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)\lambda\rho_x^{(g)} + \left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)\lambda\rho_y^{(g)} + \left(\lambda d_a \rho^{(g)} + \lambda \frac{D\rho^{(a)}}{Dt}\right)\lambda\rho_z^{(g)}\right]_z.\end{aligned}$$

The form of the terms above that are not multiplied by ε is identical to the form of (5.8) for the formal variable $\phi^{(1)}$. Therefore, setting $\lambda = 1$ as before, (5.18) can be

written concisely as

$$
\begin{aligned}
&d_g(\Delta\phi^{(a)}) + (\Delta\phi_y^{(g)})\phi_x^{(a)} - (\Delta\phi_x^{(g)})\phi_y^{(a)} \\
&= \nabla(d_g\phi_x^{(g)})\cdot\nabla\phi_x^{(g)} + \nabla(d_g\phi_y^{(g)})\cdot\nabla\phi_y^{(g)} + \nabla(d_g\phi_z^{(g)})\cdot\nabla\phi_z^{(g)} \\
&\quad + d_g(\nabla\phi^{(g)})\cdot\nabla(\Delta\phi^{(g)}) + \Big[d_g(d_g\phi_y^{(g)})\Big]_x - \Big[d_g(d_g\phi_x^{(g)})\Big]_y - \varepsilon Q\,,
\end{aligned}
\tag{5.20}
$$

where

(5.21)

$$
\begin{aligned}
Q \equiv{}& \left[d_g\left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)\right]_x + \left[d_g\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)\right]_y + \left[d_g\frac{Dw^{(a)}}{Dt}\right]_z \\
&- \left[-\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)v_x^{(g)} + \left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)v_y^{(g)} + \left(d_a\rho^{(g)} + \frac{D\rho^{(a)}}{Dt}\right)v_z^{(g)}\right]_x \\
&+ \left[-\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)u_x^{(g)} + \left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)u_y^{(g)} + \left(d_a\rho^{(g)} + \frac{D\rho^{(a)}}{Dt}\right)u_z^{(g)}\right]_y \\
&+ \left[-\left(d_a v^{(g)} + \frac{Dv^{(a)}}{Dt}\right)\rho_x^{(g)} + \left(d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right)\rho_y^{(g)} + \left(d_a\rho^{(g)} + \frac{D\rho^{(a)}}{Dt}\right)\rho_z^{(g)}\right]_z \\
&- \left[d_a v^{(a)}\right]_x + \left[d_a u^{(a)}\right]_y + \left[d_a\rho^{(a)}\right]_z .
\end{aligned}
\tag{5.21}
$$

In the next section, we will use (5.10) and (5.20) to get an estimate for the difference $(\mathbf{U}^{(a)}(\varepsilon) - \mathbf{U}^{(1)})$ in $H^2(B)$.

6. Higher accuracy approximation of SPE solutions. In this section we present a refinement to Theorem 4.5. We show in Theorem 6.2 below that if SPE is initialized to suppress fast solutions, then the QGS solution, together with its first correction in ε, provides $O(\varepsilon^2)$ approximations of SPE solutions. As before, we assume a linear density profile and zero β-factor.

6.1. Suppression of fast-scale motion. The main part of the proof of our refinement theorem is showing that on any prespecified time interval $[0,T]$, if we keep ε sufficiently small, $|\mathbf{U}_t^{(a)}|_3$ is bounded on $[0,T]$ independently of ε. We now give a simple condition on the initial SPE data that is sufficient for such suppression of fast-scale motion.

THEOREM 6.1. *Assume initial* QGS *data* $\mathbf{U}_0^{(g)} \equiv (\mathbf{u}_0^{(g)}, \rho_0^{(g)})$ *in* $H^{s+1}(B)$, *some* $s \geq 3$, *of the form*

$$
\mathbf{u}_0^{(g)} = (-\phi_{0y}^{(g)}, \phi_{0x}^{(g)}, 0), \qquad \rho_0^{(g)} = -\phi_{0z}^{(g)}\,.
$$

For $\lambda=1$ *and* $\beta_0=0$, *let* $\mathbf{U}^{(g)} \equiv (\mathbf{u}^{(g)}, \rho^{(g)})$ *be the corresponding solution of* QGS (3.1)–(3.7), *where*

$$
\mathbf{u}^{(g)} = (-\phi_y^{(g)}, \phi_x^{(g)}, 0)\,, \qquad \rho^{(g)} = -\phi_z^{(g)}
$$

Suppose initial SPE *data* $\mathbf{U}_0(\varepsilon) \equiv (\mathbf{u}_0(\varepsilon), \rho_0(\varepsilon))$ *is given in* $H^s(B)$ *such that*

$$
\begin{aligned}
&|v_0 - (\phi_{0x}^{(g)} - \varepsilon[d_g\phi_y^{(g)}]_{t=0})|_s = O(\varepsilon^2)\,,\\
&|u_0 + (\phi_{0y}^{(g)} + \varepsilon[d_g\phi_x^{(g)}]_{t=0})|_s = O(\varepsilon^2)\,,\\
&|\rho_0 + \phi_{0z}^{(g)}|_s = O(\varepsilon^2)\,,\\
&|w_0 + \varepsilon[d_g\phi_z^{(g)}]_{t=0}|_s = O(\varepsilon^2)\,,
\end{aligned}
\tag{6.1}
$$

where $d_{g|t=0} = \partial_t - \phi_{0y}^{(g)}\partial_x + \phi_{0x}^{(g)}\partial_y$. *Then the corresponding solutions* $\mathbf{U}(\varepsilon) = \mathbf{U}^{(g)} + \varepsilon\mathbf{U}^{(a)}(\varepsilon)$ *of* SPE (4.1)–(4.8) *have the property that* $\mathbf{U}_t^{(a)}(\varepsilon)$ *is initially bounded in* $H^{s-1}(B)$ *independently of* ε.

Proof. Using (6.1) to substitute for u, v, w, and ρ at time $t = 0$ in the elliptic problem (4.9) for ϕ, we find

$$
\begin{aligned}
&|\Delta(\phi_0(\varepsilon) - \phi_0^{(g)})|_{s-1} = O(\varepsilon^2)\,,\\
&|(\phi_0(\varepsilon) - \phi_0^{(g)})_z|_{s-1/2,\partial B} = O(\varepsilon^2)\,,\\
&\int_B (\phi_0(\varepsilon) - \phi_0^{(g)}) = 0\ .
\end{aligned}
\tag{6.2}
$$

From elliptic theory, (6.2) implies

$$|\phi_0(\varepsilon) - \phi_0^{(g)}|_{s+1} = O(\varepsilon^2)\,.$$

Then, for example, SPE (4.1) at time $t = 0$ can be written

$$\varepsilon[d_g u^{(g)}]_{t=0} + \varepsilon^2\left[d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right]_{t=0} - v_0 = -\phi_{0x}^{(g)} + O(\varepsilon^2)$$

in $H^s(B)$. Now replacing the geostrophic term $d_g u^{(g)}$ with its equivalent $-d_g\phi_y^{(g)}$, and using the first equation in (6.1), we have

$$\left[d_a u^{(g)} + \frac{Du^{(a)}}{Dt}\right]_{t=0} = O(1) \tag{6.3}$$

in $H^s(B)$. This implies that $u_t^{(a)}(\varepsilon)$ is initially bounded in $H^s(B)$ independently of ε, since the same is true of all the other terms in (6.3). In exactly the same way, we can use SPE (4.2)–(4.4) with (6.1) to prove that $v_t^{(a)}$, $w_t^{(a)}$, and $\rho_t^{(a)}$ are initially bounded in $H^s(B)$ independently of ε. □

6.2. The refined convergence theorem. We now state and prove a refined convergence theorem.

THEOREM 6.2. *Assume we are given time* $T > 0$ *and initial* QGS *velocity and density* $\mathbf{U}_0^{(g)} \equiv (\mathbf{u}_0^{(g)}, \rho_0^{(g)})$ *in* $H^6(B)$ *of the form* $\mathbf{u}_0^{(g)} = (-\phi_{0y}^{(g)}, \phi_{0x}^{(g)}, 0)$, $\rho_0^{(g)} = -\phi_{0z}^{(g)}$ *for some pressure function* $\phi_0^{(g)}$ *satisfying* (3.32). *For* $\lambda = 1$ *and* $\beta_0 = 0$, *let* $\mathbf{U}^{(g)} \equiv (\mathbf{u}^{(g)}, \rho^{(g)})$ *in* $C([0,T]; H^6(B))$ *be the solution of* QGS (3.1)–(3.7) *with initial data* $\mathbf{U}_0^{(g)}$. *Define* $\mathbf{U}_0^{(1)} \equiv (\mathbf{u}_0^{(1)}, \rho_0^{(1)})$ *in* $H^5(B)$ *by*

$$
\begin{aligned}
&u_0^{(1)} = -[d_g\phi_x^{(g)}]_{t=0}\,, \qquad v_0^{(1)} = -[d_g\phi_y^{(g)}]_{t=0}\,,\\
&w_0^{(1)} = -[d_g\phi_z^{(g)}]_{t=0}\,, \qquad \rho_0^{(1)} = 0\,,
\end{aligned}
\tag{6.4}
$$

and let $\mathbf{U}^{(1)} \equiv (\mathbf{u}^{(1)}, \rho^{(1)})$ *in* $C([0,T]; H^5(B))$ *be the corresponding formal first-order velocity and density guaranteed by Theorem* 5.1. *We consider initial* SPE *data* $\mathbf{U}_0(\varepsilon) \equiv (\mathbf{u}_0(\varepsilon), \rho_0(\varepsilon))$ *in* $H^5(B)$ *satisfying* (4.10), *such that*

$$|\mathbf{U}_0(\varepsilon) - (\mathbf{U}_0^{(g)} + \varepsilon \mathbf{U}_0^{(1)})|_4 = O(\varepsilon^2) . \tag{6.5}$$

Let $\mathbf{U}(\varepsilon) \equiv (\mathbf{u}(\varepsilon), \rho(\varepsilon))$ *in* $C([0,T]; H^5(B))$ *be the corresponding solutions of* SPE (4.1)–(4.8), *for all* $\varepsilon \le \varepsilon_0$, *as guaranteed by Theorem* 4.5. *Then the* SPE *solutions* $\mathbf{U}(\varepsilon)$ *converge to* $\mathbf{U}^{(g)} + \varepsilon \mathbf{U}^{(1)}$ *in* $C([0,T]; H^3(B))$ *with* $O(\varepsilon^2)$ *accuracy, i.e.,*

$$|\mathbf{U}(\varepsilon) - (\mathbf{U}^{(g)} + \varepsilon \mathbf{U}^{(1)})|_{3,T} = O(\varepsilon^2) . \tag{6.6}$$

Proof. We will show at the end of this section that if $|\mathbf{U}_t^{(a)}(\varepsilon)|_3$ is bounded on $[0,T]$ independently of ε, then $|\mathbf{U}^{(a)}(\varepsilon) - \mathbf{U}^{(1)}|_{3,T}$ is $O(\varepsilon)$, which gives us (6.6). So our current task is reduced to estimating $|\mathbf{U}_t^{(a)}|_3$. Theorem 6.1 implies that it is bounded independently of ε at $t = 0$. We proceed as we did for estimating $|\mathbf{U}^{(a)}|_4$ in §4. The argument here closely parallels the one given there, and to avoid redundancy, some of the details are omitted. We begin with the time derivative of equations (4.21)–(4.24), which are

$$\left(\frac{Du^{(g)}}{Dt}\right)_{t,\alpha} + \varepsilon \left(\frac{Du^{(a)}}{Dt}\right)_{t,\alpha} - v_{t,\alpha}^{(a)} = -D^\alpha \phi_{xt}^{(a)} , \tag{6.7}$$

$$\left(\frac{Dv^{(g)}}{Dt}\right)_{t,\alpha} + \varepsilon \left(\frac{Dv^{(a)}}{Dt}\right)_{t,\alpha} + u_{t,\alpha}^{(a)} = -D^\alpha \phi_{yt}^{(a)} , \tag{6.8}$$

$$\varepsilon \left(\frac{Dw^{(a)}}{Dt}\right)_{t,\alpha} + \rho_{t,\alpha}^{(a)} = -D^\alpha \phi_{zt}^{(a)} , \tag{6.9}$$

$$\left(\frac{D\rho^{(g)}}{Dt}\right)_{t,\alpha} + \varepsilon \left(\frac{D\rho^{(a)}}{Dt}\right)_{t,\alpha} - w_{t,\alpha}^{(a)} = 0 , \tag{6.10}$$

where we use the subscript (t,α) to mean $\partial_t D^\alpha$. Multiplying these equations by $u_{t,\alpha}^{(a)}$, $v_{t,\alpha}^{(a)}$, $w_{t,\alpha}^{(a)}$, and $\rho_{t,\alpha}^{(a)}$, respectively, then summing,

$$\begin{aligned} &\varepsilon \left\{ u_{t,\alpha}^{(a)} \left(\frac{Du^{(a)}}{Dt}\right)_{t,\alpha} + v_{t,\alpha}^{(a)} \left(\frac{Dv^{(a)}}{Dt}\right)_{t,\alpha} + w_{t,\alpha}^{(a)} \left(\frac{Dw^{(a)}}{Dt}\right)_{t,\alpha} + \rho_{t,\alpha}^{(a)} \left(\frac{D\rho^{(a)}}{Dt}\right)_{t,\alpha} \right\} \\ &\qquad + u_{t,\alpha}^{(a)} \left(\frac{Du^{(g)}}{Dt}\right)_{t,\alpha} + v_{t,\alpha}^{(a)} \left(\frac{Dv^{(g)}}{Dt}\right)_{t,\alpha} + \rho_{t,\alpha}^{(a)} \left(\frac{D\rho^{(g)}}{Dt}\right)_{t,\alpha} \\ &\quad = -\left(u_{t,\alpha}^{(a)} D^\alpha \phi_{xt}^{(a)} + v_{t,\alpha}^{(a)} D^\alpha \phi_{yt}^{(a)} + w_{t,\alpha}^{(a)} D^\alpha \phi_{zt}^{(a)} \right). \end{aligned} \tag{6.11}$$

This equation is analogous to (4.25), and we must similarly work on the "large" terms

$$u_{t,\alpha}^{(a)} \left(\frac{Du^{(g)}}{Dt}\right)_{t,\alpha} + v_{t,\alpha}^{(a)} \left(\frac{Dv^{(g)}}{Dt}\right)_{t,\alpha} + \rho_{t,\alpha}^{(a)} \left(\frac{D\rho^{(g)}}{Dt}\right)_{t,\alpha} . \tag{6.12}$$

Using (6.7)–(6.9), we substitute for $u_{t,\alpha}^{(a)}$, $v_{t,\alpha}^{(a)}$, and $\rho_{t,\alpha}^{(a)}$ in (6.12), and follow precisely the same steps as we did in §4 to rewrite (6.11) as

(6.13)

$$
\begin{aligned}
\varepsilon \Bigg\{ & u_{t,\alpha}^{(a)} \left(\frac{Du^{(a)}}{Dt} \right)_{t,\alpha} + v_{t,\alpha}^{(a)} \left(\frac{Dv^{(a)}}{Dt} \right)_{t,\alpha} + w_{t,\alpha}^{(a)} \left(\frac{Dw^{(a)}}{Dt} \right)_{t,\alpha} + \rho_{t,\alpha}^{(a)} \left(\frac{D\rho^{(a)}}{Dt} \right)_{t,\alpha} \\
& + \left(\frac{Du^{(a)}}{Dt} \right)_{t,\alpha} \left(d_g v^{(g)} \right)_{t,\alpha} - \left(\frac{Dv^{(a)}}{Dt} \right)_{t,\alpha} \left(d_g u^{(g)} \right)_{t,\alpha} \\
& - \left(\frac{Dw^{(a)}}{Dt} \right)_{t,\alpha} \left(d_g \rho^{(g)} \right)_{t,\alpha} \\
& - \left(\frac{Du^{(g)}}{Dt} - v_{t,\alpha}^{(a)} \right)_{t,\alpha} \left(d_a v^{(g)} \right)_{t,\alpha} + \left(\frac{Dv^{(g)}}{Dt} + u_{t,\alpha}^{(a)} \right)_{t,\alpha} \left(d_a u^{(g)} \right)_{t,\alpha} \\
& + \rho_{t,\alpha}^{(a)} \left(d_a \rho^{(g)} \right)_{t,\alpha} \Bigg\} \\
= & -\nabla \phi_{t,\alpha}^{(a)} \cdot \partial_t D^\alpha (d_g v^{(g)}, -d_g u^{(g)}, -d_g \rho^{(g)}) \\
& - \left(u_{t,\alpha}^{(a)} D^\alpha \phi_{xt}^{(a)} + v_{t,\alpha}^{(a)} D^\alpha \phi_{yt}^{(a)} + w_{t,\alpha}^{(a)} D^\alpha \phi_{zt}^{(a)} \right) .
\end{aligned}
$$

This equation is analogous to (4.30), and we will integrate (6.13) to get the desired growth estimate for $|\mathbf{U}_t^{(a)}|_3$. The first term on the right-hand side of (6.13) becomes, after integrating by parts and applying the divergence theorem,

$$
\begin{aligned}
& - \int \int_{\partial B} \phi_{t,\alpha}^{(a)} \, \partial_t D^\alpha (d_g v^{(g)}, -d_g u^{(g)}, -d_g \rho^{(g)}) \cdot \mathbf{n} \, ds \\
& \quad + \int \int \int_B \phi_{t,\alpha}^{(a)} \, \partial_t D^\alpha \left[d_g (v_x^{(g)} - u_y^{(g)} - \rho_z^{(g)}) \right] d\mathbf{x} .
\end{aligned}
$$

The last integral is zero because the quasigeostrophic equation (2.23) holds in B for all $t \in [0, T]$. The boundary integral reduces to

$$
\int \int_{\Sigma_h} \phi_{t,\alpha}^{(a)} \, \partial_t D^\alpha (d_g \rho^{(g)}) \, dx \, dy - \int \int_{\Sigma_0} \phi_{t,\alpha}^{(a)} \, \partial_t D^\alpha (d_g \rho^{(g)}) \, dx \, dy . \tag{6.14}
$$

The same argument used for (4.32) can be used to show that the boundary integrals (6.14) are zero for $|\alpha| \leq 3$. Also as before, the second term on the right-hand side of (6.13) vanishes upon integration.

We proceed with our energy estimate for $\mathbf{U}_t^{(a)}$. To integrate, for example, the term $u_{t,\alpha}^{(a)} (Du^{(a)}/Dt)_{t,\alpha}$ in (6.13), we write

$$
\begin{aligned}
u_{t,\alpha}^{(a)} \left(\frac{Du^{(a)}}{Dt} \right)_{t,\alpha} = & u_{t,\alpha}^{(a)} \left(\frac{Du_t^{(a)}}{Dt} \right)_\alpha + u_{t,\alpha}^{(a)} \left(\mathbf{u}_t \cdot \nabla u^{(a)} \right)_\alpha \\
= & u_{t,\alpha}^{(a)} \frac{\partial u_{t,\alpha}^{(a)}}{\partial t} + u_{t,\alpha}^{(a)} \left(\mathbf{u} \cdot \nabla u_t^{(a)} \right)_\alpha + u_{t,\alpha}^{(a)} \left(\mathbf{u}_t \cdot \nabla u^{(a)} \right)_\alpha \\
= & u_{t,\alpha}^{(a)} \frac{\partial u_{t,\alpha}^{(a)}}{\partial t} + u_{t,\alpha}^{(a)} \mathbf{u} \cdot \nabla u_{t,\alpha}^{(a)} + u_{t,\alpha}^{(a)} G_{1,\alpha} + u_{t,\alpha}^{(a)} \left(\mathbf{u}_t \cdot \nabla u^{(a)} \right)_\alpha ,
\end{aligned}
$$

where

$$
G_{1,\alpha} = \left(\mathbf{u} \cdot \nabla u_t^{(a)} \right)_\alpha - \mathbf{u} \cdot \nabla u_{t,\alpha}^{(a)} , \tag{6.15}
$$

and the integral of $u_{t,\alpha}^{(a)}\,(Du^{(a)}/Dt)_{t,\alpha}$ is

$$\left(u_{t,\alpha}^{(a)}\,,\,\left(\frac{Du^{(a)}}{Dt}\right)_{t,\alpha}\right)_{L_2} = \frac{1}{2}\frac{d}{dt}|u_{t,\alpha}^{(a)}|_0^2 + \left(u_{t,\alpha}^{(a)}\,,\,G_{1,\alpha}\right)_{L^2} + \left(u_{t,\alpha}^{(a)}\,,\,\left(\mathbf{u}_t\cdot\nabla u^{(a)}\right)_\alpha\right)_{L^2}.$$

With similar definitions for $G_{i,\alpha}$, $i = 2,3,4$, we integrate (6.13) over B for all α, $|\alpha| \le 3$, to get

(6.16)

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left\{ |u_{t,\alpha}^{(a)}|_0^2 + |v_{t,\alpha}^{(a)}|_0^2 + |w_{t,\alpha}^{(a)}|_0^2 + |\rho_{t,\alpha}^{(a)}|_0^2 \right\} + \left(\left(\frac{Du^{(a)}}{Dt}\right)_{t,\alpha}\,,\,\left(d_g v^{(g)}\right)_{t,\alpha}\right)_{L^2} \\
&\quad - \left(\left(\frac{Dv^{(a)}}{Dt}\right)_{t,\alpha}\,,\,\left(d_g u^{(g)}\right)_{t,\alpha}\right)_{L^2} - \left(\left(\frac{Dw^{(a)}}{Dt}\right)_{t,\alpha}\,,\,\left(d_g \rho^{(g)}\right)_{t,\alpha}\right)_{L^2} \\
&\quad + \left(\left(\frac{Du^{(g)}}{Dt}\right)_{t,\alpha}\,,\,\left(d_a v^{(g)}\right)_{t,\alpha}\right)_{L^2} + \left(\left(\frac{Dv^{(g)}}{Dt}\right)_{t,\alpha}\,,\,\left(d_a u^{(g)}\right)_{t,\alpha}\right)_{L^2} \\
&\quad + \left(v_{t,\alpha}^{(a)}\,,\,\left(d_a v^{(g)}\right)_{t,\alpha}\right)_{L^2} + \left(u_{t,\alpha}^{(a)}\,,\,\left(d_a u^{(g)}\right)_{t,\alpha}\right)_{L^2} + \left(\rho_{t,\alpha}^{(a)}\,,\,\left(d_a \rho^{(g)}\right)_{t,\alpha}\right)_{L^2} \\
&= -\left(u_{t,\alpha}^{(a)}\,,\,G_{1,\alpha}\right)_{L^2} - \left(v_{t,\alpha}^{(a)}\,,\,G_{2,\alpha}\right)_{L^2} - \left(w_{t,\alpha}^{(a)}\,,\,G_{3,\alpha}\right)_{L^2} - \left(\rho_{t,\alpha}^{(a)}\,,\,G_{4,\alpha}\right)_{L^2} \\
&\quad - \left(u_{t,\alpha}^{(a)}\,,\,\left(\mathbf{u}_t\cdot\nabla u^{(a)}\right)_\alpha\right)_{L^2} - \left(v_{t,\alpha}^{(a)}\,,\,\left(\mathbf{u}_t\cdot\nabla v^{(a)}\right)_\alpha\right)_{L^2} \\
&\quad - \left(w_{t,\alpha}^{(a)}\,,\,\left(\mathbf{u}_t\cdot\nabla w^{(a)}\right)_\alpha\right)_{L^2} - \left(\rho_{t,\alpha}^{(a)}\,,\,\left(\mathbf{u}_t\cdot\nabla \rho^{(a)}\right)_\alpha\right)_{L^2}.
\end{aligned}$$

Equation (6.16) is analogous to (4.35), with the G's analogous to the F's defined there. In preparation for the time integration we rewrite, for example, the term

$$\left(\left(u_{tt}^{(a)}\right)_\alpha\,,\,\left(d_g v^{(g)}\right)_{t,\alpha}\right)_{L^2}$$

occurring in the second term of (6.16) as

$$\frac{d}{dt}\left(u_{t,\alpha}^{(a)}\,,\,\left(d_g v^{(g)}\right)_{t,\alpha}\right)_{L^2} - \left(u_{t,\alpha}^{(a)}\,,\,\left(d_g v^{(g)}\right)_{tt,\alpha}\right)_{L^2}.$$

Doing the same for the third and fourth terms in (6.16), and integrating the entire equation with respect to time, we get

$$\text{(6.17)}\quad \frac{1}{2}\left\{ |u_{t,\alpha}^{(a)}|_0^2 + |v_{t,\alpha}^{(a)}|_0^2 + |w_{t,\alpha}^{(a)}|_0^2 + |\rho_{t,\alpha}^{(a)}|_0^2 \right\}\Big|_0^t = B_{1,\alpha}+B_{2,\alpha}+B_{3,\alpha}+B_{4,\alpha}+B_{5,\alpha},$$

where

$$B_{1,\alpha} = \left\{ -\left(u_{t,\alpha}^{(a)}, \left(d_g v^{(g)}\right)_{t,\alpha}\right)_{L^2} + \left(v_{t,\alpha}^{(a)}, \left(d_g u^{(g)}\right)_{t,\alpha}\right)_{L^2} + \left(w_{t,\alpha}^{(a)}, \left(d_g \rho^{(g)}\right)_{t,\alpha}\right)_{L^2} \right\}\Big|_0^t$$
$$+ \int_0^t \left\{ \left(u_{t,\alpha}^{(a)}, \left(d_g v^{(g)}\right)_{tt,\alpha}\right)_{L^2} - \left(v_{t,\alpha}^{(a)}, \left(d_g u^{(g)}\right)_{tt,\alpha}\right)_{L^2} \right.$$
$$\left. - \left(w_{t,\alpha}^{(a)}, \left(d_g \rho^{(g)}\right)_{tt,\alpha}\right)_{L^2} \right\} d\tau,$$

$$B_{2,\alpha} = \int_0^t \left\{ \left(-\left(\mathbf{u}\cdot\nabla u^{(a)}\right)_{t,\alpha}, \left(d_g v^{(g)}\right)_{t,\alpha}\right)_{L^2} + \left(\left(\mathbf{u}\cdot\nabla v^{(a)}\right)_{t,\alpha}, \left(d_g u^{(g)}\right)_{t,\alpha}\right)_{L^2} \right.$$
$$\left. + \left(\left(\mathbf{u}\cdot\nabla w^{(a)}\right)_{t,\alpha}, \left(d_g \rho^{(g)}\right)_{t,\alpha}\right)_{L^2} \right\} d\tau,$$

$$B_{3,\alpha} = -\int_0^t \left\{ \left(\left(\frac{Du^{(g)}}{Dt}\right)_{t,\alpha}, \left(d_a v^{(g)}\right)_{t,\alpha}\right)_{L^2} - \left(\left(\frac{Dv^{(g)}}{Dt}\right)_{t,\alpha}, \left(d_a u^{(g)}\right)_{t,\alpha}\right)_{L^2} \right\} d\tau$$
$$- \int_0^t \left\{ \left(v_{t,\alpha}^{(a)}, \left(d_a v^{(g)}\right)_{t,\alpha}\right)_{L^2} + \left(u_{t,\alpha}^{(a)}, \left(d_a u^{(g)}\right)_{t,\alpha}\right)_{L^2} \right.$$
$$\left. + \left(\rho_{t,\alpha}^{(a)}, \left(d_a \rho^{(g)}\right)_{t,\alpha}\right)_{L^2} \right\} d\tau,$$

$$B_{4,\alpha} = -\int_0^t \left\{ \left(u_{t,\alpha}^{(a)}, G_{1,\alpha}\right)_{L^2} + \left(v_{t,\alpha}^{(a)}, G_{2,\alpha}\right)_{L^2} + \left(w_{t,\alpha}^{(a)}, G_{3,\alpha}\right)_{L^2} + \left(\rho_{t,\alpha}^{(a)}, G_{4,\alpha}\right)_{L^2} \right\} d\tau,$$

$$B_{5,\alpha} = -\int_0^t \left\{ \left(u_{t,\alpha}^{(a)}, \left(\mathbf{u}_t\cdot\nabla u^{(a)}\right)_\alpha\right)_{L^2} + \left(v_{t,\alpha}^{(a)}, \left(\mathbf{u}_t\cdot\nabla v^{(a)}\right)_\alpha\right)_{L^2} \right.$$
$$\left. + \left(w_{t,\alpha}^{(a)}, \left(\mathbf{u}_t\cdot\nabla w^{(a)}\right)_\alpha\right)_{L^2} + \left(\rho_{t,\alpha}^{(a)}, \left(\mathbf{u}_t\cdot\nabla \rho^{(a)}\right)_\alpha\right)_{L^2} \right\} d\tau.$$

If we define

$$S^2 = |u_t^{(a)}|_3^2 + |v_t^{(a)}|_3^2 + |w_t^{(a)}|_3^2 + |\rho_t^{(a)}|_3^2,$$

then upon summing over α, $|\alpha| \le 3$, the left-hand side of (6.17) becomes $\frac{1}{2}[S^2(t) - S^2(0)]$. We now proceed to estimate $B_{i,\alpha}$, $i = 1, 2, 3$, in terms of S.

For $B_{1,\alpha}$, we get

$$\sum_{|\alpha|\le 3} |B_{1,\alpha}| \le C\left(S(0) + S(t) + \int_0^t S(\tau)\, d\tau\right), \tag{6.18}$$

where C depends on $|\mathbf{U}_0^{(g)}|_6$.

In expression $B_{2,\alpha}$, terms like $(\mathbf{u}\cdot\nabla u^{(a)})_{t,\alpha}$ involve fourth-order spatial derivatives of $u_t^{(a)}$ when $|\alpha| = 3$, and estimation in $H^3(B)$ requires the same trick we used in §4 for the term $A_{2,\alpha}$ in (4.36). Assuming that $\alpha_3 \ne 3$, and, for example, $\alpha_1 \ge 1$, we integrate by parts once in x, leaving a second derivative of $(\mathbf{u}\cdot\nabla u^{(a)})_t$. The resulting term can be estimated by $CS + C\varepsilon S^2$. When $\alpha_3 = 3$, we integrate by parts in z. The resulting interior terms can be estimated as before, but the surface integrals involve third-order derivatives of $u_t^{(a)}$ on the boundary, in the term $\partial_z^2(\mathbf{u}\cdot\nabla u^{(a)})_t$. However, these

surface integrals can be estimated in terms of $|u_t^{(a)}|_3$ due to the boundary condition $w = 0$. After these modifications, each surface integral involves at most second-order derivatives of $u_t^{(a)}$, and from the trace theorem, the L^2 norm of these second-order terms on the surface is dominated by $|u_t^{(a)}|_3$.

In summary, we obtain

$$\sum_{|\alpha|\le 3} |B_{2,\alpha}| \;\le\; C\int_0^t [\, S(\tau) + \varepsilon S^2(\tau)\,]\, d\tau\,. \tag{6.19}$$

We can directly estimate $B_{3,\alpha}$ to get

$$\sum_{|\alpha|\le 3} |B_{3,\alpha}| \;\le\; C\int_0^t [\, S(\tau) + S^2(\tau)\,]\, d\tau\,. \tag{6.20}$$

For $B_{4,\alpha}$, we need estimates for $G_{i,\alpha}$, $i = 1,2,3,4$. From definition (6.15) for $G_{1,\alpha}$ and calculus inequality (3.15), we obtain

$$|G_{1,\alpha}|_0 \;\le\; C(\,|\mathbf{u}|_3\,|\nabla u_t^{(a)}|_\infty + |\nabla \mathbf{u}|_\infty\,|u_t^{(a)}|_2) \;\le\; C|u_t^{(a)}|_3 \;\le\; CS\,,$$

where we have used (4.46), and C now depends on $|\mathbf{U}_0^{(g)}|_6$, $|\mathbf{U}_0^{(a)}(\varepsilon)|_4$, ε_0, and T. Recall that ε_0 is chosen to satisfy (4.45); therefore, we have

$$\sum_{|\alpha|\le 3} |B_{4,\alpha}| \;\le\; C\int_0^t S^2(\tau)\, d\tau\,. \tag{6.21}$$

Finally, from expression $B_{5,\alpha}$, we get

$$\sum_{|\alpha|\le 3} |B_{5,\alpha}| \;\le\; C\int_0^t [\, S(\tau) + \varepsilon S^2(\tau)\,]\, d\tau\,. \tag{6.22}$$

Combining (6.17) with estimates (6.18)–(6.22), we have

$$S^2(t) \;\le\; C\left(\, S(0) + S^2(0) + S(t) + \int_0^t [\, S(\tau) + S^2(\tau)\,]\, d\tau\right)\,.$$

Using the inequality $ab \le (\delta a^2/2) + (b^2/2\delta)$, this inequality reduces to

$$S^2(t) \;\le\; C\left(1 + S^2(0) + \int_0^t S^2(\tau)\, d\tau\right)\,.$$

The Gronwall inequality gives us $S^2(t) \le Ce^{Ct}$, where C now also depends on $S(0) \equiv |[\mathbf{U}_t^{(a)}]_{t=0}|_3$. Recall that initial condition (6.5), together with Theorem 6.1, ensures $S(0)$ is bounded uniformly in ε. Thus we have established that there is a constant M, depending only on $|\mathbf{U}_0^{(g)}|_6$, $|\mathbf{U}_0^{(a)}(\varepsilon)|_4$, $|[\mathbf{U}_t^{(a)}(\varepsilon)]_{t=0}|_3$, ε_0, and T, such that

$$|\mathbf{U}_t^{(a)}(\varepsilon)|_{3,T} \;\le\; M\,, \qquad 0 < \varepsilon \le \varepsilon_0\,. \tag{6.23}$$

We will now use (6.23) to show that $|\mathbf{U}^{(a)}(\varepsilon) - \mathbf{U}^{(1)}|_{3,T}$ is $O(\varepsilon)$. We begin by subtracting the exact vorticity equation (5.20) from the formal vorticity equation (5.10), which gives us

$$d_g\Big[\Delta(\phi^{(1)} - \phi^{(a)})\Big] + \Delta\phi_y^{(g)}(\phi^{(1)} - \phi^{(a)})_x - \Delta\phi_x^{(g)}(\phi^{(1)} - \phi^{(a)})_y = \varepsilon Q\,.$$

Defining $\xi \equiv \Delta(\phi^{(1)} - \phi^{(a)})$ and recalling that $\mathbf{u}^{(g)} = (-\phi_y^{(g)}, \phi_x^{(g)}, 0)$, we can write

$$\xi_t + \mathbf{u}^{(g)} \cdot \nabla \xi - (\Delta \mathbf{u}^{(g)}) \cdot \nabla(\Delta^{-1}\xi) = \varepsilon Q .$$

Differentiating, we have

$$D^\alpha \xi_t + D^\alpha(\mathbf{u}^{(g)} \cdot \nabla \xi) - D^\alpha\Big[(\Delta \mathbf{u}^{(g)}) \cdot \nabla(\Delta^{-1}\xi)\Big] = \varepsilon D^\alpha Q ,$$

which can be rewritten as

$$D^\alpha \xi_t + \mathbf{u}^{(g)} \cdot D^\alpha \nabla \xi = -F_\alpha + D^\alpha\Big[(\Delta \mathbf{u}^{(g)}) \cdot \nabla(\Delta^{-1}\xi)\Big] + \varepsilon D^\alpha Q ,$$

where

$$F_\alpha = D^\alpha(\mathbf{u}^{(g)} \cdot \nabla \xi) - \mathbf{u}^{(g)} \cdot D^\alpha \nabla \xi .$$

Multiplying by $D^\alpha \xi$ and integrating results in

$$\frac{1}{2}\frac{d}{dt}|\xi_\alpha|_0^2 = \Big(F_\alpha, \xi_\alpha\Big)_{L^2} + \Big(D^\alpha\Big[(\Delta \mathbf{u}^{(g)}) \cdot \nabla(\Delta^{-1}\xi)\Big], \xi_\alpha\Big)_{L^2} + \Big(\varepsilon D^\alpha Q, \xi_\alpha\Big)_{L^2} . \tag{6.24}$$

For $|\alpha| \le 2$, we conclude from the definition of F_α that

$$|F_\alpha|_0 \le C|\xi|_2 . \tag{6.25}$$

In Appendix B we show that

$$|Q|_2 \le C\Big(|\mathbf{U}^{(a)}|_4 + |\mathbf{U}^{(a)}|_4^2 + |\mathbf{U}_t^{(a)}|_3 + \varepsilon|\mathbf{U}^{(a)}|_4 |\mathbf{U}_t^{(a)}|_3\Big) ,$$

where C depends only on $|\mathbf{U}_0^{(g)}|_6$. Therefore we have

$$|Q|_2 \le C\Big(1 + |\mathbf{U}_t^{(a)}|_3\Big) , \tag{6.26}$$

where C now also depends on $|\mathbf{U}_0^{(a)}|_4$, ε_0, and T.

Taking absolute values in (6.24), using (6.25) and (6.26), and summing over α, $|\alpha| \le 2$, we have

$$\frac{1}{2}\frac{d}{dt}|\xi|_2^2 \le C|\xi|_2^2 + C\varepsilon\Big(1 + |\mathbf{U}_t^{(a)}|_3\Big)|\xi|_2 .$$

With initial condition $\xi_0 = O(\varepsilon)$, we conclude

$$|\xi|_{2,T} = O(\varepsilon) . \tag{6.27}$$

Finally, using (5.7) and (5.19) for the difference $(\mathbf{U}^{(a)} - \mathbf{U}^{(1)})$, it is easy to see that (6.27) implies

$$|\mathbf{U}^{(a)}(\varepsilon) - \mathbf{U}^{(1)}|_{3,T} = O(\varepsilon) ,$$

which proves (6.6). □

Appendix A. The L^∞ Estimate. Here we derive (3.25):

$$|\nabla \mathbf{u}|_{\mathrm{L}^\infty} + |\nabla \rho|_{\mathrm{L}^\infty} \leq C\,|\omega|_{\mathrm{L}^\infty} \left(1 + \log^+ \frac{|\omega|_{\mathrm{H}^s}}{|\omega|_{\mathrm{L}^\infty}}\right),$$

where $s \geq 2$ and $\mathbf{u}$, ρ, and ω satisfy the QGS equations (3.1)–(3.4), and (3.6). Establishing this estimate amounts to estimating all second derivatives of ϕ. The QGS pressure ϕ satisfies the boundary value problem

$$\begin{aligned} L\phi = \phi_{xx} + \phi_{yy} + (\lambda(z)\phi_z)_z = \omega \quad &\text{in } B \\ \phi_z = 0 \quad &\text{on } \Sigma_0 \quad \text{and} \quad \Sigma_h\,. \end{aligned}$$

Recall that $\lambda(z) \geq \lambda_0 > 0$. Letting $\lambda = \sigma^2$, we can make a change of variable in the z coordinate that transforms the elliptic operator L to the Laplace operator to leading order. Defining the new vertical coordinate ζ so that $d\zeta/dz = \sigma(z)^{-1}$ and $\zeta = 0$ when $z = 0$, we have

$$\sigma \frac{\partial}{\partial z} = \frac{\partial}{\partial \zeta}\,, \qquad \sigma\sigma_z \frac{\partial}{\partial z} = \tilde{\sigma}^{-1}\tilde{\sigma}_\zeta \frac{\partial}{\partial \zeta}\,,$$

where $\tilde{\sigma}(\zeta) = \sigma(z)$. Using the notation $\tilde{\phi}(x, y, \zeta) = \phi(x, y, z)$ and $\tilde{\omega}(x, y, \zeta) = \omega(x, y, z)$, $\tilde{\phi}$ is determined by the transformed boundary value problem

$$\begin{aligned} \Delta\tilde{\phi} - \tilde{\sigma}^{-1}\tilde{\sigma}_\zeta\tilde{\phi}_\zeta = \tilde{\omega} \quad &\text{in } \tilde{B}, \\ \tilde{\phi}_\zeta = 0 \quad &\text{on } \Sigma_0 \quad \text{and} \quad \Sigma_{\zeta_0}\,, \end{aligned} \tag{A.1}$$

where Δ is the Laplacian in the (x, y, ζ) system, and ζ_0 is the value of ζ corresponding to $z = h$. Then it suffices to estimate second derivatives of $\tilde{\phi}$ in the new coordinates.

Defining $f \equiv \tilde{\omega} + \tilde{\sigma}^{-1}\tilde{\sigma}_\zeta\tilde{\phi}_\zeta$, we view (A.1) as $\Delta\tilde{\phi} = f$, and seek a Green's function $G(\mathbf{x}, \mathbf{y})$ for the corresponding Neumann problem. This will actually be a modified Green's function, due to the lack of unique solvability. $G(\mathbf{x}, \mathbf{y})$ can be expressed in terms of the Green's function $G^{per}(\mathbf{x}, \mathbf{y}) = G^{per}(\mathbf{x} - \mathbf{y})$ satisfying the fully periodic problem

$$\Delta_{\mathbf{x}} G^{per}(\mathbf{x} - \mathbf{y}) = \delta^{per}(\mathbf{x} - \mathbf{y}) - (2\zeta_0)^{-1}\,, \tag{A.2}$$

where G^{per} and δ^{per} are periodic in x, y, and ζ on the periodic box

$$\tilde{B}_p = \left\{ \mathbf{x} : -\tfrac{1}{2} < x < \tfrac{1}{2},\ -\tfrac{1}{2} < y < \tfrac{1}{2},\ -\zeta_0 < z < \zeta_0 \right\},$$

and δ^{per} represents the periodic Dirac delta function. The Laplacian of any periodic function must have average value zero; thus the right-hand side of (A.2) has been adjusted by subtracting the average value of δ^{per} on B_p.

Using the method of images, we define

$$G(\mathbf{x}, \mathbf{y}) \equiv G^{per}(\mathbf{x} - \mathbf{y}) + G^{per}(\mathbf{x} - \mathbf{y}^*)\,, \tag{A.3}$$

where $\mathbf{y}^*$ is the image point of $\mathbf{y}$ reflected across the boundary Σ_0. Thus by construction $G(\mathbf{x}, \mathbf{y})$ is symmetric, with normal derivative zero on the surface Σ_0. Also, in the original box B we have

$$\Delta_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x} - \mathbf{y}) - {\zeta_0}^{-1}\,.$$

Applying Green's second identity in the usual way, we obtain the representation

$$\tilde{\phi}(\mathbf{x}) = \int_B G(\mathbf{x},\mathbf{y})\, f(\mathbf{y})\, d\mathbf{y} + \Phi\,, \tag{A.4}$$

where Φ is the average value of $\tilde{\phi}$ over B.

Before proceeding, we need to establish some facts about the smoothness of $G(\mathbf{x},\mathbf{y})$. Define the cut-off function $\xi \in C^\infty(B_p)$ such that $\xi(\mathbf{x}) = 1$ for $|\mathbf{x}| < r_0 \equiv \min\{1/4, \zeta_0/4\}$ and $\xi(\mathbf{x}) = 0$ for $|\mathbf{x}| > 2r_0$. Let $H(\mathbf{x})$ be the periodic extension of $\xi(\mathbf{x})\,(4\pi\,|\mathbf{x}|)^{-1}$. Then

$$\Delta H = g^{per}(\mathbf{x}) - \delta^{per}(\mathbf{x})\,,$$

where the function $g^{per}(\mathbf{x})$ is periodic and C^∞. Combining this with (A.2) results in

$$\Delta\Big[G^{per}(\mathbf{x}) + H\Big] = g^{per}(\mathbf{x}) - (2\zeta_0)^{-1};$$

therefore, by elliptic theory, $G^{per}(\mathbf{x}) + H$ is a smooth periodic function. For $\mathbf{x}$ and $\mathbf{y}$ in B we have the pointwise estimates $|\partial_{\mathrm{x}_j} G^{per}(\mathbf{x}-\mathbf{y})| \le C|\mathbf{x}-\mathbf{y}|^{-2}$, where ∂_{x_j} represents differentiation with respect to the jth coordinate of $\mathbf{x}$. Since $|\mathbf{x}-\mathbf{y}^*| \ge |\mathbf{x}-\mathbf{y}|$, it is also true that $|\partial_{\mathrm{x}_j} G^{per}(\mathbf{x}-\mathbf{y}^*)| \le C|\mathbf{x}-\mathbf{y}|^{-2}$. Together with the remarks above, these estimates imply

$$|D^\alpha G(\mathbf{x},\mathbf{y})| \le C|\mathbf{x}-\mathbf{y}|^{-1-|\alpha|}\,, \qquad |\alpha| \le 2\,, \qquad \mathbf{x},\mathbf{y} \in B,\ \mathbf{x} \ne \mathbf{y}\,. \tag{A.5}$$

From expression (A.4) for $\tilde{\phi}$ we write

$$\partial_{\mathrm{x}_j}\tilde{\phi}(\mathbf{x}) = \int_B [\partial_{\mathrm{x}_j} G(\mathbf{x},\mathbf{y})]\, f(\mathbf{y})\, d\mathbf{y}\,. \tag{A.6}$$

Our goal is to estimate $|\partial_{\mathrm{x}_k}\partial_{\mathrm{x}_j}\tilde{\phi}|_{\mathrm{L}^\infty}$ in terms of $|f|_{\mathrm{L}^\infty}$ and $|f|_{\mathrm{H}^s}$ using expression (A.6) with estimate (A.5). We introduce a cut-off function $\eta(\mathbf{x})$ that satisfies $\eta(\mathbf{x}) = 1$ for $|\mathbf{x}| < r$, $\eta(\mathbf{x}) = 0$ for $|\mathbf{x}| > 2r$, and $|D^\alpha \eta(\mathbf{x})| \le (C/r)^{|\alpha|}$. Here $r \le r_0 \equiv \min\{1/4, \zeta_0/4\}$ is a radius to be chosen suitably small later on. We use η to express $\partial_{\mathrm{x}_j}\tilde{\phi}$ as the sum of the terms Q_j and R_j by setting

$$\begin{aligned} Q_j(\mathbf{x}) &= \textstyle\int_B [\partial_{\mathrm{x}_j} G(\mathbf{x},\mathbf{y})]\, \eta(\mathbf{x}-\mathbf{y})\, f(\mathbf{y})\, d\mathbf{y}\,, \\ R_j(\mathbf{x}) &= \textstyle\int_B [\partial_{\mathrm{x}_j} G(\mathbf{x},\mathbf{y})]\, [1-\eta(\mathbf{x}-\mathbf{y})]\, f(\mathbf{y})\, d\mathbf{y}\,. \end{aligned} \tag{A.7}$$

We first estimate $\partial_{\mathrm{x}_k} Q_j$ and $\partial_{\mathrm{x}_k} R_j$ for the restricted case in which the first derivative is horizontal (i.e., $j = 1, 2$). This is necessary because the domain B has horizontal boundaries. We claim that Q_j can be written as

$$Q_j(\mathbf{x}) = \int_B G(\mathbf{x},\mathbf{y})\, \partial_{\mathrm{y}_j}[\eta(\mathbf{x}-\mathbf{y})\, f(\mathbf{y})]\, d\mathbf{y}, \qquad j = 1, 2\,. \tag{A.8}$$

Observe that

$$\begin{aligned} \int_B [\partial_{\mathrm{x}_j} G^{per}(\mathbf{x}-\mathbf{y})]\eta(\mathbf{x}-\mathbf{y}) f(\mathbf{y}) d\mathbf{y} &= -\int_B [\partial_{\mathrm{y}_j} G^{per}(\mathbf{x}-\mathbf{y})]\eta(\mathbf{x}-\mathbf{y}) f(\mathbf{y}) d\mathbf{y} \\ &= \int_B G^{per}(\mathbf{x}-\mathbf{y})\, \partial_{\mathrm{y}_j}[\eta(\mathbf{x}-\mathbf{y})\, f(\mathbf{y})]\, d\mathbf{y}, \qquad j = 1, 2\,. \end{aligned} \tag{A.9}$$

The last step is justified by the horizontal periodicity of all functions involved. Similarly, since $\mathbf{y}_j^* = \mathbf{y}_j$ for $j = 1, 2$, (A.9) holds if $G^{per}(\mathbf{x}-\mathbf{y})$ is replaced by $G^{per}(\mathbf{x}-\mathbf{y}^*)$.

Then (A.9) and (A.3) verify (A.8), which we now use to estimate $\partial_{\mathrm{x}_k} Q_j$, $j = 1, 2$. From (A.8) we have $\partial_{\mathrm{x}_k} Q_j = S_{j,k} + T_{j,k}$, where

$$S_{j,k}(\mathbf{x}) = \int_B [\partial_{\mathrm{x}_k} G(\mathbf{x},\mathbf{y})]\, \partial_{\mathrm{y}_j} [\eta(\mathbf{x}-\mathbf{y}) f(\mathbf{y})]\, d\mathbf{y}, \tag{A.10}$$

$$T_{j,k}(\mathbf{x}) = \int_B G(\mathbf{x},\mathbf{y})\, \partial_{\mathrm{y}_j} [f(\mathbf{y})\, \partial_{\mathrm{x}_k} \eta(\mathbf{x}-\mathbf{y})]\, d\mathbf{y}. \tag{A.11}$$

To estimate $S_{j,k}(\mathbf{x})$, we notice from (A.5) that $\partial_{\mathrm{x}_k} G(\mathbf{x},\mathbf{y})$, as a function of $\mathbf{y}$, belongs to $L^p(\{\mathbf{y} : |\mathbf{x}-\mathbf{y}| < 2r\})$ for $p < 3/2$. Using Hölder's inequality with $p = 4/3$ results in

$$\begin{aligned} |S_{j,k}(\mathbf{x})| &\le |\partial_{\mathrm{x}_j} G|_{\mathrm{L}^{4/3}}\, |\nabla f|_{\mathrm{L}^4} + |f|_{\mathrm{L}^\infty} \int_r^{2r} \rho^{-2} r^{-1} \rho^2\, d\rho \\ &\le C r^{1/4} |\nabla f|_{\mathrm{L}^4} + C|f|_{\mathrm{L}^\infty}, \qquad j = 1, 2. \end{aligned}$$

Sobolev's inequality yields $|\nabla f|_{\mathrm{L}^4} \le C|\nabla f|_{\mathrm{H}^1} \le C|f|_{\mathrm{H}^2}$; therefore, it follows that

$$|S_{j,k}(\mathbf{x})| \le C r^{1/4} |f|_{\mathrm{H}^2} + C|f|_{\mathrm{L}^\infty}, \qquad j = 1, 2.$$

For $T_{j,k}(\mathbf{x})$ we have

$$\begin{aligned} |T_{j,k}(\mathbf{x})| &\le C r^{-1} |G|_{\mathrm{L}^{4/3}}\, |\nabla f|_{\mathrm{L}^4} + |f|_{\mathrm{L}^\infty} \int_r^{2r} \rho^{-1} r^{-2} \rho^2\, d\rho \\ &\le C r^{1/4} |\nabla f|_{\mathrm{L}^4} + C|f|_{\mathrm{L}^\infty}, \qquad j = 1, 2 \end{aligned}$$

as above for $S_{j,k}$, and we conclude that

$$|\partial_{\mathrm{x}_k} Q_j(\mathbf{x})| \le C r^{1/4} |f|_{\mathrm{H}^2} + C|f|_{\mathrm{L}^\infty}, \qquad j = 1, 2. \tag{A.12}$$

We now estimate derivatives of the second term $R_j(\mathbf{x})$. From (A.7) we write

$$\partial_{\mathrm{x}_k} R_j(\mathbf{x}) = \int_B \partial_{\mathrm{x}_k} \Big\{ [1 - \eta(\mathbf{x}-\mathbf{y})]\, [\partial_{\mathrm{x}_j} G(\mathbf{x},\mathbf{y})] \Big\} f(\mathbf{y})\, d\mathbf{y}.$$

Notice that the term $\partial_{\mathrm{x}_k}\{[1 - \eta(\mathbf{x}-\mathbf{y})]\, [\partial_{\mathrm{x}_j} G(\mathbf{x},\mathbf{y})]\}$ is smooth since $[1 - \eta(\mathbf{x}-\mathbf{y})]$ is zero when $\mathbf{x}-\mathbf{y}$ is near zero. Using (A.5) to bound second derivatives of $G(\mathbf{x},\mathbf{y})$, we have

$$\begin{aligned} |\partial_{\mathrm{x}_k} R_j(\mathbf{x})| &\le C \Big\{ \int_r^{\zeta_0} \rho^{-3} \rho^2\, d\rho + \int_r^{2r} \rho^{-2} r^{-1} \rho^2\, d\rho \Big\}\, |f|_{\mathrm{L}^\infty} \\ &\le C(1 + \log(\zeta_0/r))\, |f|_{\mathrm{L}^\infty}. \end{aligned} \tag{A.13}$$

Combining estimates (A.12) and (A.13) for $\partial_{\mathrm{x}_k} Q_j$ and $\partial_{\mathrm{x}_k} R_j$, respectively, yields

$$|\partial_{\mathrm{x}_k} \partial_{\mathrm{x}_j} \tilde{\phi}(\mathbf{x})| \le C \Big\{ r^{1/4} |f|_{\mathrm{H}^2} + (1 + \log(\zeta_0/r))\, |f|_{\mathrm{L}^\infty} \Big\}, \tag{A.14}$$

which holds for $j = 1, 2$ and any k. Now from problem (A.1), the relation $\Delta\tilde{\phi} = f$ implies

$$|\partial^2_{\mathrm{x}_3} \tilde{\phi}| \le |\partial^2_{\mathrm{x}_1} \tilde{\phi}| + |\partial^2_{\mathrm{x}_2} \tilde{\phi}| + |f|.$$

Combining this with (A.14), we have established the estimate (A.14) for all second derivatives of $\tilde{\phi}$.

Now we need to eliminate f in the right-hand side of (A.14) in favor of $\tilde{\omega}$. Since $f \equiv \tilde{\omega} + \tilde{\sigma}^{-1}\tilde{\sigma}_\zeta \tilde{\phi}_\zeta$, it follows from (A.1) that

$$|f|_{H^2} \leq C|\nabla\tilde{\phi}|_{H^2} + |\tilde{\omega}|_{H^2} \leq C|\tilde{\omega}|_{H^2} . \tag{A.15}$$

We clearly have the L^∞ estimate

$$|f|_{L^\infty} \leq C|\nabla\tilde{\phi}|_{L^\infty} + C|\tilde{\omega}|_{L^\infty} . \tag{A.16}$$

From Sobolev's inequality we can write

$$|\nabla\tilde{\phi}|_{L^\infty} \leq C|\nabla\tilde{\phi}|_{W^{1,p}} \quad \text{for } p > 3 .$$

We also know from elliptic theory that

$$|\nabla\phi|_{W^{1,p}} \leq |\phi|_{W^{2,p}} \leq C_p|\omega|_{L^p} \leq C_p|\omega|_{L^\infty} ,$$

where $W^{1,p}(B)$ represents the Sobolev space of functions in $L^p(B)$ whose first derivatives are also in $L^p(B)$. These last three inequalities imply

$$|f|_{L^\infty} \leq C|\tilde{\omega}|_{L^\infty} . \tag{A.17}$$

Then plugging (A.15) and (A.17) into (A.14) results in

$$|\partial_{x_k}\partial_{x_j}\tilde{\phi}(\mathbf{x})| \leq C\Big\{ r^{1/4}|\tilde{\omega}|_{H^2} + (1 + \log(\zeta_0/r))\,|\tilde{\omega}|_{L^\infty}\Big\}. \tag{A.18}$$

Finally, choosing r in (A.18) such that $r^{1/4} = \min\{\zeta_0^{1/4}, |\tilde{\omega}|_{L^\infty}/|\tilde{\omega}|_{H^2}\}$ results in

$$|\partial_{x_k}\partial_{x_j}\tilde{\phi}(\mathbf{x})| \leq C\left(1 + \log^+ \frac{|\tilde{\omega}|_{H^s}}{|\tilde{\omega}|_{L^\infty}}\right)|\tilde{\omega}|_{L^\infty} .$$

Appendix B. Estimate for Q. We will prove here that

$$|Q|_2 \leq C\Big(|\mathbf{U}^{(a)}|_4 + |\mathbf{U}^{(a)}|_4^2 + |\mathbf{U}_t^{(a)}|_3 + \varepsilon|\mathbf{U}^{(a)}|_4\,|\mathbf{U}_t^{(a)}|_3\Big) , \tag{B.1}$$

where Q is defined by (5.21), and the constant C depends only on $|\mathbf{U}^{(g)}|_6$. In the expression for Q, we have the following five representative types of terms to estimate:

$$\text{(I)} \quad \Big[d_g\Big(d_a u^{(g)}\Big)\Big]_x , \qquad \text{(II)} \quad \left[d_g\left(\frac{Du^{(a)}}{Dt}\right)\right]_x , \qquad \text{(III)} \quad \Big[v_x^{(g)} d_a v^{(g)}\Big]_x ,$$

$$\text{(IV)} \quad \left[v_x^{(g)}\left(\frac{Dv^{(a)}}{Dt}\right)\right]_x , \qquad \text{(V)} \quad \Big[d_a v^{(a)}\Big]_x .$$

We begin with terms of type (III) since they are the easiest to estimate. Recall that we are estimating Q in $H^2(B)$. A typical subterm of $[d_g(d_a u^{(g)})]_x$ is $(v_x^{(g)} u^{(a)} v_x^{(g)})_x$. Taking its $H^2(B)$ norm, we have $|(v_x^{(g)} u^{(a)} v_x^{(g)})_x|_2 \leq C\,|\mathbf{U}^{(a)}|_3$. Notice that we have bounded the norms of the purely geostrophic terms by constants. In this way we find that

$$\left|\Big[v_x^{(g)} d_a v^{(g)}\Big]_x\right|_2 \leq C\,|\mathbf{U}^{(a)}|_3 , \tag{B.2}$$

and the same is true of the other similar type (III) terms in (5.21).

For the type (IV) term

$$\left[v_x^{(g)}\left(\frac{Dv^{(a)}}{Dt}\right)\right]_x,$$

we choose to look at the subterms

$$\begin{aligned}&\left[v_x^{(g)}\left(v_t^{(a)}+u^{(g)}v_x^{(a)}+\varepsilon u^{(a)}v_x^{(a)}\right)\right]_x\\&\quad=\left(v_x^{(g)}v_t^{(a)}\right)_x+\left(v_x^{(g)}u^{(g)}v_x^{(a)}\right)_x+\varepsilon\left(v_x^{(g)}u^{(a)}v_x^{(a)}\right)_x.\end{aligned}$$

Estimating each term in succession, we have

$$\left|\left[v_x^{(g)}\left(\frac{Dv^{(a)}}{Dt}\right)\right]_x\right|_2 \le C\left(|\mathbf{U}_t^{(a)}|_3+|\mathbf{U}^{(a)}|_4+\varepsilon|\mathbf{U}^{(a)}|_4^2\right). \tag{B.3}$$

A typical subterm of type (V) term $[d_a v^{(a)}]_x$ is $u_x^{(a)}v_x^{(a)}+u^{(a)}v_{xx}^{(a)}$, and we easily get

$$\left|\left[d_a v^{(a)}\right]_x\right|_2 \le C|\mathbf{U}^{(a)}|_4^2. \tag{B.4}$$

The relevant subterms of type (I) term $[d_g(d_a u^{(g)})]_x$ are

$$\left(u^{(a)}u_x^{(g)}\right)_{tx},\qquad u^{(g)}\left(u^{(a)}u_x^{(g)}\right)_{xx}.$$

From these terms we get

$$\left|\left[d_g\left(d_a u^{(g)}\right)\right]_x\right|_2 \le C\left(|\mathbf{U}^{(a)}|_4+|\mathbf{U}_t^{(a)}|_3\right). \tag{B.5}$$

Finally, the type (II) term

$$\left[d_g\left(\frac{Du^{(a)}}{Dt}\right)\right]_x$$

produces the two relevant terms

$$d_g\left[\left(\frac{Du^{(a)}}{Dt}\right)_x\right],\qquad u_x^{(a)}\left(\frac{Du^{(a)}}{Dt}\right)_x.$$

From $d_g[(Du^{(a)}/Dt)_x]$ we look at $d_g[(Du_x^{(a)}/Dt)+u_x^{(g)}u_x^{(a)}+\varepsilon u_x^{(a)}u_x^{(a)}]$. We need not estimate the high derivative term $d_g(Du_x^{(a)}/Dt)$ because, due to the incompressibility of the flow, it vanishes in the sum $d_g(Du_x^{(a)}/Dt)+d_g(Dv_y^{(a)}/Dt)+d_g(Dw_z^{(a)}/Dt)$ in expression (5.21). From the term $d_g(u_x^{(g)}u_x^{(a)})$, we have

$$\left|\left(u_x^{(g)}u_x^{(a)}\right)_t\right|_2+\left|u^{(g)}\left(u_x^{(g)}u_x^{(a)}\right)_x\right|_2 \le C\left(|\mathbf{U}_t^{(a)}|_3+|\mathbf{U}^{(a)}|_4\right).$$

From the term $d_g(\varepsilon u_x^{(a)} u_x^{(a)})$, we have

$$\varepsilon \left| \left(u_x^{(a)} u_x^{(a)} \right)_t \right|_2 + \varepsilon \left| u^{(g)} \left(u_x^{(a)} u_x^{(a)} \right)_x \right|_2 \leq \varepsilon C \left(|\mathbf{U}^{(a)}|_3 |\mathbf{U}_t^{(a)}|_3 + |\mathbf{U}^{(a)}|_4^2 \right).$$

For $u_x^{(g)} (\frac{Du^{(a)}}{Dt})_x$, we have

$$\left| u_x^{(g)} u_{tx}^{(a)} \right|_2 + \left| u_x^{(g)} \left(u^{(g)} u_x^{(a)} \right)_x \right|_2 + \varepsilon \left| u_x^{(g)} \left(u^{(a)} u_x^{(a)} \right)_x \right|_2 \leq C \left(|\mathbf{U}_t^{(a)}|_3 + |\mathbf{U}^{(a)}|_4 + \varepsilon |\mathbf{U}^{(a)}|_4^2 \right).$$

Summarizing for the type (II) term, we write

$$\text{(B.6)} \quad \left| d_g \left[\left(\frac{Du^{(a)}}{Dt} \right) \right]_x \right|_2 \leq C \left(|\mathbf{U}^{(a)}|_4 + |\mathbf{U}_t^{(a)}|_3 + \varepsilon |\mathbf{U}^{(a)}|_3 |\mathbf{U}_t^{(a)}|_3 + \varepsilon |\mathbf{U}^{(a)}|_4^2 \right).$$

Combining inequalities (B.2)–(B.6) we have (B.1) as claimed.

Acknowledgment. The authors wish to express their gratitude to Robert Miller and Andrew Bennett of Oregon State University for introducing them to this rich class of problems.

REFERENCES

[1] J. W. BARKER, *Interaction of fast and slow waves in problems with two time scales*, SIAM J. Math. Anal., 15 (1984), pp. 500–513.

[2] J. T. BEALE, T. KATO, AND A. MADJA, *Remarks on the breakdown of smooth solutions for the 3-D Euler equations*, Comm. Math. Phys., 94 (1984), pp. 61–66.

[3] A. F. BENNETT AND P. E. KLODEN, *The periodic quasigeostrophic equations: existence and uniqueness of strong solutions*, Proc. Roy. Soc. Edinburgh, 91A (1982), pp. 185–203.

[4] ———, *The quasigeostrophic equations: approximation predictability and equilibrium spectra of solutions*, Quart. J. Roy. Met. Soc., 107 (1981), pp. 121–136.

[5] ———, *The simplified quasigeostrophic equations: existence and uniqueness of strong solutions*, Mathematika, 27 (1980), pp. 287–311.

[6] F. P. BRETHERTON AND M. KARWEIT, *Mid-ocean mesoscale modelling*, in Numerical Models of Ocean Circulation, National Academy of Sciences, Washington, D.C., 1975.

[7] G. L. BROWNING, W. R. HOLLAND, H.-O. KREISS, AND S. J. WORLEY, *An accurate hyperbolic system for approximately hydrostatic and incompressible oceanographic flows*, Dyn. Atm. Oceans. 14 (1990), pp. 303–332.

[8] G. L. BROWNING, A. KASAHARA, AND H.-O. KREISS, *Initialization of the primitive equations by the bounded derivative method*, J. Atmos. Sci., 37 (1980), pp. 1424–1436.

[9] G. L. BROWNING AND H.-O. KREISS, *Problems with different time scales for nonlinear partial differential equations*, SIAM J. Math. Anal., 42 (1982), pp. 704–718.

[10] R. CAMASSA AND D. HOLM, *Dispersive barotropic equations for stratified mesoscale ocean dynamics*, preprint.

[11] J. G. CHARNEY, *Geostrophic turbulence*, J. Atmos. Sci., 28 (1971), pp. 1087–1095.

[12] J. A. DUTTON, *Ceaseless Wind, an Introduction to the Theory of Atmospheric Motion*, McGraw-Hill, New York, 1976, pp. 1–579.

[13] ———, *The nonlinear quasigeostrophic equations: existence and uniqueness of solutions on a bounded domain*, J. Atmos. Sci., 31 (1974), pp. 422–433.

[14] G. B. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976, pp. 1–349.

[15] T. KATO AND G. PONCE, *Well-posedness of the Euler and Navier–Stokes equations in Lebesgue spaces $L_s^p(\mathbf{R}^2)$*, Rev. Mat. Iberoamericana, 2 (1986), pp. 73–88.

[16] S. KLAINERMAN AND A. MAJDA, *Singular limits of qualilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 43 (1981), pp. 481–524.

[17] A. MAJDA, *Compressible fluid flow and systems of conservation laws in several space variables*, in Applied Math. Sciences Series, Vol. 53, Springer, Berlin, Heidelberg, New York, Tokyo, 1984.

[18] ———, *Vorticity and the mathematical theory of incompressible fluid flow*, Comm. Pure Appl. Math., 39 (1986), pp. S187–S220.

[19] J. C. McWilliams and P. R. Gent, *Intermediate models of planetary circulations in the atmosphere and ocean*, J. Atmos. Sci., 37 (1980), pp. 1657–1678.

[20] J. Pedlosky, *Geophysical Fluid Dynamics*, second ed., Springer-Verlag, New York, 1987, pp. 1–710.

[21] ———, *Geophysical fluid dynamics*, in Mathematical Problems in the Geophysical Sciences, American Mathematical Society, Providence, RI, pp. 1–60.

[22] ———, *The stability of currents in the atmosphere and the ocean, Part* I, J. Atmos. Sci., 21 (1964), pp. 201–219.

[23] S. H. Schochet, *Singular limits in bounded domains for quasilinear symmetric hyperbolic systems having a vorticity equation*, J. Differential Equations, 68 (1987), pp. 400–428.

[24] R. Temam, *On the Euler equations of incompressible perfect fluids*, J. Funct. Anal., 20 (1975), pp. 32–43.

[25] A. Valli and W. M. Zajaczkowski, *About the motion of nonhomogeneous ideal incompressible fluids*, Nonlinear Anal., 12 (1988), pp. 43–50.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1069–1084, July 1994

002

REACTION AND DIFFUSION AT A GAS/LIQUID INTERFACE, II*

HEIKKI HAARIO† AND THOMAS I. SEIDMAN‡

Abstract. The authors consider a reaction/diffusion system consisting of parabolic partial differential equations coupled through boundary conditions with ordinary differential equations. The specific model example arises in connection with the "film model" for mass transport in a chemical bubble reactor. Well-posedness is shown for the general system and convergence to steady state is shown for the specific problem.

Key words. partial differential equation, system, reaction-diffusion, film model, well-posedness, asymptotics

AMS subject classifications. 35K60, 35K57, 92E20

1. Introduction.

We will primarily be concerned with the approach to steady state for the reaction/diffusion system

$$\begin{aligned} u_t &= Du_{xx} - \lambda uv - \mu uw, \\ v_t &= Ev_{xx} - \lambda uv, \\ w_t &= Fw_{xx} + \lambda uv - \mu uw, \end{aligned} \tag{1}$$

with the boundary conditions

$$u = u^*, \qquad v_x = 0, \qquad w_x = 0 \tag{2}$$

at $x = 0$ and with boundary conditions at $x = 1$ of the form

$$\begin{aligned} -Du_x &= [u_t + \lambda uv + \mu uw]/\sigma, \\ -Ev_x &= [v_t + \lambda uv]/\sigma, \\ -Fw_x &= [w_t - \lambda uv + \mu uw]/\sigma. \end{aligned} \tag{3}$$

The particular system discussed here corresponds to a (comparatively simple) chemical process involving two irreversible reactions:

$$A + B \xrightarrow{\lambda} C, \qquad A + C \xrightarrow{\mu} D = \text{ "product"} \tag{4}$$

with $[u, v, w]$ representing concentrations for the reactants $[A, B, C]$; our eventual hypotheses will reflect the consideration that concentrations must be nonnegative. The treatment here continues the considerations of [2]; as in [2], the general form of the system (1), (3) is suggested by the "film model" [8] (or, e.g., [1]) for a gas/liquid interface.

The involvement of the time derivatives on the right-hand side of (3) is the most unusual feature of the mathematics here. If we were to introduce auxiliary scalar variables

$$U := u|_{x=1}, \qquad V := v|_{x=1}, \qquad W := w|_{x=1},$$

* Received by the editors July 20, 1992; accepted for publication (in revised form) April 12, 1993.
† Department of Mathematics, University of Helsinki, Hallituskatu 15, 00100 Helsinki, Finland (haario@convex.csc.fi).

‡ Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21228 (seidman@math.umbc.edu) or (seidman@umbc.bitnet). This research has been partially supported by National Science Foundation grant ECS-8814788 and by U.S. Air Force Office of Scientific Research grant AFOSR-91-0008.

then these boundary conditions would take the form of a system of ordinary differential equations for U, V, W involving as "sources" the *transport terms*

$$\zeta_1 := -\sigma \left[Du_x|_{x=1} \right], \quad \zeta_2 := -\sigma \left[Ev_x|_{x=1} \right], \quad \zeta_3 := -\sigma \left[Fw_x|_{x=1} \right]. \tag{5}$$

The physical interpretation is that U, V, W represent chemical concentrations in the (stirred, homogeneous) "bulk liquid" while u, v, w are concentrations within a "film" at a gas/liquid interface; σ is a units-dependent normalization of the ratio of total bubble surface to liquid volume. It is actually U, V, W (and the transport terms ζ_j) which are of primary interest to the chemical engineer.

There is no real surprise as to the eventual steady state to be reached for (4). Since the first reaction uses up the reactant B without replenishment and since "after this" there would be no replenishment for C, used up then by the second reaction, one can expect ultimate depletion of B, C while A continues to be supplied at $x = 0$ at the level u^*. Thus one can expect the steady state

$$u \equiv u^*; \quad v \equiv 0; \quad w \equiv 0. \tag{6}$$

Nevertheless, the mathematical arguments for well-posedness of the system and for the approach to steady state are not entirely trivial, and these will be our present concern. We can consider a system such as (1), (3) as an example of a more general abstract form—see (7) or (11), below—and will formulate the problem in this context, since it readily provides insight as to the significant aspects of the nonlinearities. The well-posedness holds in much greater generality, both for the chemical kinetics and the form of the transport terms. The arguments for convergence to the steady state (6) and for the exponential decay rate of that convergence are a bit more specialized and will be presented specifically in the context of (1), (3)—although it is clear that variations on the argument would apply to systems with suitably similar properties. In addition, we make some brief comments in §5 (following [2] and [3]) on the modeling and related computation.

The technical difficulties associated with well-posedness are related to the lack of a uniform Lipschitz bound for the quadratic nonlinearity; one cannot, then, a priori exclude the possibility of blowup in finite time. In particular, the positive term $+\lambda uv$ in the third equation of (1) makes it more difficult to obtain a uniform bound for w: there seems to be no obvious invariant set.[1] From (4), we expect to obtain bounds by "conservation" but these are inherently L^1 bounds and some effort is needed to see that we will actually have exponential decay to the steady state in L^∞. Furthermore, it is not a priori clear that the regularity of the transport terms supports a classical interpretation for the ordinary differential equations governing the bulk concentrations and it will be necessary to reformulate the problem so as to defer consideration of this regularity.

2. Formulation of the problem. It will be convenient to formulate the problem abstractly as an ordinary differential equation on a suitable space and apply semigroup methods. For a quasilinear equation

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{F}(t, \mathbf{z}), \qquad \mathbf{z}(0) = \overset{\circ}{\mathbf{z}}, \tag{7}$$

[1] We remark that if we had $E = F$ in (1), then there would be a separate equation for $[v + w]$ and it would be easy to bound this in terms of its initial data and so bound w as $v \geq 0$. For $E = F$, however, this trick is unavailable.

with $\mathbf{A}$ generating a C_0 semigroup one obtains an equivalent integral equation

$$\mathbf{z}(t) = \mathbf{S}(t)\,\mathring{\mathbf{z}} + \int_0^t \mathbf{S}(t-s)\mathbf{F}(s,\mathbf{z}(s))\,ds \tag{8}$$

by the variation of parameters formula for the "mild solution." A fairly standard argument gives a unique solution of (8) if, e.g., $\mathbf{F}$ is uniformly Lipschitzian with respect to $\mathbf{z}$. Our objective, in this section, is to put the problem

$$\begin{aligned} u_t &= Du_{xx} - \lambda uv - \mu uw, & U_t &= \sigma\left[-Du_x\big|_{x=1}\right] - \lambda UV - \mu UW, \\ v_t &= Ev_{xx} - \lambda uv, & V_t &= \sigma\left[-Ev_x\big|_{x=1}\right] - \lambda UV, \\ w_t &= Fw_{xx} + \lambda uv - \mu uw, & W_t &= \sigma\left[-Fw_x\big|_{x=1}\right] + \lambda UV - \mu UW, \end{aligned} \tag{9}$$

$$\begin{aligned} u = u^* = 1, \quad v_x = 0, \quad w_x = 0 \quad &\text{at } x = 0, \\ u = U, \quad v = V, \quad w = W \quad &\text{at } x = 1, \end{aligned} \tag{10}$$

into the form (7) with $\mathbf{F}$ uniformly Lipschitzian in $\mathbf{z}$ so that we can consider (8). The construction will clearly have greater generality than (9), extending to the range of problems arising from the film model; a variant, corresponding to a new transport model [3] will be sketched in §5.

If we set

$$\mathbf{u} := \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \qquad \mathbf{U} := \begin{pmatrix} U \\ V \\ W \end{pmatrix}, \qquad \mathbf{D} := \begin{pmatrix} D & 0 & 0 \\ 0 & E & 0 \\ 0 & 0 & F \end{pmatrix},$$

then the system (9) takes the form

$$\dot{\mathbf{u}} = [\mathbf{D}\mathbf{u}_x]_x + \mathbf{f}^0(\cdot,\mathbf{u}), \qquad \dot{\mathbf{U}} = \boldsymbol{\zeta} + \mathbf{f}^1(\cdot,\mathbf{U}) \quad \left(\boldsymbol{\zeta} := -\sigma\left[\mathbf{D}\mathbf{u}_x\right]\big|_{x=1}\right) \tag{11}$$

with $\mathbf{f}^0 : \mathbb{R}_+ \times [0,1] \times \mathbb{R}^3 \to \mathbb{R}^3$, $\mathbf{f}^1 : \mathbb{R}_+ \times \mathbb{R}^3 \to \mathbb{R}^3$ given by

$$\mathbf{f}^0\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} -\lambda uv - \mu uw \\ -\lambda uv \\ \lambda uv - \mu uw \end{pmatrix}, \qquad \mathbf{f}^1\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \begin{pmatrix} -\lambda UV - \mu UW \\ -\lambda UV \\ \lambda UV - \mu UW \end{pmatrix}. \tag{12}$$

To (11) we adjoin the boundary conditions (10) and, of course, initial conditions

$$\mathbf{u}\big|_{t=0} = \mathring{\mathbf{u}}, \qquad \mathbf{U}(0) = \mathring{\mathbf{U}}\,. \tag{13}$$

We require that $\mathring{\mathbf{u}}$ be bounded and that both $\mathring{\mathbf{u}}$ and $\mathring{\mathbf{U}}$ be nonnegative:

$$0 \le \mathring{u}_j(\cdot) \le M, \qquad 0 \le \mathring{U}_j \le M, \tag{14}$$

i.e., the initial data satisfy $0 \le u, v, w, U, V, W \le M$ for a suitable choice of $M \ge 1$—although we do not a priori require that the initial data be consistent with either the boundary conditions (taken at $t = 0$) or the coupling condition $\mathbf{U} = \mathbf{u}\big|_{x=1}$ or even that $\mathring{\mathbf{u}}$ be regular enough for such consistency to be meaningful. By $\mathcal{P}$ we will refer to the system (9)—or, equivalently, (11) using (12)—together with (10) and with (13) subject to (14).

We now introduce the spaces $\mathcal{X} := L^2(0,1) \times \mathbb{R}$ and $\mathcal{Z} := \mathcal{X}^3$ with the norms

$$\|[y,Y]\|_{\mathcal{X}}^2 := \sigma\|y\|^2 + |Y|^2, \qquad \|\mathbf{z}\|^2 := \sum_{j=1}^{3} \|[y_j, Y_j]\|_{\mathcal{X}}^2 \text{ for } \mathbf{z} = [\mathbf{y}, \mathbf{Y}] \in \mathcal{Z} \tag{15}$$

and define a linear operator $\mathbf{A}$ by

$$\mathbf{A} : [\mathbf{y}, \mathbf{Y}] \mapsto [(\mathbf{D}\mathbf{y}')', \boldsymbol{\zeta}] \qquad (\boldsymbol{\zeta} := -\sigma \mathbf{D}\mathbf{y}'|_{x=1}) \tag{16}$$

(here $'$ denotes d/dx) for $[\mathbf{y}, \mathbf{Y}] \in \mathcal{D}_{\mathbf{A}} \subset \mathcal{Z}$, where we set

$$\mathcal{D}_{\mathbf{A}} := \left\{ [\mathbf{y}, \mathbf{Y}] \in \mathcal{Z}: \begin{array}{l} Y_j = y_j(1) \text{ and} \\ y_j \in H^2(0,1) \text{ so } (\mathbf{D}\mathbf{y}')' \in \left[L^2(0,1)\right]^3; \\ \text{one has } \begin{cases} y_j(0) = 0 & \text{for Dirichlet conditions;} \\ y_j'(0) = 0 & \text{for Neumann conditions} \end{cases} \end{array} \right\}. \tag{17}$$

Setting

$$\mathbf{z}^* = [\mathbf{u}^*, \mathbf{U}^*] := \left[\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\right], \qquad \overset{\circ}{\mathbf{z}} := [\overset{\circ}{\mathbf{u}}, \overset{\circ}{\mathbf{U}}] - \mathbf{z}^*, \tag{18}$$

our first observation is that (10) is enforced by having $[\mathbf{u}, \mathbf{U}] - \mathbf{z}^* =: \mathbf{z} = [\mathbf{y}, \mathbf{Y}]$ be in $\mathcal{D}_{\mathbf{A}}$ and that (13) then corresponds to having $\mathbf{z}(0) = \overset{\circ}{\mathbf{z}}$ in (7).

A bit of manipulation, using the inner product of $\mathcal{Z}$ corresponding to (15), gives

$$\begin{aligned} \langle \mathbf{z}, \mathbf{A}\hat{\mathbf{z}} \rangle &= \sum_{j=1}^{3} \left[\sigma \int_0^1 y_j \left((D_j \hat{y}_j')' \right) dx + Y_j \left([\hat{\zeta}_j] \right) \right] \\ &= -\sum_{j=1}^{3} \sigma \left[\int_0^1 D_j \hat{y}_j' y_j' \, dx \right] \end{aligned}$$

for $\mathbf{z} = [\mathbf{y}, \mathbf{Y}]$, $\hat{\mathbf{z}} = [\hat{\mathbf{y}}, \hat{\mathbf{Y}}]$ in $\mathcal{D}_{\mathbf{A}}$—(17) then ensures that $y_j \hat{y}_j'$ vanishes at $x = 0$ for either Dirichlet or Neumann conditions and that the ζ_j terms at $x = 1$ cancel. Just as for the standard operator-theoretic treatment of $y \mapsto y''$ with boundary conditions, this shows that $\mathbf{A}$ is a negative self-adjoint operator on the Hilbert space $\mathcal{Z}$; it follows [5] that $\mathbf{A}$ is the infinitesimal generator of an analytic contraction semigroup $\mathbf{S}(\cdot)$ on $\mathcal{Z}$.

Thus, if we set $\mathbf{z} := [\mathbf{y}, \mathbf{Y}]$ taking values in $\mathcal{Z}$ and corresponding to $[\mathbf{u} - \mathbf{u}^*, \mathbf{U} - \mathbf{U}^*]$ and define

$$\begin{array}{rccc} \mathbf{F}: & \mathbb{R}_+ \times [0,1] \times \mathcal{Z} & \to & \mathcal{Z}, \\ & [t, \cdot, [\mathbf{y}, \mathbf{Y}]] & \mapsto & [\mathbf{f}^0(t, \cdot, \mathbf{y} + \mathbf{u}^*), \mathbf{f}^1(t, \mathbf{Y} + \mathbf{U}^*)], \end{array} \tag{19}$$

from (12), then we can write $\boldsymbol{\mathcal{P}}$ more succinctly as (7)—again subject to (14), noting (18).

Since $\mathbf{F}$, as given above, is not uniformly Lipschitzian, we will actually work with a modified version of this system. In the specific case of (12) the functions $\mathbf{f}^0$ and $\mathbf{f}^1$ are identical pointwise and we set

$$\hat{\mathbf{f}}(\mathbf{U}) = \hat{\mathbf{f}}\left(t; \begin{matrix} U \\ V \\ W \end{matrix} \right) := \mathbf{f}^1 \begin{pmatrix} \tilde{U} \\ \tilde{V} \\ \tilde{W} \end{pmatrix}, \tag{20}$$

where, with M a constant and $\tilde{M} = \tilde{M}(t) := M + \lambda M^2 t$ (M to be determined later), we set

$$\tilde{U} := \begin{cases} 0 & U \le 0, \\ M & U \ge M, \\ U & \text{else}, \end{cases} \quad \tilde{V} := \begin{cases} 0 & V \le 0, \\ M & V \ge M, \\ V & \text{else}, \end{cases} \quad \tilde{W} := \begin{cases} 0 & W \le 0, \\ \tilde{M} & W \ge \tilde{M}, \\ W & \text{else}; \end{cases} \tag{21}$$

abusing notation slightly, we use $\hat{\mathbf{f}}(\mathbf{u}) = \hat{\mathbf{f}}(t, \mathbf{u})$ as defined pointwise. We then replace $\mathbf{F}$ in (7) by $\hat{\mathbf{F}}$, given by

$$\hat{\mathbf{F}}(\mathbf{z}) = \hat{\mathbf{F}}(t, [\mathbf{y}, \mathbf{Y}]) := [\hat{\mathbf{f}}(\mathbf{y} + \mathbf{u}^*), \hat{\mathbf{f}}(\mathbf{Y} + \mathbf{U}^*)]. \tag{22}$$

(Note that $\hat{\mathbf{F}}$ is, indeed, explicitly dependent on t since $\tilde{M}$ depends on t in defining $\tilde{W}$ from W as in (20); nevertheless, we will feel free to suppress this t-dependence as above.) We denote this modified problem as $\hat{\mathcal{P}}$.

3. Well-posedness. We return now to the system (9) and will show that $\hat{\mathcal{P}}$ is well-posed, noting that, on any fixed time interval $[0, T]$, $\tilde{M}$ is bounded so the variables $\tilde{U}, \tilde{V}, \tilde{W}$ appearing in (20) are restricted to a compact set and $\hat{\mathbf{F}}$, as defined by (22), is a uniformly Lipschitzian function. We will then show by a weak maximum principle argument that the solution of $\hat{\mathcal{P}}$ necessarily satisfies

$$0 \le u \le M, \quad 0 \le v \le M, \quad 0 \le w \le \tilde{M}, \tag{23}$$

so $\tilde{U} = U$, $\tilde{V} = V$, $\tilde{W} = W$, and similarly for u, v, w so (22) coincides with (19), meaning that we will then have well-posedness[2] for the problem $\mathcal{P}$, i.e., (9), etc., as originally given.

LEMMA 3.1. *The problem $\hat{\mathcal{P}}$ is well-posed.*

Proof. We have noted that $\hat{\mathbf{F}}$ is uniformly Lipschitzian (for $t \in [0, T]$ with $T > 0$ arbitrary) and now sketch the standard argument that this gives contractivity for the Picard map $\mathbf{T}$, given by

$$[\mathbf{Tz}](t) := \mathbf{S}(t)\, \mathring{\mathbf{z}} + \int_0^t \mathbf{S}(t-s)\hat{\mathbf{F}}(s, \mathbf{z}(s))\, ds,$$

as a selfmap of $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{Z}}_T := C([0, T] \to \mathcal{Z})$ with respect to the norm

$$\|\mathbf{z}\|_{\boldsymbol{\mathcal{Z}}} := \sup\{e^{-as}\|\mathbf{z}(s)\|_{\mathcal{Z}} : 0 \le s \le T\};$$

here $a := L/\vartheta$ for some $\vartheta < 1$ with L the Lipschitz constant for $\hat{\mathbf{F}}$. For *any* functions $\mathbf{z}$, $\mathbf{z}'$ in $\boldsymbol{\mathcal{Z}}$ one has

$$\begin{aligned} \|e^{-at}[\mathbf{Tz} - \mathbf{Tz}'](t)\|_{\mathcal{Z}} &\le \int_0^t e^{-a(t-s)}\, \|\mathbf{S}(t-s)\, e^{-as}[\hat{\mathbf{F}}(\mathbf{z}(s)) - \hat{\mathbf{F}}(\mathbf{z}'(s))]\, \|\, ds \\ &\le \int_0^t e^{-a(t-s)} L \|e^{-as}[\mathbf{z}(s) - \mathbf{z}'(s)]\, \|\, ds \\ &\le L\|\mathbf{z} - \mathbf{z}'\|_{\boldsymbol{\mathcal{Z}}} \int_0^t e^{-a(t-s)}\, ds \\ &\le L\|\mathbf{z} - \mathbf{z}'\|_{\boldsymbol{\mathcal{Z}}} \int_0^\infty e^{-ar}\, dr = \vartheta \|\mathbf{z} - \mathbf{z}'\|_{\boldsymbol{\mathcal{Z}}}. \end{aligned}$$

[2] A technical delicacy intrudes at this point: while we show that the unique solution of (7) using (22) satisfies (23), one might envision the possibility of there existing another solution of (9) which did not satisfy (23) and so would not be a solution of $\hat{\mathcal{P}}$. We will not provide details here but do note that, interpreting the notion "solution of (9)" in a fairly general weak sense, one can still show uniqueness for $\mathcal{P}$ so the solution we construct here is *the* solution.

Taking the sup over t on the left, we obtain

$$\|\mathbf{Tz} - \mathbf{Tz}'\|_{\mathcal{Z}} \leq \vartheta \|\mathbf{z} - \mathbf{z}'\|_{\mathcal{Z}}$$

so $\mathbf{T}$ is contractive and there exists a unique solution on $[0,T]$ to $\hat{\mathcal{P}}$. Since $T > 0$ was arbitrary, unique existence on $[0,\infty)$ follows. To see that this solution depends Lipschitz continuously on the initial data, we estimate similarly: for solutions $\mathbf{z}$, $\mathbf{z}'$ with different initial data one has

$$\begin{aligned}
\|e^{-at}[\mathbf{z} - \mathbf{z}'](t)\| &\leq \|e^{-at}\mathbf{S}(t)[\mathbf{z} - \mathbf{z}'](0)\| \\
&\quad + \int_0^t e^{-a(t-s)} \|\mathbf{S}(t-r)e^{-as}[\hat{\mathbf{F}}(\mathbf{z}(s)) - \hat{\mathbf{F}}(\mathbf{z}'(s))]\| \, ds \\
&\leq \|[\mathbf{z} - \mathbf{z}'](0)\| + \int_0^t e^{-a(t-s)} L \|e^{-as}[\mathbf{z}(s) - \mathbf{z}'(s)]\| \, ds \\
&\leq \|\mathbf{z}(0) - \mathbf{z}'(0)\| + \vartheta \|\mathbf{z} - \mathbf{z}'\|_{\mathcal{Z}}.
\end{aligned}$$

so $\|\mathbf{z} - \mathbf{z}'\|_{\mathcal{Z}} \leq \|\mathbf{z}(0) - \mathbf{z}'(0)\|/(1-\vartheta)$, and the problem $\hat{\mathcal{P}}$ is well-posed. □

THEOREM 3.2. *Subject to* (14), *the problem* $\mathcal{P}$ *has a unique solution satisfying* (23); *this depends continuously on the initial data.*

Proof. Using Lemma 3.1, it is sufficient to show that, given (14) and the corresponding choice of M for (21), the solution $\mathbf{z} = [\mathbf{y}, \mathbf{Y}]$ of $\hat{\mathcal{P}}$ satisfies (23) with $\mathbf{u} := \mathbf{y} + \mathbf{u}^*$, etc. To this end, it is useful to note that the semigroup solution obtained above for $\hat{\mathcal{P}}$ necessarily satisfies the system (9)—modified corresponding to the passage via (21) from $\mathbf{F}$ as in (19) and (12) to $\hat{\mathbf{F}}$ as in (22) and (20)—in its weak formulation. One easily sees that this takes the form of the identities

$$\begin{aligned}
&\text{(i)} \quad \sigma \int \varphi u_t + \Phi U_t + \sigma \int D\varphi_x u_x \\
&\qquad = -\lambda[\sigma \int \varphi \tilde{u}\tilde{v} + \Phi \tilde{U}\tilde{V}] - \mu[\sigma \int \varphi \tilde{u}\tilde{w} + \Phi \tilde{U}\tilde{W}], \\
&\text{(ii)} \quad \sigma \int \varphi v_t + \Phi V_t + \sigma \int E\varphi_x u_x \\
&\qquad = -\lambda[\sigma \int \varphi \tilde{u}\tilde{v} + \Phi \tilde{U}\tilde{V}], \\
&\text{(iii)} \quad \sigma \int \varphi w_t + \Phi W_t + \sigma \int F\varphi_x v_x \\
&\qquad = \lambda[\sigma \int \varphi \tilde{u}\tilde{v} + \Phi \tilde{U}\tilde{V}] - \mu[\sigma \int \varphi \mathbf{u}\tilde{w} + \Phi \tilde{U}\tilde{W}],
\end{aligned} \tag{24}$$

each valid, e.g., for any moderately smooth φ satisfying the coupling condition $\varphi|_{x=1} = \Phi$ and also, in the case of (24(i)), satisfying the homogeneous boundary condition $\varphi|_{x=0} = 0$.

We show, first, the nonnegativity in (23). Take $\varphi = u_- = u \wedge 0 := \min\{u, 0\}$, $\Phi = U_-$ in (24(i)), noting that the boundary condition $u|_{x=0} = 1$ ensures that $\varphi|_{x=0} = 0$. By a theorem of Stampacchia [11], we have almost everywhere

$$\varphi_t = (u_-)_t = \{u_t \quad \text{if } u < 0; \quad 0 \quad \text{if } u \geq 0\},$$

etc. Thus $\varphi u_t = \left[\frac{1}{2}\varphi^2\right]_t$ and $\varphi_x u_x = [\varphi_x]^2$. Since $\varphi \le 0$ with $\varphi = 0$ when $\tilde{u} \ne 0$, etc., (24(i)) gives

$$\frac{1}{2}\left[\sigma \int \varphi^2 + \Phi^2\right]_t + \sigma \int D[\varphi_x]^2 = 0$$

and, since (14) ensures that $\varphi\big|_{t=0} = 0$ and $\Phi\big|_{t=0} = 0$, we may integrate this in t to get $\varphi^2 \equiv 0$, $\Phi^2 \equiv 0$ so $u \ge 0$, etc. Using $\varphi = v_-$, $\Phi = V_-$ in (24(ii)), the same argument gives $v \ge 0$, noting that here $\varphi_x = \{v_x \text{ if } v < 0;\ 0 \text{ if } v \ge 0\}$ gives $\varphi_x\big|_{x=0} = 0$ in either case, and, similarly, using $\varphi = w_-$, $\Phi = W_-$ in (24(iii)) gives $w \ge 0$, noting that $\tilde{u}\tilde{v} \ge 0$.

Next, we take $\varphi := (u - M)_+ := \max\{u - M, 0\}$, $\Phi = (U - M)_+$ in (24(i)), noting that the boundary condition $u\big|_{x=0} = 1$ so any choice of $M \ge 1$ ensures that φ, Φ are admissible since $u\big|_{x=0} = y_1\big|_{x=0} - 1 = Y_1 - 1 = U$. Again $\varphi u_t = \left[\frac{1}{2}\varphi^2\right]_t$, etc., and we now have $\varphi \ge 0$ and $\tilde{u}\tilde{v} = uv \ge 0$, etc., so (24(i)) gives, using (14),

$$\frac{1}{2}\left[\sigma \int \varphi^2 + \Phi^2\right]_t + \sigma \int D[\varphi_x]^2 \le 0.$$

As before, this gives $\varphi \equiv 0$, etc., so $u, U \le M$ and, similarly, we get $v, V \le M$ by using $\varphi := (v - M)_+$, $\Phi := (V - M)_+$. Finally, we will set $\varphi = (w - \hat{M})_+$, $\Phi = (W - \hat{M})_+$ in (24(iii)). Now

$$\begin{aligned} \varphi w_t &= (w - \hat{M})_+ (w_t - \lambda M^2) + (w - \hat{M})_+ \lambda M^2 \\ &= \left(\tfrac{1}{2}\varphi^2\right)_t + \lambda M^2 \varphi \quad \text{a.e.} \end{aligned}$$

Since $w_t \le F w_{xx} + \lambda M^2$, it follows that $(\frac{1}{2}\varphi^2)_t \le \varphi F w_{xx}$ when

$$\left(\frac{1}{2}\|\varphi\|^2\right)_t + F\|\varphi_x\|^2 \le F\Phi w_x\big|_{x=1}.$$

Similarly, for Φ we have

$$\left(\frac{1}{2}\Phi^2\right)_t + \lambda M^2 \Phi = \Phi W_t \le -\sigma \Phi F w_x\big|_{x=1} + \lambda M^2 \Phi.$$

We conclude that these φ and Φ also vanish identically so $w, W \le \tilde{M}$. □

4. Long-term behavior. In this section we study the asymptotic behavior of the specific system given by (9). We have already noted preceding (6) the heuristics for expecting that approach to steady state. Note, however, that the correctness of this naïve reasoning is not completely evident—certainly the "sequencing" cannot be literally true. Furthermore, we note that the estimates used in the previous sections do not even give a uniform bound for the component C—the bound $\tilde{M}$ for w, W in (23) grows unboundedly with t.

To prove convergence, we begin by defining the following functions of t:

$$\begin{aligned} \varphi_1 &= \sigma \int_0^1 v\,dx + V, \\ \varphi_2 &= \varphi_1 + \sigma \int_0^1 w\,dx + W, \\ \varphi_3 &= \sigma\|u - 1\|^2 + (U - 1)^2. \end{aligned} \tag{25}$$

The function φ_1 describes the total amount of the component B at a given time so we expect it to be decreasing and finally vanishing; this and similar statements for φ_2 and φ_3 must now be more exactly formulated.

LEMMA 4.1. *Each of the functions φ_1, φ_2 and φ_3 converges as $t \to \infty$. The functions φ_1, φ_2 are decreasing with φ_1', φ_2' integrable in t on $(0,\infty)$. Moreover, the functions $\|w\|$, W are bounded and $\|u_x\|^2$, $\|v_x\|^2$, $\|w_x\|^2$ are integrable in t on $(0,\infty)$.*

Proof. Set $v_\nu := Ev_x|_{x=1}$. Noting the boundary conditions, we have

$$\left(\int_0^1 v\,dx\right)_t = v_\nu - \lambda \int_0^1 uv\,dx,$$

and, since $V_t = -\sigma v_\nu - \lambda UV$, we have

$$\varphi_1' = -\lambda\left(\sigma \int_0^1 uv\,dx + UV\right) \le 0.$$

As a nonnegative decreasing function, φ_1 has a limit φ_1^* as $t \to \infty$. The integrability of φ_1' then follows from the observation that, for arbitrary $t > 0$,

$$-\int_0^t \varphi_1' = \varphi_1(0) - \varphi_1(t) \le \varphi_1(0) - \varphi_1^* < \infty.$$

A similar calculation with analogous conclusion is then possible for the function φ_2. It follows that $\int_0^1 uv$, UV, $\int_0^1 uw$, UW must each be an integrable function of t on $(0,\infty)$.

We next consider the function φ_3. Multiplying the equations for u and U in (9) by $u-1$ and $U-1$, respectively, and integrating, we easily see that

$$\varphi_3(t) - \varphi_3(s) + 2D\int_s^t \|u_x\|^2 \le 2\lambda \int_s^t \left(\sigma \int_0^1 uv + UV + \sigma \int_0^1 uw + UW\right).$$

Denote the integrand $(\cdots)$ on the right-hand side by G and note that we have just shown that G is integrable in t on $(0,\infty)$. Taking $s = 0$ it follows that $\int_0^t \|u_x\|^2 \le \varphi_3(0) + \int_0^t G$. Thus we see that the function $\|u_x\|^2$ is also integrable in t on $(0,\infty)$. We have the inequality

$$|\varphi_3(t) - \varphi_3(s)| \le 2D\int_s^\infty \|u_x\|^2 + 2\lambda \int_s^\infty G \qquad (t \ge s),$$

and, since the right-hand side goes to 0 as $s \to \infty$, we see that $\lim_{t\to\infty} \varphi_3(t)$ exists.

It remains to prove the boundedness of $\|w\|$ and W as functions of t. Multiplying the equations of the third line of (9) by w, W and integrating by parts, we readily obtain the estimate

$$\sigma\|w\|^2 + 2F\sigma \int_0^t \|w_x\|^2 + W^2 \le 2\lambda \int_0^t \left(\int_0^1 \sigma uvwUVW\right) + \sigma\|w_0\|^2 + W_0^2.$$

Since $H^1(0,1)$ embeds compactly in $C[0,1]$, one has, for $\varepsilon > 0$, an inequality $\|w\|_\infty \le C_\varepsilon + \varepsilon(\|w_x\|^2 + W^2)$, etc. Thus we can obtain

$$0 \le \sigma \int uvw + UVW \le \varepsilon M^2 \|w_x\|^2 + \beta\eta + C_\varepsilon,$$

where we have set

$$\eta(t) := \sigma\|w\|^2 + F\sigma\int_0^t \|w_x\|^2 + W^2,$$

$$\beta(t) := \lambda\left[\sigma\int_0^1 uv + UV\right].$$

Choosing ε so $\varepsilon M^2\lambda \leq F$, we then have $\eta(t) \leq const + \int \beta\eta$ and by Gronwall's inequality this gives

$$\eta(t) \leq (const)\exp\left[\int_0^t \beta\right].$$

This is uniformly bounded as $\beta = -\varphi_1'$ is integrable which shows that $\|w\|$, W are uniformly bounded and that $\|w_x\|^2$ is integrable in t on $(0,\infty)$. □

LEMMA 4.2. *Fixing $\varepsilon > 0$, there is a fixed compact subset of $C[0,1]$ containing the values of $u(t,\cdot)$, $v(t,\cdot)$, $w(t,\cdot)$ for all $t \geq \varepsilon$.*

Proof. As before, let $\mathbf{z} = (u-1, v, w, U-1, V, W)$ and write the state of the system at time t as

$$\mathbf{z}(t) = \mathbf{S}(t-s)\mathbf{z}(s) + \int_s^t \mathbf{S}(t-r)\mathbf{F}(\mathbf{z}(r))\,dr,$$

where $\mathbf{S}$ is the analytic semigroup given by the linear system, etc. We first show that u, v, w belong to a compact subset of $C[0,1]$. To this end, we first consider the operators[3] $\mathbf{A}^\vartheta$ $(0 \leq \vartheta < 1)$ for which we have

$$\mathbf{A}^\vartheta\mathbf{z}(t) = \mathbf{A}^\vartheta\mathbf{S}(t-s)\mathbf{z}(s) + \int_s^t \mathbf{A}^\vartheta\mathbf{S}(t-r)\mathbf{F}(\mathbf{z}(r))\,dr.$$

As $\mathbf{S}(\cdot)$ is an analytic semigroup, we have $\|\mathbf{A}^\vartheta\mathbf{S}(\tau)\| \leq M_1\tau^{-\vartheta}$ (see, e.g., [5, p. 26]) and so obtain the estimate

$$\|\mathbf{A}^\vartheta\mathbf{z}(t)\| \leq M(t-s)^{-\vartheta}\|\mathbf{z}(s)\| + \int_s^t M(t-r)^{-\vartheta}\|\mathbf{F}(\mathbf{z}(r))\|\,dr.$$

By Lemma 4.1 we know that $\mathbf{z}(\cdot)$ and so also the term $\mathbf{F}(\mathbf{z})$ are bounded in t (one has $\|-\mu uw\|_{L^2} \leq \mu M\|w\|_{L^2} \leq const$, etc.). Now let M_1 be a bound so $\|z(\cdot)\|$, $\|\mathbf{F}(\mathbf{z}(\cdot))\| \leq M_1$ and fix $\varepsilon > 0$. For any given $t \geq \varepsilon$, take $s = t - \varepsilon$ and one has

$$\|\mathbf{A}^\vartheta\mathbf{z}(t)\| \leq MM_1\left(\varepsilon^{-\vartheta} + \int_0^\varepsilon r^{-\vartheta}\,dr\right),$$

where the right-hand side is a fixed constant for $\vartheta < 1$. Thus, $\mathbf{z}(\cdot)$ is bounded in $\mathcal{D}_{\mathbf{A}^\vartheta}$.

Looking only at the components $\mathbf{y} = (u, v, w)$ and noting that these are in $L^2(0,1)$ for $\vartheta = 0$ and in $H^2(0,1)$ for $\vartheta = 1$, an interpolation argument (see, e.g., [13]) shows the uniform bound on u, v, w in $H^{2\vartheta}(0,1)$ on $\{t \geq \varepsilon\}$for $\vartheta < 1$. By the Sobolev embedding theorem, with $\vartheta > 1/4$, the functions u, v, w then take values in a compact subset of $C[0,1]$. □

[3] One typically avoids technical difficulties with fractional powers in semigroup theory by requiring that the generator be invertible. Here we have no difficulties since the generator is self-adjoint.

We are now ready to state the theorem describing the long-term behavior of our system.

THEOREM 4.3. *The solution* (u, v, w, U, V, W) *of the system* (9) *converges* (*uniformly in* x *for* u, v, w) *to the steady state* $(u_*, v_*, w_*, U_*, V_*, W_*)$ *given by* (6).

Proof. By Lemma 4.2, for any sequence $t_n \to \infty$ we can find a subsequence (t_k) for which $\mathbf{z}(t_k)$ converges to some state $\mathbf{z}_*$. We need only show that this limit state is always as given by (6)—which also shows that the convergence is independent of any extraction of a subsequence.

Consider a new initial value problem with the original initial state $\overset{\circ}{\mathbf{z}}$ replaced by the state $\mathbf{z}_*$. Denote by φ_1^* the corresponding function for (25). The function φ_1^* is again decreasing; we show that it is in fact constant. Suppose, to the contrary, that there would be some $\hat{t}$ for which $\varphi_1^*(\hat{t}) < \varphi_1^*(0)$. Since $\varphi_1(\hat{t})$ clearly depends continuously on $\mathbf{z}(t)$ which, in turn, depends continuously on $\overset{\circ}{\mathbf{z}}$ by Theorem 3.2, we may consider "initial value problems" starting with $\overset{\circ}{\mathbf{z}} = \mathbf{z}(t_k)$ and with $\overset{\circ}{\mathbf{z}} = \mathbf{z}_*$, respectively, and observe that we must then have $\varphi_1(t_k + \hat{t}) \to \varphi_1^*(\hat{t})$ since $\mathbf{z}(t_k) \to \mathbf{z}_*$. On the other hand, $\varphi_1(t_k + \hat{t}) \to \lim_{t\to\infty} \varphi_1(t) = \varphi_1^*(0)$ so $\varphi_1^*(\hat{t}) \equiv \varphi_1^*(0)$.

It follows that $(\varphi_1^*)_t = -\lambda(\sigma \int_0^1 u_*(t)v_*(t) + U_*(t)V_*(t))$ vanishes identically. Hence, noting their nonnegativity, we see that $u_*v_*(t) \equiv 0$, $U_*V_*(t) \equiv 0$ and, from the equation, we conclude that $v_*(t, \cdot)$ and $V_*(t)$ remain constant. These constants must each be 0, since otherwise one would necessarily have $u_* \equiv 0$, contradicting the boundary condition at $x = 0$. The steady state for the components v, V has thus been found and the asymptotics for w, W and u, U are subsequently found in a similar way—using φ_2 and φ_3, respectively, in place of φ_1. □

We next show that the decay to the steady state is exponential (in $C[0,1]$ for u, v, w). We know at this point that a unique solution exists with $0 \le u, v, w$, $0 \le U, V, W$, with u, v, w taking values in a given compact subset of $C[0,1]$ such that $u \to u^*$, $v \to 0$, $w \to 0$ uniformly on $[0,1]$ as $t \to \infty$. We first deal with v and V. Fixing $\epsilon > 0$ arbitrary, there is some $\tau = \tau(\epsilon)$ and $M > 0$ such that

$$|u - u^*| \le \epsilon, \qquad 0 \le v \le M, \qquad 0 \le w \le M' \quad \text{for } x \in [0,1], \quad t \ge \tau.$$

Now define $\bar{\lambda} := \lambda(u^* - \epsilon)$ so that $\lambda u \ge \bar{\lambda}$ for $t \ge \tau$ and $\bar{v}(t, \cdot) = \bar{V}(t) := Me^{-\bar{\lambda}(t-\tau)}$. We set

$$\varphi := (v - \bar{v})_+, \qquad \Phi := (V - \bar{V})_+.$$

Note that $\Phi = \varphi\big|_{x=1}$. If $\varphi_x \not\equiv 0$ one has $v_x \equiv \varphi_x$ (ae), and where $\varphi \not\equiv 0$ one has

$$\begin{aligned} \varphi_t &= v_t - \bar{v}_t = v_t + \bar{\lambda}\bar{v} = v_t - \lambda\varphi + \lambda v, \\ \Phi_t &= V_t - \lambda\Phi + \lambda V. \end{aligned}$$

Hence, using these φ, Φ in the identity (24(ii)), one has

$$\begin{aligned} &\tfrac{1}{2}[\sigma\|\varphi\|^2 + \Phi^2]_t + \bar{\lambda}[\sigma\|\varphi\|^2 + \Phi^2] + \sigma E\|\varphi_x\|^2 \\ &\quad = -\left[\sigma \int \varphi(\lambda u)v + \Phi(\lambda U)V\right] + \bar{\lambda}\left[\sigma \int \varphi v + \Phi V\right] \le \quad 0 \end{aligned}$$

which (since $\varphi, \Phi = 0$ at $t = \tau$) gives $\varphi, \Phi \equiv 0$, or

$$0 \le v(t, \cdot), V(t) \le Me^{-\lambda(t-\tau)}.$$

Next choose $\bar{\mu} \le \mu(u^* - \epsilon)$, $\bar{\mu} \le \lambda$ and set $\bar{c} := \lambda(u^* + \epsilon)/(\lambda - \bar{\mu})$, $\mu u \ge \bar{\mu}$ and $(\lambda - \bar{\mu}\bar{c}\bar{V}) \ge \lambda uv$. Now set

$$\varphi := ([w + \bar{c}\bar{V}] - [M' + \bar{c}M]e^{-\bar{\mu}(t-\tau)})_+, \qquad \Phi := \varphi_{x=1},$$

and note that $\varphi|_{t=\tau} = (w|_{t=\tau} - M')_+ = 0$. As above, $w_x = \varphi_x$ where $\varphi_x \ne 0$ and where $\varphi \ne 0$ one has (a.e.)

$$\varphi_t = w_t - \bar{\lambda}\bar{c}\bar{V} + \bar{\mu}[M' + \bar{c}M]e^{-\bar{\mu}(t-\tau)} = w_t - \bar{\lambda}\bar{c}\bar{V} - \bar{\mu}[\varphi - (w + \bar{c}\bar{V})],$$

so

$$w_t = \varphi_t + \bar{\mu}\varphi + [(\bar{\lambda} - \bar{\mu})\bar{c}\bar{V} - \bar{\mu}w] \quad \text{when } \varphi \ne 0.$$

Similarly,

$$W_t = \Phi_t + \bar{\mu}\Phi + [(\bar{\lambda} - \bar{\mu})\bar{c}\bar{V} - \bar{\mu}W] \quad \text{when } \Phi \ne 0.$$

Using this in (24(iii)) gives

$$\begin{aligned} &\tfrac{1}{2}[\sigma\|\varphi\|^2 + \Phi^2]_t + \bar{\mu}[\sigma\|\varphi\|^2 + \Phi^2] + \sigma F\|\varphi_x\|^2 \\ &\quad = -\left[\sigma\int \varphi(\bar{\lambda} - \bar{\mu})\bar{c}\bar{V} - \lambda u V\right] + \Phi[(\bar{\lambda}\bar{\mu})\bar{c}\bar{V} - \lambda UV] \\ &\qquad - \sigma\int \varphi[\mu u - \bar{\mu}]w + \Phi[\mu U - \bar{\mu}]W \le 0. \end{aligned}$$

Thus $\varphi, \Phi \equiv 0$ and we have

$$0 \le w, W \le (M' + \bar{c}M)e^{-\bar{\mu}(t-\tau)} =: \bar{W}(t).$$

A small diversion is necessary before proceeding to show exponential decay for $u - u^*$, $U - u^*$.

LEMMA 4.4. *Let y be given on $[0,1]$ with $y(0) = 0$. Then*

$$\sigma\|y\|^2 + y(1)^2 \le \beta(\sigma)\|y'\|^2,$$

with $\beta = \sigma/\rho^2$ for the smallest positive root of the equation $\rho\tan\rho = \sigma$.

Proof. Define $\mathbf{L} : z \to [y, y(1)]$ with $y(x) = \int_0^x z$ and note that the adjoint $\mathbf{L}^*$ is given by $\mathbf{L}^* : [\hat{y}, \hat{Y}] \to \hat{Y} + \sigma\int_x^1 \hat{y}$ (using the $\mathcal{X}$ inner product). Then β is the largest eigenvalue of the (positive, self-adjoint) operator $\mathbf{L}^*\mathbf{L} : L^2(0,1) \to L^2(0,1)$. Thus $\mathbf{L}^*\mathbf{L}z = \beta z$. With $z = y'$ this gives $y(1) + \sigma\int_x^1 y = \beta y'$. On the other hand, we get the boundary conditions $y(0) = 0$ and $y(1) = \beta y'(1)$. Differentiating, we also get the equation $-\sigma y = \beta y''$. Since $\beta > 0, \sigma > 0$, we have $y(x) = \sin(\sqrt{\sigma/\beta}\,x)$ and at $x = 1$ the condition $\sin\sqrt{\sigma/\beta} = \beta\sqrt{\sigma/\beta}\cos\sqrt{\sigma/\beta}$, from which the claim follows. Observe that $\rho \sim \sqrt{\beta}$ as $\sigma \to 0$, so $\beta(\sigma) \to 1$ as $\sigma \to 0$. □

Perhaps the simplest way to treat the exponential decay for $\hat{u} := u - u^*$, $\hat{U} := U - u^*$ is to use semigroup methods, noting that we have

$$\begin{aligned} &\hat{u}_t = D\hat{u}_{xx} + g, \qquad \hat{u}\big|_{x=0} = 0, \quad \hat{u}_{x=1} = \hat{U}, \\ &\hat{U}_t = \sigma[-D\hat{u}_x\big|_{x=1}] + G \end{aligned} \tag{26}$$

where $g = -\lambda uv - \mu uw$, $G = -\lambda UV - \mu UW$. From our previous estimates, we have (with a suitable M)

$$0 \le g, G \le Me^{-\bar{\lambda}(t-\tau)} \quad \text{for } t \ge \tau.$$

We may write (26) in semigroup form: setting

$$\hat{\mathbf{A}}[y, Y] := [-Dy'', \sigma Dy'|_{x=1}]$$

with the domain $\mathcal{D}_{\hat{\mathbf{A}}} := \{[y, Y] : y'' \in L^2(0,1), y(0) = 0, y(1) = Y\}$, we observe that $\hat{\mathbf{A}}$ is a self-adjoint positive definite operator on $\mathcal{X}$. Any eigenvalue $\hat{\lambda}$ for $\hat{\mathbf{A}}$ satisfies

$$\hat{\lambda}\|[y, Y]\|^2 \ge \sigma D\|y'\|^2 \ge \sigma D \frac{1}{\beta(\sigma)}\|[y, Y]\|^2,$$

i.e., $\hat{\lambda} \ge \sigma D\beta(\sigma)$. Letting $\hat{\mathbf{S}}(t)$ denote the semigroup generated by $-\hat{\mathbf{A}}$, one then has

$$\begin{aligned}\|\hat{\mathbf{A}}^{\vartheta}\hat{\mathbf{S}}(t)\| &= \sup\{\hat{\lambda}^{\vartheta}e^{-\hat{\lambda}t} : \hat{\lambda} \text{ an eigenvalue of } \hat{\mathbf{A}}\} \\ &\le \sup\{\hat{\lambda}^{\vartheta}e^{-\hat{\lambda}t} : \lambda \ge \sigma D\beta(\sigma)\}.\end{aligned}$$

By using the semigroup representation of the solution of (26) we have, with $0 \le \vartheta \le 1$,

$$\begin{aligned}\|\hat{\mathbf{A}}^{\vartheta}[\hat{u}, \hat{U}](t)\| &\le \|\hat{\mathbf{A}}^{\vartheta}\hat{\mathbf{S}}(t-\tau)\|\|[\hat{u}, \hat{U}](\tau)\| + \int_{\tau}^{t} \|\hat{\mathbf{A}}^{\vartheta}\hat{\mathbf{S}}(t-s)\|\|[g, G](s)\|ds \\ &\le K\left(r^{-\vartheta}e^{-\bar{\beta}r} + \int_0^r (r-\rho)^{-\vartheta}e^{-\bar{\beta}(r-\rho)}e^{-\bar{\mu}\rho}d\rho\right) \\ &\le Ke^{-\bar{\beta}r}\left(r^{-\vartheta} + \int_0^r (r-\rho)^{-\vartheta}e^{-\bar{\mu}-\bar{\beta})\rho}d\rho\right),\end{aligned}$$

provided $\bar{\beta} < \sigma D\beta(\sigma)$ so that one has an estimate

$$\|\hat{\mathbf{A}}^{\vartheta}\hat{\mathbf{S}}(t)\| \le K(\beta)t^{-\vartheta}e^{-\bar{\beta}t}.$$

The term above will be bounded uniformly in $r \ge r_0 > 0$ – if $\bar{\beta} < \bar{\mu}(< \bar{\lambda})$. As before, we have $\|\hat{u}\|_{\infty} \le K\|\hat{\mathbf{A}}^{\vartheta}[\hat{u}, \hat{U}]\|$ for $\vartheta > 1/4$, so we obtain

$$\|\hat{u}\|_{\infty}, |\hat{U}| \le Ke^{-\bar{\beta}t} \quad \text{for } t \ge \tau + r_0,$$

provided $\bar{\beta} < \bar{\mu}$ and also $\bar{\beta} < \sigma D\beta(\sigma)$. We have now obtained the desired asympotically exponential convergence.

THEOREM 4.5.

$$\begin{aligned} V, \|v\|_{\infty} &\to 0 \text{ at a rate } e^{-\bar{\lambda}t} && (\text{any } \bar{\lambda} < \lambda), \\ W, \|w\|_{\infty} &\to 0 \text{ at a rate } e^{-\bar{\mu}t} && (\text{any } \bar{\mu} < \min\{\lambda, \mu\}), \\ U - u^*, \|u - u^*\|_{\infty} &\to 0 \text{ at a rate } e^{-\bar{\beta}t} && (\text{any } \bar{\beta} < \min\{\lambda, \mu, \sigma D\beta(\sigma)\}).\end{aligned}$$

It is not difficult to see that the "shape" of $u - u^*$ is asymptotically proportional to the eigenfunction we found earlier: $\sin\sqrt{\sigma/\beta}\, x$. Thus, if σ is taken to be a small parameter we have $\sin\sqrt{\sigma/\beta}\, x \sim \sqrt{\sigma/\beta}\, x$ for $0 \le x \le 1$ and, uniformly for large t,

$$\begin{aligned} u(t,x) &= u^* + (U - u^*)x + o(e^{-\bar{\beta}t}), \\ u_x(t,1) &= (U - u^*) + o(e^{-\bar{\beta}t}).\end{aligned} \tag{27}$$

The usual "film theory" analysis takes u as corresponding to its quasi-steady-state form which (for $v, w \sim 0$) is just the straight line, neglecting the $o(e^{-\bar{\beta}t})$ terms in (27). This analysis takes

$$U_t = \sigma D u_x\big|_{x=1} \sim -\sigma D(U - u^*),$$

giving

$$U - u^* \sim ce^{-\sigma D t} \quad \text{as } t \to \infty.$$

Comparing this to the decay rate obtained above, we see that $\beta(\sigma) > 1$ so $\sigma D\beta(\sigma) > \sigma D$ and we actually get (slightly) faster decay than the film theory indicates (assuming $\lambda, \mu > \sigma D\beta(\sigma)$)—but with asymptotic agreement as $\sigma \to 0$, giving $\beta(\sigma) \searrow 1$.

4.1. Another example. To show that the applicability of the methods above is not confined to irreversible chemical kinetics, we sketch briefly the corresponding argument for the situation in which the second reaction in (4) is reversible:

$$A + B \longrightarrow C, \qquad A + C \rightleftharpoons D. \tag{28}$$

To make this analysis feasible, we will assume that the diffusion coefficients are equal (e.g., all 1) and that the Dirichlet boundary condition applies to B, rather than to A. Thus, we are considering

$$\begin{aligned} u_t &= u_{xx} - \lambda uv - q, & U_t &= \sigma\left[-u_x\big|_{x=1}\right] - \lambda UV - Q, \\ v_t &= v_{xx} - \lambda uv, & V_t &= \sigma\left[-v_x\big|_{x=1}\right] - \lambda UV, \\ w_t &= w_{xx} + \lambda uv - q, & W_t &= \sigma\left[-w_x\big|_{x=1}\right] + \lambda UV - Q, \\ z_t &= z_{xx} + q, & Z_t &= \sigma\left[-z_x\big|_{x=1}\right] + Q, \end{aligned} \tag{29}$$

where, for convenience, we have abbreviated

$$q := \mu uw - \nu z, \qquad Q := \mu UW - \nu Z, \tag{30}$$

and with

$$\begin{aligned} &u_x = 0, \quad v = 1, \quad w_x = 0, \quad z_x = 0 \quad \text{at } x = 0, \\ &u = U, \quad v = V, \quad w = W, \quad z = Z \quad \text{at } x = 1, \end{aligned} \tag{31}$$

and specification of bounded, nonnegative initial data.

We pass over the well-posedness argument, which parallels our discussion of §§2 and 3 in using a semigroup construction for a modified problem. The nonnegativity of all components of the solution is just as above. The bound on v, V (by M, with $M \geq \overset{\circ}{v}, \overset{\circ}{V}, 1$) is obtained by the standard argument, as earlier, but the other upper bounds are a bit different. Set $a := u + w + 2z$, $A := U + W + 2Z$ so one has

$$a_t = a_{xx}, \qquad A_t = \sigma\left[-a_x\big|_{x=1}\right] \tag{32}$$

with $a\big|_{x=1} = A$ and $a_x\big|_{x=0} = 0$. Multiplication by $(a - M)_+$ (with M a bound on $\overset{\circ}{a} = \overset{\circ}{u} + \overset{\circ}{w} + 2\overset{\circ}{z}$ and on $\overset{\circ}{A}$) then gives, as earlier, that $a, A \leq M$; since the components are nonnegative, this gives the separate bounds $u, w, 2z, U, W, 2Z \leq M$. Next, setting $\varphi := \sigma \int (w + z) + (W + Z)$ we see that $d\varphi/dt = \lambda[\sigma \int uv + UV] \geq 0$. Since we already know that φ is bounded, we must have $uv \in L^1([0,1] \times \mathbb{R}_+)$, and $UV \in L^1(\mathbb{R}_+)$.

As for (9), we get a precompact semi-orbit for $[v, V]$ and $v \to \bar{v}$ with $\bar{v}_{xx} = 0$ (as uv integrable) so $\bar{v} \equiv 1$; also $V \to 1$. Since uv is integrable and $v \to 1$, we have u integrable, so uw is integrable as w is bounded. Now setting $\zeta := \sigma \int z + Z$, we get $d\zeta/dt + \nu\zeta = \textit{integrable}$ so ζ is integrable (with $\zeta \to 0$) and so $z, Z \to 0$ also. At this point we note that there is a semigroup associated with (32), which converges to the $\mathcal{X}$-orthogonal projection onto the (one-dimensional) nullspace of the generator so

$$a \to \bar{a} \equiv \bar{A}, \qquad A \to \bar{A} := \frac{\sigma \int \overset{\circ}{a}\, dx + \overset{\circ}{A}}{\sigma + 1}. \tag{33}$$

Since we have already shown that $u, z \to 0$, we must have $w \to \bar{a} \equiv \bar{A}$ and, similarly, $W \to \bar{A}$.

Altogether, we have shown convergence for (29)–(31) to the steady state

$$u \to 0, \qquad v \to 1, \qquad w \to \bar{A}, \qquad z \to 0$$

with $\bar{A}$ as in (33). As for (9) above, we can show that this convergence is in $C[0, 1]$ and obtain exponential decay rates.

5. Discussion. In the chemical literature, film theory is used for gas/liquid reactions to calculate the mass transfer from gas to liquid at the interface. From this point of view, the "film" is a theoretical construct since only the bulk concentrations are of practical interest. In the film theory this mass transfer is given by the flux at $x = 0$, normalized by the ratio σ of total gas surface to bulk fluid volume.

We remark that the treatments in the literature and discussions with chemical engineers have not resolved the question of whether, as is implicit here, one actually has a physically "real" film of liquid moving with the bubble or whether this model may be a metaphor for a singular perturbation analysis involving a boundary layer for the pure diffusion. The former case is, indeed, plausible for a bubble reactor—one would expect from fluid dynamics considerations that there would be a boundary layer at each bubble of relatively motionless fluid within which transport would be dominated by molecular diffusion: this is the "film." Allowing for the fact that one must expect some statistical fluctuation of layer thickness, interpretable as interaction of the layer with the bulk, alternative ways to model transient gas/liquid mass transfers are the "penetration theory" and the "surface renewal theory," cf. [6], [1], as well as a variant that we sketch below but treat in more detail in [3].

The use of the film theory in chemical engineering is mainly restricted to steady-state cases. Approximate ways of computing the "enhancement factor" due to fast reactions in the interfacial layer have been considered in the chemical literature [14] in special cases with steady state and simple kinetics. Here, on the other hand, we have assumed the system to be *dynamic*, with a given initial state at $t = 0$. Such a system adequately models two types of reactors: (i) a fully "batch" reactor characterized by fixed amount of reactants in which the gas is brought into the liquid all at once, so all the bubbles begin the reaction with the same initial conditions at $t = 0$; (ii) steady-state operation of a co-current column reactor for which the role of the time variable t is played by the height coordinate of the column. A practical example is discussed in [4].

In the more common "semi-batch" reactor the bubbles are fed into the liquid continuously and we note that our model does not quite fit this. The situation differs from the full-batch one in that the bubbles have varying initial conditions: a bubble coming in at time $t = \tau$ has the initial conditions as in (13) where the right-hand

sides are to be replaced by $U(\tau)$, $V(\tau)$, $W(\tau)$. One then has $u = u(t, x, \tau)$ for $t \geq \tau$, and the flux into U involves an integral with respect to τ, etc. We intend to treat such systems elsewhere.

In using nonsteady film models to compute the enhancement factor, perhaps the most straightforward way to solve the parabolic system of equations numerically is to use the method of lines: discretize the system with respect to the spatial variable with a finite difference formula and then integrate the resulting ordinary differential equation (ODE) system with any standard ODE solver. One may encounter numerical problems of accuracy with very fast reactions, and one expects that some care in selection of the difference scheme will be needed. Useful difference formulae are discussed, e.g., in [9]. In systems like $A + B \mapsto C$ the concentrations are bounded by the initial values. In such cases the algorithm described should satisfy the *discrete maximum principle*: the computed solutions also stay between the limits given by the initial values. For more details, see [12] and [2].

In our treatment we have only considered in detail the specific system (9) but the approach here is not restricted to this. Other kinetics can be handled by the same methods. The number of components may be arbitrarily large. The essential properties of the nonlinearity $\mathbf{F}$ appearing in §§2 and 3 are that the modified $\hat{\mathbf{F}}$ be Lipschitzian and that its components satisfy such conditions as $u_- \hat{f}_1(u, v, w) \leq 0$, $v_- \hat{f}_2(u, v, w) \leq 0$, $\hat{f}_3 \leq const.$ The boundary conditions may also be modified.

A special case of interest might be to consider the situation when the first reaction $A + B \to C$ is "very fast" compared to $A + C \to D$. With an increasing reaction rate constant λ, the concentration profiles of A and B in the film become more and more $\vee$-shaped since the reactants "eat" each other almost immediately. In the limit we can expect A and B to meet at one moving point only. This naïve picture thus leads to (i) a free boundary problem for the "reduced problem"; (ii) a singular perturbation problem for the asymptotics of the relation between the problems. These questions will be studied in [7].

5.1. A variant model. A related "surface renewal" model suggests the possible consideration of an (additional) distributed transport interaction in which a term of the form $-\gamma_1[u - U]$ would be included on the right-hand side of the first reaction/diffusion equation in (1) with the corresponding transport term ζ_1 modified to be

$$\zeta_1 := \sigma \left[-Du_x\big|_{x=1} + \int_0^1 \gamma_1 u \, dx \right]$$

and a new term $-\Gamma_1 U$ (with $\Gamma_1 = \sigma \int \gamma_1 \, dx$) included in the ordinary differential equation for U—with similar treatment of the other components. Thus we let

$$\begin{aligned} \zeta_j &:= \sigma \left[-D_j [u_j]_x\big|_{x=1} + \int_0^1 \gamma_j u_j \, dx \right], \\ \mathbf{\Gamma} &:= \operatorname{diag} \left\{ \sigma \int_0^1 \gamma_1 \, dx, \ldots, \sigma \int_0^1 \gamma_m \, dx \right\} \end{aligned} \tag{34}$$

be the the coupling terms. Here we suppose that σ and each D_j is a positive constant and that each $\gamma_j = \gamma_j(\cdot)$ is in $L^\infty[0, 1]$:

$$|\gamma_j(x)| \leq M \quad \text{a.e. on } [0, 1]. \tag{35}$$

Remark. We do not pursue the possibility of a more refined treatment that could permit, e.g., $\gamma_j \in H^{-1}(0,1)$ with a one-sided bound; compare [10].

We are now considering systems of the form

$$\begin{aligned} \dot{\mathbf{u}} &= [\mathbf{D}\mathbf{u}_x]_x - \boldsymbol{\gamma}[\mathbf{u} - \mathbf{U}] + \mathbf{f}^0(\cdot, \mathbf{u}), \\ \dot{\mathbf{U}} &= \boldsymbol{\zeta} - \boldsymbol{\Gamma}\mathbf{U} + \mathbf{f}^1(\cdot, \mathbf{U}), \end{aligned} \tag{36}$$

together with the further coupling $\mathbf{U} = \mathbf{u}|_{x=1}$ and for this we introduce the linear operator $\mathbf{A}$ on $\boldsymbol{\mathcal{Z}} := \left[L^2(0,1) \times \mathbb{R}\right]^4$, given by

$$\begin{aligned} \mathbf{A} : [\mathbf{y}, \mathbf{Y}] &\mapsto [(\mathbf{D}\mathbf{y}')' - \boldsymbol{\gamma}(\mathbf{y} - \mathbf{Y}),\ \boldsymbol{\zeta} - \boldsymbol{\Gamma}\mathbf{Y}] \\ &\text{with } \boldsymbol{\zeta} := \sigma \left[[-\mathbf{D}\mathbf{y}']|_{x=1} + \int_0^1 \boldsymbol{\gamma}\mathbf{y}\, dx \right] \end{aligned} \tag{37}$$

for $[\mathbf{y}, \mathbf{Y}] \in \mathcal{D}_{\mathbf{A}}$ with $\mathcal{D}_{\mathbf{A}} \subset \boldsymbol{\mathcal{Z}}$ given essentially by (17).

Along the same lines as for (11) one can see, using the inner product of $\boldsymbol{\mathcal{Z}}$ and some manipulation, that $\mathbf{A}$ is a negative self-adjoint operator which generates an analytical contraction semigroup. The well-posedness and long-term behavior of (36) can also be established as was shown earlier.

REFERENCES

[1] E. L. CUSSLER, *Diffusion; Mass Transfer in Fluid Systems,* Cambridge University Press, Cambridge, 1984.

[2] H. HAARIO AND T. I. SEIDMAN, *Reaction and diffusion at a gas/liquid interface,* in Proceedings of the minisymposium on numerical methods for semiconductors and magnets, Bericht 42, Univ. Jyväskylä, Math. Inst., 1988.

[3] ———, *A modification of the film model,* Chem. Eng. Sci., to appear.

[4] H. HAARIO AND I. TURUNEN, *The simulation of a co-current bubble reactor,* in Proceeding of the Fourth European Conference on Mathematics in Industry, H. Wacker and W. Zulehner, eds., Kluwer, Norwell, MA, 1991.

[5] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations,* Lecture Notes in Math., 840, Springer-Verlag, New York, 1981.

[6] C-J. HUANG AND C-H. KUO, *Mathematical models for mass transfer accompanied by reversible chemical reaction,* A. I. Ch. J., 11 (1965), pp. 901–910.

[7] L. KALACHEV AND T. I. SEIDMAN, *Asymptotic analysis of a diffusion/reaction system with one fast reaction,* in preparation.

[8] W. NERNST, *Theorie der Reaktionsgeschwindigkeit in heterogenen Systemer,* Z. Phys. Chem., 47 (1904), pp. 52–55.

[9] W. E. SCHIESSER, *Numerical Method of Lines. Integration of Partial Differential Equations,* Academic Press, New York, 1991.

[10] T. I. SEIDMAN, *A convergent approximation scheme for the inverse Sturm–Liouville problem,* Inverse Problems, 1 (1985), pp. 251–262.

[11] G. STAMPACCHIA, *Equations elliptiques du second ordre á coefficients discontinues,* Les Presses de l'Université de Montréal, 1966.

[12] J. STRIKWERDA, *Finite Difference Schemes and Partial Differential Equations,* Wadsworth & Brooks/Cole, Math. Series, 1989.

[13] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators,* North Holland, Amsterdam, 1978.

[14] K. R. WESTERTERP, W. P. M. VAN SWAAIJ, AND A. A. C. M. BEENACKERS, *Chemical Reactor Design and Operation,* Wiley, New York, 1984.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1085–1111, July 1994

003

ON A NONLINEAR PARABOLIC PROBLEM ARISING IN SOME MODELS RELATED TO TURBULENT FLOWS *

JESUS ILDEFONSO DIAZ† AND FRANCOIS DE THELIN‡

Abstract. This paper studies the Cauchy–Dirichlet problem associated with the equation

$$b(u)_t - \operatorname{div}\left(|\nabla u - K(b(u))\,\mathbf{e}|^{p-2}(\nabla u - K(b(u))\,\mathbf{e})\right) + g(x,u) = f(t,x).$$

This problem arises in the study of some turbulent regimes: flows of incompressible turbulent fluids through porous media and gases flowing in pipes of uniform cross sectional areas. The paper focuses on the class of bounded weak solutions, and shows (under suitable assumptions) their stabilization, as $t \to \infty$, to the set of bounded weak solutions of the associated stationary problem. The existence and comparison properties (implying uniqueness) of such solutions are also investigated.

Key words. nonlinear parabolic equations, degenerate parabolic and elliptic equations, stabilization, existence and uniqueness of bounded weak solutions

AMS subject classifications. 35K65, 35K60, 76S05

Introduction. Physical models. The purpose of this paper is the study of the following nonlinear boundary value problem:

$$(0.1)\qquad \begin{cases} b(u)_t - \operatorname{div}\phi(\nabla u - K(b(u))\,\mathbf{e}) + g(x,u) = f(t,x) & \text{in } (0,\infty)\times\Omega, \\ u = 0 & \text{on } (0,\infty)\times\partial\Omega, \\ b(u(0,x)) = b(u_o(x)) & \text{on } \Omega, \end{cases}$$

where Ω is a bounded regular open set of $\mathbb{R}^N$, b is a nondecreasing continuous function, $K(\cdot)$ and $g(x,\cdot)$ are continuous functions satisfying some additional assumptions, and

$$(0.2)\qquad \phi(\zeta) = |\zeta|^{p-2}\zeta \quad \text{for some } p > 1 \text{ and any } \zeta \in \mathbb{R}^N$$

(in (0.1) $\mathbf{e}$ denotes a given unit vector in $\mathbb{R}^N$).

When b is strictly increasing and $p = 2$ the partial differential equation of (0.1) is of the parabolic type. Nevertheless, it becomes degenerate when $p > 2$ or $b'(0) = +\infty$ (for instance) and singular if $1 < p < 2$ or $b'(0) = 0$ (for example).

Problem (0.1), or some special cases of it, arises in many different physical contexts. Here we shall mention two of them which are related with turbulent flows, thus justifying the title of this article.

Model 1. *Flow through porous media in turbulent regime.* The infiltration of an incompressible fluid in laminar regime through a porous medium (assumed homogeneous for simplicity) is governed by the continuity equation

$$\theta_t + \operatorname{div}\mathbf{v} = 0$$

*Received by the editors August 14, 1991; accepted for publication (in revised form) March 19, 1993.

†Departamento de Matemàtica Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain. This author's research was supported in part by Dirección General de Investigación Cientifica y Technológica project PB90/0620.

‡Laboratoire d'Analyse Numerique, Universitè Paul Sabatier, 31062 Toulouse, France.

and the Darcy law

$$\mathbf{v} = -K(\theta)\,\text{grad}\,\Phi(\theta),$$

where $\theta(x,t)$ is the volumetric moisture content, $K(\theta)$ is the hydraulic conductivity and the total potential Φ is given by $\Phi(\theta) = \psi(\theta) + z$ with $\psi(\theta)$ the hydrostatic potential and z the gravitational potential (obviously we have simplified the exposition by assuming several constants equal to one: see details in Bear [7]). In turbulent regimes (which appear, for instance, in the flow through rock filled dams) the flow rate is different from that which can be predicted by the Darcy law, and so several authors proposed a nonlinear relation between $\mathbf{v}$ and $K(\theta)$ grad Φ (nonlinear Darcy's law)

$$|\mathbf{v}|^{q-2}\,\mathbf{v} = -K(\theta)\,\text{grad}\,\Phi(\theta) \quad \text{for some } q > 2 \tag{0.3}$$

(see Ahmed and Sunada [1], Hannoura and Barends [35], and Volker [54]). If $\mathbf{e}$ denotes the unit vector in the vertical direction, by introducing

$$\varphi(\theta) = \int_0^\theta K(s)\,\Phi'(s)\,ds, \qquad p = q/(q-1) \tag{0.4}$$

(notice that $1 < p < 2$), we obtain

$$\theta_t - \text{div}\,\left(|\nabla\varphi(\theta) - K(\theta)\,\mathbf{e}|^{p-2}\,(\nabla\varphi(\theta) - K(\theta)\,\mathbf{e})\right) = 0. \tag{0.5}$$

The functions φ and K are given by physical experiments (see the above references). Usually they are nondecreasing functions, being φ strictly increasing for unsaturated media. In the unsaturated case the function $u = \varphi(\theta)$ satisfies the equation of (0.1) with $b = \varphi^{-1}$ and $g = f = 0$. The case of partially saturated media leads to the same equation (for a different unknown u) but with b a strictly increasing function on $(-\infty, u^*)$ and identically constant ($\equiv b(u^*)$) on the set $[u^*,\infty)$ for some $u^* \in \mathbb{R}$ (see Bear [7]). The interest of the presence of the term $g(x,u)$ appears when the action of the roots of plants into soil is taken into account (see Gilding [33] and his references). We mention that if $p = 2$ and $N = 1$ (0.5) is also known as the nonlinear Fokker–Planck equation and has been intensively treated in the mathematical literature (see the works Kalashnikov [39], Diaz and Kersner [25], Gilding [34] and their references). Finally we point out that equation (0.1) also arises if the fluid is assumed to be compressible and (again) turbulent (see Leibenson [44] and Bear [7]).

Model 2. *Turbulent gas flowing in pipelines.* Let ρ, p, v, and T be the density pressure velocity and temperature of a perfect gas flowing in a pipe of uniform cross sectional area. In the practical cases of interest the flow is turbulent, and so ρ, p, v, and T can be assumed to depend on the scalar x (the distance along the pipe) and time t (see, e.g., Shapiro [50]). The conservation of the mass and linear momentum leads to the system

$$\rho_t + (\rho v)_x = 0, \tag{0.6}$$

$$\rho v_t + \rho v v_x + \rho_x = -\frac{\lambda}{2}\,\rho|v|v, \tag{0.7}$$

to which we add the equation of the conservation of the energy and the constitutive law $p/\rho = T$ (after suitable normalizations). In (0.6) the term $(\lambda/2)\rho|v|v$ models the

frictional forces (λ is known as the Darcy-Weissbach coefficient). Using asymptotic methods it was shown in Diaz and Liñan [26] that if the lengh L of the pipeline is considerably greater than the diameter D of the cross section, for large values of time the term $\rho v_t + \rho v v_x$ can be neglected obtaining

$$p_x = -\frac{\lambda}{2}\rho|v|v. \tag{0.8}$$

An easy computation allows to see that $u = |p|p$ satisfies the equation (0.1) with $b(u) = u^{1/2}$ sign $u, K = g = f = 0$ and the exponent p of (0.2) given by $p = 3/2$. The study of incompressible flow leads to a similar equation (0.1) but with a linear ϕ (see Liñan [45]). Finally, we mention that the interest of the presence of the term $g(x,u)$ in this context is motivated by the study of the behavior of solutions near the extinction time (see Diaz and Liñan [26, Thm. 3]).

Problems like (0.1) appear in a variety of different settings (see Bermudez, Durany, and Saguez [10], Diaz and Herrero [24], van Duijn and Hilhorst [28], Martinson and Pavlov [47] and the monographs by Diaz [22], [23]).

This paper deals with the mathematical treatement of problem (0.1) (which sometimes will be referred to as *the model problem*). Motivated by the physical models, we shall resctrict ourselves to the study of *bounded weak solutions*. This class of solutions is introduced in §1, where we also show that under suitable hypothesis those solutions stabilize as $t \to +\infty$ to the set of bounded weak solutions of the associated stationary problem. This is done for a general class of nonlinear equations including the one of (0.1). We extend the result of Langlais and Phillips [43] concerning the special case $p = 2$ and $K \equiv 0$ by passing to the limit by a variant of the already classical Minty argument (see Lions [46]). The rest of the paper is devoted to the study of the model problem (0.1).

The comparison properties (and uniqueness) of bounded weak solutions of (0.1) and its stationary problem is analysed in §2. In the case of problem (0.1) we extend the result of Alt and Lukhaus [3] valid only for $p \geq 2$ by giving a comparison criterion for $1 < p \leq 2$. The results for the stationary problem needs a different type of assumptions depending on whether $g(x,u)$ is a strictly increasing function or not.

The existence of a bounded weak solution of (0.1) is carried out in §3 by coupling regularity and sub-supersolutions arguments. The boundedness result of Boccardo and Giachetti [17] for a general class of stationary problems is shown to be applicable to our case, thus being systematically used in order to formulate our assumptions on the data f and u_0. References to some existence results for similar problems are collected in Remark 6.

We finish the article by coming back to the stabilization question and checking the assumptions of §1 for the concrete case of problem (0.1). In the first part we prove this property by using the comparison principle and the uniqueness of the associated stationary problem. That extends the result of Kröner and Rodrigues [41] concerning the case $p = 2, b$ and K Lipschitz continuous functions (b being also assumed to be bounded). Finally we treat the case of $K(b(s)) = \lambda s$ by purely energy methods generalizing several results in the literature for special cases of b, p, and $K \equiv 0$.

Some notation used through the paper follows: given $p > 1$ we associate to it the exponents $p' = p/(p-1), p^* = Np/(N-p)$ if $p < N$ and p^* arbitrary in $(p, +\infty)$ if $p \geq N$ and finally $p\# = \max(p, 2)$. The symbol $\langle , \rangle$denotes the duality product between the Sobolev space $W_0^{1,p}(\Omega)$ and its dual $(W_0^{1,p}(\Omega))^* = W^{-1,p'}(\Omega)$. Finally, we shall use the common letter C to denote different constants if no other specification is needed in the context.

1. Notion of solution and a stabilization result for a general class of equations. Let Ω be a regular open bounded set of $\mathbb{R}^N$. In this section we consider the following problem:

$$\begin{cases} b(u)_t + \mathcal{A}u + g(x,u) = f(t,x) & \text{in } (0,\infty)\times\Omega, \\ u = 0 & \text{on } (0,\infty)\times\partial\Omega, \\ b(u(0,x)) = b(u_0(x)) & \text{in } \Omega, \end{cases} \tag{1.1}$$

where $\mathcal{A}u$ denotes the operator

$$\mathcal{A}u = -\operatorname{div} \mathbf{A}(x,u,\nabla u) \tag{1.2}$$

for $\mathbf{A} : \Omega\times\mathbb{R}\times\mathbb{R}^N \to \mathbb{R}^N$, a Caratheodory function (i.e., continuous in $(\eta,\xi)\in\mathbb{R}\times\mathbb{R}^N$ and measurable in x) satisfying

$$|\mathbf{A}(x,\eta,\xi)| \le C_0\left(|\eta|^{p^*/p'} + |\xi|^{p-1}\right) + k_0(x) \tag{1.3}$$

for some $C_0 > 0$ and $k_0 \in L^{p'}(\Omega)$ and

$$\left(\mathbf{A}(x,\eta,\xi) - \mathbf{A}\left(x,\eta,\hat{\xi}\right)\right)\cdot\left(\xi - \hat{\xi}\right) > 0, \tag{1.4}$$

for any $\eta\in\mathbb{R}, \xi,\hat{\xi}\in\mathbb{R}^N, \xi\neq\hat{\xi}$ and almost every $x\in\Omega$. Obviously, the model problem (0.1) corresponds to the special case

$$\mathbf{A}(x,\eta,\xi) = \phi(\xi - K(b(\eta))\,\mathbf{e}), \qquad \phi(\xi) = |\xi|^{p-2}\xi. \tag{1.5}$$

In that case (1.4) is trivially satisfied and (1.3) holds under an additional condition (see (3.1)).

Here and throughout the rest of the paper we assume the following conditions:

$$b : \mathbb{R}\to\mathbb{R} \text{ is a continuous nondecreasing function with } b(0) = 0, \tag{1.6}$$

$$\begin{cases} g : \Omega\times\mathbb{R}\to\mathbb{R} \text{ is a Caratheodory function such that } |g(x,\eta)| \le \beta(|\eta|)(1 + d(x)) \\ \text{for some } d\in L^1(\Omega) \text{ and some continuous increasing function } \beta, \end{cases} \tag{1.7}$$

$$f \in L^1((0,T)\times\Omega) + L^{p'}\left(0,T : W^{-1,p'}(\Omega)\right), \text{ for any } T > 0, \tag{1.8}$$

$$u_0 \in L^\infty(\Omega). \tag{1.9}$$

We shall use the notion of a weak solution introduced in [3]. By a *bounded weak solution* of the problem (1.1) we mean a function $u\in L^p(0,T : W_0^{1,p}(\Omega))\cap L^\infty((0,T)\times\Omega)$, satisfying

$$\begin{cases} b(u)_t \in L^{p'}\left(0,T : W^{-1,p'}(\Omega)\right) \quad \text{and} \quad \int^T \langle b(u)_t, v\rangle + \int^T\int [b(u) - b(u_0)]\, v_t = 0, \\ \text{for any } v\in L^p(0,T : W_0^{1,p}(\Omega))\cap W^{1,1}(0,T : L^1(\Omega)), \quad \text{with} \quad v(T,\cdot) = 0, \end{cases} \tag{1.10}$$

$$\text{(1.11)}\qquad \begin{cases} \int_0^T \langle b(u)_t, v\rangle + \int_0^T \int_\Omega \mathbf{A}(\cdot, u, \nabla u)\cdot \nabla v + \int_0^T \int_\Omega g(\cdot, u)\, v = \int_0^T \int_\Omega fv, \\ \text{for any } v \in L^p(0,T : W_0^{1,p}(\Omega)) \cap L^\infty((0,T)\times\Omega), \end{cases}$$

where T is any positive number. In (1.11) we have used the notation

$$\text{(1.12)}\qquad \int_0^T \int_\Omega fv = \int_0^T \int_\Omega f_1 v + \int_0^T \langle f_2, v\rangle$$

if $f = f_1 + f_2$ with $f_1 \in L^1((0,T)\times\Omega)$ and $f_2 \in L^{p'}(0,T : W^{-1,p'}(\Omega))$.

The rest of this section is devoted to the study of the stabilization, as $t \to \infty$, of any bounded weak solution u of (1.1). As usual, we define the ω limit set associated to u by

$$\omega(u) = \Big\{ u_\infty \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega) : \exists t_n \to \infty \text{ such that } u(t_n, .) \to u_\infty \text{ in } L^p(\Omega), \text{ as } n \to \infty \Big\}.$$

In order to state our result we need some additional assumptions on f:

$$\text{(1.13)}\qquad \begin{cases} \text{there exists } f_\infty \in L^1(\Omega) + W^{-1,p'}(\Omega) \text{ such that } f(t,\cdot) \to f_\infty \text{ as} \\ t \to \infty \text{ in the sense that } \int_{t+1}^{t-1} \|f(\tau,\cdot) - f_\infty\|_{L^1+W^{-1,p}} \to 0 \quad \text{as } t \to \infty. \end{cases}$$

Finally if $f_\infty \in L^1(\Omega) + W^{-1,p'}(\Omega)$ (i.e., $f_\infty = f_{\infty,1} + f_{\infty,2}, f_{\infty,1} \in L^1(\Omega), f_{\infty,2} \in W^{-1,p'}(\Omega)$) we say that u_∞ is a *bounded weak solution* of the stationary problem

$$\text{(1.14)}\qquad \begin{cases} -\operatorname{div} A(x, u_\infty, \nabla u_\infty) + g(x, u_\infty) = f_\infty & \text{in } \Omega, \\ u_\infty = 0 & \text{on } \partial\Omega, \end{cases}$$

if $u \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ and satisfies

$$\text{(1.15)}\qquad \int_\Omega \mathbf{A}(x, u_\infty, \nabla u_\infty)\cdot \nabla v + \int_\Omega g(\cdot, u_\infty)\, v = \int_\Omega f_\infty v$$

for any $v \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ (we have used again the same abuse in the notation as in (1.12)).

THEOREM 1. *Let u be a bounded weak solution of* (1.1) *such that*

$$\text{(1.16)}\qquad u \in L^\infty(t_0, +\infty; W_0^{1,p}(\Omega)) \quad \text{for some } t_0 > 0.$$

Then $\omega(u) \neq \phi$. Moreover, if $u_\infty \in \omega(u)$ satisfies

$$\text{(1.17)}\quad \exists t_n \to +\infty \text{ such that } u(t_n + s, \cdot) \to u_\infty \text{ in } L^r(-1,1; L^p(\Omega)) \text{ for any } r \geq 1,$$

then u_∞ is a bounded weak solution of the stationary problem (1.14).

Proof. Let $t_n \to +\infty$. As $\{u(t_n,\cdot)\}$ is bounded in $W_0^{1,p}(\Omega)$ there is some subsequence (denoted again by t_n) such that $u(t_n,\cdot)$ converges in $L^p(\Omega)$ and so $\omega(u) \neq \phi$. For any $\zeta \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ and for any $\varphi \in \mathcal{D}(-1,1), \varphi \geq 0$ such that $\int_{-1}^1 \varphi(s)\, ds = 1$, we define the function $v(t,x) = \zeta(x)\varphi(t - t_n)$. For $T \geq t_n + 1$, we have

$$\int_0^T \int_\Omega [b(u) - b(u_0)]\, v_t = \int_{t_n - 1}^{t_n + 1} \int_\Omega b(u)\, \zeta \varphi'(t - t_n)$$

and from conditions (1.10) and (1.11) we get

$$\int_{t_n-1}^{t_n+1}\int_\Omega b(u)\,\zeta\varphi'(t-t_n)+\int_{t_n-1}^{t_n+1}\int_\Omega \mathbf{A}(\cdot,u,\nabla u)\cdot\nabla\zeta\varphi(t-t_n)$$
$$=\int_{t_n-1}^{t_n+1}\int_\Omega (f(t,\cdot)-g(\cdot,u))\,\zeta\varphi(t-t_n).$$

Changing variables, namely $s=t-t_n$ and defining $U_n(s,x)=u(t_n+s,x)$ we obtain

$$\int_{-1}^{1}\int_\Omega b(U_n)\,\zeta\varphi'(s)+\int_{-1}^{1}\int_\Omega \mathbf{A}(\cdot,U_n,\nabla U_n)\cdot\nabla\zeta\varphi(s)$$
$$=\int_{-1}^{1}\int_\Omega (f(s,\cdot)-g(\cdot,U_n))\,\zeta\varphi(s). \tag{1.18}$$

Since U_n is bounded in $L^\infty(-1,1;W_0^{1,p}(\Omega))$, by (1.3) and the Sobolev theorem $\mathbf{A}(\cdot,U_n,\nabla U_n)$ is bounded in $L^\infty(-1,1;(L^{p'}(\Omega))^N)$. So there is a subsequence, denoted again by U_n, weakly* convergent to u_∞ in $L^\infty(-1,1;W_0^{1,p}(\Omega))$, weakly convergent to u_∞ in $L^p(-1,1;W_0^{1,p}(\Omega))$ and such that $\mathbf{A}(\cdot,U_n,\nabla U_n)$ converges weakly* to $\mathbf{Y}$ in $L^\infty(-1,1;(L^{p'}(\Omega))^N)$. Moreover, from the assumptions on b and g the sequence $b(U_n)$ converges to $b(u_\infty)$ and $g(\cdot,U_n)$ converges to $g(\cdot,u_\infty)$ in the space $L^r((-1,1)\times\Omega)$ for any $r\in[1,+\infty)$. Moreover, we have

$$\int_\Omega \mathbf{Y}\cdot\nabla\zeta=\int_\Omega (f_\infty-g(\cdot,u_\infty))\,\zeta.$$

Due to the quasilinear character of our operator, the main difficulty is to identify $\mathbf{Y}$ as $\mathbf{A}(\cdot,u_\infty,\nabla u_\infty)$. We shall show that by means of the following inequality which is a variant of the well-known Minty argument (see [46])

$$\int_\Omega [\mathbf{Y}-\mathbf{A}(\cdot,u_\infty,\nabla\chi)]\cdot\nabla(u_\infty-\chi)\geq 0\quad\text{for any }\chi\in W_0^{1,p}(\Omega). \tag{1.19}$$

If (1.19) is verified taking $\chi=u_\infty+\lambda\xi$, with $\lambda>0$ and arbitrary $\xi\in W_0^{1,p}(\Omega)$ and letting $\lambda\to 0$ we obtain

$$\int_\Omega [\mathbf{Y}-\mathbf{A}(\cdot,u_\infty,\nabla u_\infty)]\cdot\nabla\xi\geq 0.$$

Hence $\mathcal{A}u_\infty=-\operatorname{div}\mathbf{Y}$ and the conclusion of the theorem holds. The proof of (1.19) follows from the next two lemmas. □

LEMMA 1.

$$\lim_{n\to+\infty}\int_{-1}^{1}\int_\Omega \mathbf{A}(\cdot,U_n,\nabla U_n)\cdot\nabla(U_n-u_\infty)\,\varphi(s)=0.$$

Proof of Lemma 1. Let $v(t,x)=u(t,x)\varphi(t-t_n)$. By (1.10) and (1.11) we get

$$\int_{t_n-1}^{t_n+1}\langle b(u)_t,u\rangle\,\varphi(t-t_n)+\int_{t_n-1}^{t_n+1}\int_\Omega \mathbf{A}(\cdot,u,\nabla u)\cdot\nabla u\varphi(t-t_n)$$
$$=\int_{t_n-1}^{t_n+1}\int_\Omega [f(t,\cdot)-g(\cdot,u)]\,u\varphi(t-t_n).$$

Following Alt and Luckhaus [3], we define the real function

$$B(u) = \int_0^u [b(u) - b(s)]\, ds \quad \forall u \in \mathbb{R},$$

and the time variable function

$$z_u(t) = \int_\Omega B(u(t,\cdot)).$$

As $B[u(\cdot,\cdot)]$ is bounded, then $z_u \in L^1(0,T)$ and Lemma 2 of Bamberger [5] gives

$$\int_{t_n-1}^{t_n+1} \int_\Omega \langle b(u)_t, u\rangle \varphi(t-t_n) = -\int_{t_n-1}^{t_n+1} \int_\Omega z_u(t) \varphi'(t-t_n)$$
$$= -\int_{-1}^{1} \varphi'(s) \int_\Omega \int_0^{U_n(s,x)} \{b[U_n(s,x)] - b(\sigma)\}.$$

By the dominated convergence theorem, the last term converges to

$$-\int_{-1}^{1} \varphi'(s) \int_\Omega \int_0^{u_\infty(x)} \{b[u_\infty(x)] - b(\sigma)\}$$

which is identically equal to zero. Since U_n (and, therefore, u_∞) is uniformly bounded and since $f_{\infty,1} \in L^1(\Omega)$ from the Egorov theorem, we deduce that

$$\lim_{n\to\infty} \int_{-1}^{1} \int_\Omega f_{\infty,1}(U_n - u_\infty)\varphi(s) = 0.$$

Then, by the previous results on the convergence of U_n we have

$$\lim_{n\to+\infty} \int_{-1}^{1} \int_\Omega [f(t_n+s,\cdot) - g(\cdot,U_n)] U_n \varphi(s) = \int_\Omega [f_\infty - g(\cdot,u_\infty)] u_\infty.$$

Then we get

$$\lim_{n\to+\infty} \int_{-1}^{1} \int_\Omega \mathbf{A}(\cdot,U_n,\nabla U_n)\cdot\nabla U_n \varphi(s) = \int_\Omega [f_\infty - g(\cdot,u_\infty)] u_\infty = \int_\Omega \mathbf{Y}\cdot\nabla u_\infty. \qquad \square$$

LEMMA 2.

$$\mathbf{Y} = \mathbf{A}(\cdot,u_\infty,\nabla u_\infty).$$

Proof of Lemma 2. For any $\chi \in W_0^{1,p}(\Omega)$ we have

$$\int_{-1}^{1} \int_\Omega [\mathbf{A}(\cdot,U_n,\nabla U_n) - \mathbf{A}(\cdot,u_\infty,\nabla\chi)]\cdot\nabla(u_\infty - \chi)\varphi(s) = I_1 + I_2 + I_3 + I_4,$$

where

$$I_1 = \int_{-1}^{1} \int_\Omega \mathbf{A}(\cdot,U_n,\nabla U_n)\cdot\nabla(u_\infty - U_n)\varphi(s) \to 0 \quad \text{by Lemma 1,}$$
$$I_2 = \int_{-1}^{1} \int_\Omega [\mathbf{A}(\cdot,U_n,\nabla U_n) - \mathbf{A}(\cdot,U_n,\nabla\chi)]\cdot\nabla(U_n - \chi)\varphi(s) \geq 0 \quad \text{by (1.4),}$$

and

$$I_3 = \int_{-1}^{1} \int_{\Omega} \mathbf{A}(\cdot, U_n, \nabla\chi) \cdot \nabla(u_n - u_\infty)\, \varphi(s).$$

By Lemma 2.1 of [46] $\mathbf{A}(\cdot, U_n, \nabla\chi)$ converges strongly to $\mathbf{A}(\cdot, u_\infty, \nabla\chi)$ in $(L^{p'}(\Omega))^n$, and by (1.3) there exists $C > 0$, independent of n such that

$$\operatorname*{ess\,sup}_{s\in[-1,1]} \int_{\Omega} |\mathbf{A}(\cdot, U_n, \nabla\chi) \cdot \nabla(U_n - u_\infty)| \le C < +\infty.$$

Whence, by the dominated convergence theorem, $\lim I_3 = 0$. By the same reason $\lim I_4 = 0$, where

$$I_4 = \int_{-1}^{1} \int_{\Omega} [\mathbf{A}(\cdot, U_n, \nabla\chi) - \mathbf{A}(\cdot, u_\infty, \nabla\chi)] \cdot \nabla(u_\infty - \chi)\, \varphi(s).$$

That proves the inequality (1.19) and the conclusion of the theorem follows. □

We point out that the proof of Theorem 1 also gives the information that $u(t_n, \cdot) \to u_\infty(x)$ weakly in $W_0^{1,p}(\Omega)$. The next result shows that this convergence can be improved under the following additional coercivity assumption:

(1.20)

$$\begin{cases} C\left|\xi - \hat{\xi}\right|^p \\ \le \left\{\left[\mathbf{A}(x,\eta,\xi) - \mathbf{A}\left(x,\eta,\hat{\xi}\right)\right] \cdot \left[\xi - \hat{\xi}\right]\right\}^{\alpha/2} \left\{k_1(x) + k_2\left(|\eta|^{p*} + |\xi|^p + \left|\hat{\xi}\right|^p\right)\right\}^{1-(\alpha/2)} \\ \text{for any } \eta \in \mathbb{R}, \xi, \hat{\xi} \in \mathbb{R}^N, \text{ a.e. } x \in \Omega \text{ and for some } C_1 > 0, k_1 \in L^1(\Omega), \\ k_2 > 0 \text{ and some } \alpha \in (1, 2]. \end{cases}$$

THEOREM 2. *Assume the same conditions as in Theorem* 1 *and also* (1.20). *Then, for any* $u_\infty \in \omega(u)$, *there exists a sequence* $\{\tilde{t}_n\}, \tilde{t}_n \to +\infty$ *as* $n \to +\infty$, *such that* $u(\tilde{t}_n, \cdot) \to u_\infty$ *strongly in* $W_0^{1,p}(\Omega)$.

Proof. Taking $\chi = u_\infty$ in the proof of Lemma 2 we have

$$\lim_{n\to+\infty} \int_{-1}^{1} \int_{\Omega} \mathbf{A}(\cdot, U_n, \nabla u_\infty) \cdot \nabla(U_n - u_\infty)\, \varphi(s) = 0.$$

So, by Lemma 1, we obtain $\lim_{n\to+\infty} I_n = 0$, where

$$I_n = \int_{-1}^{1} \int_{\Omega} [\mathbf{A}(\cdot, U_n, \nabla U_n) - \mathbf{A}(\cdot, u_\infty, \nabla u_\infty)] \cdot \nabla(U_n - u_\infty)\, \varphi(s).$$

Moreover, by (1.20) and Hölder's inequality

$$C \int_{-1}^{1} \int_{\Omega} |\nabla U_n - \nabla u_\infty|^p\, \varphi(s)$$
$$\le \{I_n\}^{\frac{\alpha}{2}} \left\{\int_{-1}^{1} \int_{\Omega} \left[k_1 + k_2\left(|U_n|^{p*} + |\nabla U_n|^p + |\nabla u_\infty|^p\right)\right] \varphi(s)\right\}^{1-\frac{\alpha}{2}}.$$

So, as U_n is bounded in $L^p(0,T;W_0^{1,p}(\Omega))$, for any $\varphi \in \mathcal{D}(-1,1)$ $\varphi \geq 0$ such that $\int_{-1}^{1} \varphi(s) = 1$, we have

$$\lim_{n\to+\infty} \int_{-1}^{1} \int_{\Omega} |\nabla u(t_n + s, \cdot) - \nabla u_\infty|^p \varphi(s) = 0.$$

But that is impossible if for some $\varepsilon \geq 0$,

$$\int_{\Omega} |\nabla u(t_n + s, \cdot) - \nabla u_\infty|^p \geq \varepsilon$$

for almost every $s \in (-1,1)$. Then there is a sequence $\{s_n\}, s_n \in [-1,1]$, such that

$$\lim_{n\to+\infty} \int_{\Omega} |\nabla u(t_n + s_n, \cdot) - \nabla u_\infty|^p = 0. \qquad \square$$

Remark 1. The results of this section generalize previous results in the literature: the case of $\mathcal{A} = -\Delta$ (the Laplacian operator) was treated by Langlais and Phillips [43] who showed convergence in $L^2(\Omega)$. Convergence in $L^s(\Omega)$ for some $s \geq 1$ was given in the papers Berryman and Holland [12] and Diaz and Liñan [26] for the special case of the model equation with $b(s) = s^{1/m}, K \equiv 0, (p-1)m \leq 1, p = 2$, and $p \neq 2$, respectively. The convergence in $L^1(\Omega)$ was shown in Chipot and Rodrigues [20] for $b(s) = s$ and $\mathcal{A}$ satisfying a coercivity condition stronger than (1.20). Concerning strong convergence, our result improves the one by Kröner and Rodrigues [41] (for the model equation, $p = 2$ and b Lipschitz and bounded) and the results of El Hachimi and de Thelin [29], [30] (for the model equation with $b(u) = u$ and $K \equiv 0$).

Remark 2. If $\omega(u)$ consists of a discrete number of points it is easy to see that in Theorems 1 and 2 the convergence holds for *any* subsequence t_n, i.e., when $t \to +\infty$. A more difficult task is to prove such conclusions when there is a continuum of equilibrium solutions. Some results in this direction are due to Matano [48], Alikakos and Bates [2] and Diaz and Veron [27].

The assumptions (1.16) and (1.17) hold under additional conditions on the formulation of the problem. Concerning the condition (1.17), we shall verify it (in §3) by using suitable comparison arguments. Energy type arguments also lead to this condition once we have suitable additional information on the solution. This is contained in the following result.

PROPOSITION 1. *Let $u \in L^\infty((0,\infty) \times \Omega)$. Assume that there is a continuous strictly increasing function k from $\mathbb{R}$ to $\mathbb{R}$ with $k(0) = 0$ such that $k(u) \in L^1_{\mathrm{loc}}(0,\infty : L^q(\Omega))$ for some $q \geq 1$ and*

$$\lim_{t\to+\infty} \int_{t-1}^{t+1} \int_{\Omega} |k(u)_t|^q = 0. \tag{1.21}$$

Then, if there exists a sequence $t_n \to +\infty$ satisfying

$$\lim_{n\to+\infty} u(t_n, \cdot) = u_\infty \quad \text{in } L^p(\Omega), \tag{1.22}$$

condition (1.17) *holds.*

Proof. Let $u_\infty = \lim_{n\to+\infty} u(t_n, \cdot)$ in $L^p(\Omega)$. Then there exists a subsequence (denoted again by t_n) such that $u(t_n, \cdot)$ converges almost everywhere to u_∞. For

$s \in]-1,1[$ and $x \in \Omega$, we define $U_n(s,x) = u(t_n+s,x)$. As u is bounded, by the dominated convergence theorem, $k[u(t_n,\cdot)]$ converges to $k(u_\infty)$ in $L^r(\Omega)$ for any $r \in [1,\infty)$. Moreover,

$$|k[u(t_n+s,\cdot)] - k[u(t_n,\cdot)]| = \left| \int_{t_n}^{t_n+s} k(u)_t(\tau,\cdot)\, d\tau \right|$$
$$\leq \left\{ \int_{t_n-1}^{t_n+1} |k(u)_t(\tau,\cdot)|^q \, d\tau \right\}^{\frac{1}{q}} 2^{\frac{1}{q'}}.$$

Thus using (1.21)

$$\|k(U_n) - k[u(t_n,\cdot)]\|_{L^q(\Omega)} \to 0 \quad \text{as } t_n \to \infty,$$

so, $k(U_n)$ converges to $k(u_\infty)$ in $L^q(\Omega)$. Moreover $k(U_n)$, and U_n converge in almost everywhere point to $k(u_\infty)$ and u_∞, respectively. Finally, as all these sequences are bounded the convergence holds in the space $L^r((-1,1)\times\Omega)$, for any $r \geq 1$. □

2. Comparison and continuous dependence results for the model equation. In this section, we give several results on the comparison (and then uniqueness) and continuous dependence of solutions of the model problem (0.1), i.e., (1.1) with **A** given by (1.5). We make the following additional assumptions:

$$\text{(2.1)} \qquad \begin{cases} |K[b(\eta)] - K[b(\hat\eta)]| \leq C\,|\eta - \hat\eta|^\gamma \\ \text{for any } \eta,\hat\eta \in \mathbb{R} \text{ with } \gamma \geq \frac{1}{p} \quad \text{if } 1<p<2, \gamma \geq \frac{1}{p}, \quad \text{if } p>2, \end{cases}$$

(2.2)

$$g(\cdot,\eta) - g(\cdot,\hat\eta) \geq -C^*(b(\eta) - b(\hat\eta)) \quad \text{for some } C^* \geq 0 \text{ and any } \eta,\hat\eta \in \mathbb{R}, \eta > \hat\eta.$$

Our operator is coercive in the sense that it satisfies the relation (1.20). This is a direct consequence of the well-known inequality

$$\text{(2.3)} \qquad C\left|\theta - \hat\theta\right|^p \leq \left\{ \left[|\theta|^{p-2}\theta - \left|\hat\theta\right|^{p-2}\hat\theta \right] \cdot \left[\theta - \hat\theta\right] \right\}^{\frac{\alpha}{2}} \left\{ |\theta|^p + \left|\hat\theta\right|^p \right\}^{1-\frac{\alpha}{2}},$$

which holds for any $\theta, \hat\theta \in \mathbb{R}^N$ and $p>1$ with $\alpha = p$ if $1<p\leq 2$ and $\alpha = 2$ if $p \geq 2$ (see Simon [51]). Inequality (2.3) generalizes the following one (sometimes referred as Tartar's inequality)

$$\text{(2.4)} \qquad C\left|\theta - \hat\theta\right|^p \leq \left[|\theta|^{p-2}\theta - \left|\hat\theta\right|^{p-2}\hat\theta \right] \cdot \left[\theta - \hat\theta\right]$$

which only holds for $p \geq 2$. Moreover taking $\theta = \phi(\zeta), \hat\theta = \phi(\hat\zeta)$ and changing p by p' we obtain the following inequality for any ζ and $\hat\zeta$ in $\mathbb{R}^N$

$$\text{(2.5)} \qquad \begin{cases} \left|\phi(\zeta) - \phi\left(\hat\zeta\right)\right|^{p'} \leq C\left\{ \left(\zeta - \hat\zeta\right)\left[\phi(\zeta) - \phi\left(\hat\zeta\right)\right] \right\}^{\frac{\beta}{2}} \left\{ |\zeta|^p + \left|\hat\zeta\right|^p \right\}^{1-\frac{\beta}{2}}, \\ \text{where } \beta = 2 \text{ if } 1<p\leq 2 \text{ and } \beta = p' \text{ if } p \geq 2 \text{ and } \phi(\xi) = |\xi|^{p-2}\xi. \end{cases}$$

THEOREM 3. *Assume* (2.1) *and* (2.2). *Let* (f, u_0) *and* $(\hat{f}, \hat{u}_0)$ *be a pair of data satisfying* (1.8) *and* (1.9). *Let* u *and* $\hat{u}$ *be bounded weak solutions of problem* (0.1) *corresponding to* $(f, u_0), (\hat{f}, \hat{u}_0)$, *respectively, and such that*

$$b(u)_t, b(\hat{u})_t \in L^1((0,T)\times\Omega). \tag{2.6}$$

Then

(i) *if the data are ordered* [*i.e.*, $f \le \hat{f}, u_0 \le \hat{u}_0$] *we have* $u \le \hat{u}$ *in* $(0,T)\times\Omega$,

(ii) *if* $f = f_1 + f_2, \hat{f} = \hat{f}_1 + \hat{f}_2$ *with* $f_1, \hat{f}_1 \in L^1((0,T)\times\Omega)$ *and* $f_2 = \hat{f}_2 \in L^{p'}(0,T : W^{-1,p'}(\Omega))$, *we have*

$$\|b(u(t,\cdot)) - b(\hat{u}(t,\cdot))\|_{L^1(\Omega)} \le e^{C^* t}\left(\|b(u_0) - b(\hat{u}_0)\|_{L^1(\Omega)} + \int_0^t e^{-C^* t}\left\|f_1(s,\cdot) - \hat{f}_1(s,\cdot)\right\|_{L^1(\Omega)}\right).$$

Remark 3. Conclusion (i) of Theorem 3 for $p \ge 2$ is a direct consequence of Theorem 2.2 of [3] (see also Artola [4] and Chipot and Rodrigues [20] for $b(u) = u$, and $p = 2$ and $p \ge 2$, respectively). Indeed, from (2.5) we deduce that if $p \ge 2$, then

$$\left|\phi(\zeta) - \phi\left(\hat{\zeta}\right)\right| \le C\left|\zeta - \hat{\zeta}\right|\left(|\zeta|^p + \left|\hat{\zeta}\right|^p\right)^{2-p/p}.$$

Applying this inequality to $\zeta = \xi - K(b(\eta)), \hat{\zeta} = \xi - K(b(\hat{\eta}))$ and using the assumption (2.1) we arrive to the hypothesis of the mentioned result (remark that $2 - p' \in (0,1)$). The case $1 < p < 2$ needs a new treatment because the assumptions of the mentioned papers are not satisfied.

Proof. (i) We only consider the case $1 < p < 2$. For small $\delta > 0$, we introduce the test function

$$v = \psi_\delta(u - \hat{u}), \quad \text{where } \psi_\delta(\eta) = \min\left(1, \max\left(0, \frac{\eta}{\delta}\right)\right) \quad \text{for } \eta \in \mathbb{R}.$$

We get

$$\int_0^t\int_\Omega [b(u)_t - b(\hat{u})_t]\,\psi_\delta(u-\hat{u}) + I_1(\delta) + I_2(\delta) + \int_0^t\int_\Omega [g(\cdot,u) - g(\cdot,\hat{u})]\,\psi_\delta(u-\hat{u}) = \int_0^t\int_\Omega \left(f - \hat{f}\right)\psi_\delta(u-\hat{u}), \tag{2.7}$$

where

$$I_1(\delta) = \frac{1}{\delta}\int_0^t\int_{A_\delta} \{\phi[\nabla u - K(b(u))\,\mathbf{e}] - \phi[\nabla\hat{u} - K(b(\hat{u}))\,\mathbf{e}]\} \cdot [\nabla u - K(b(u))\,\mathbf{e} - \nabla\hat{u} + K(b(\hat{u}))\,\mathbf{e}]$$

$$I_2(\delta) = \frac{1}{\delta}\int_0^t\int_{A_\delta} \{\phi[\nabla u - K(b(u))\,\mathbf{e}] - \phi[\nabla\hat{u} - K(b(\hat{u}))\,\mathbf{e}]\} \cdot \mathbf{e}\,[K(b(u)) - K(b(\hat{u}))],$$

with $A_\delta = \{(t,x) : 0 < u(t,x) - \hat{u}(t,x) < \delta\}$. By Young's inequality, we have that for any $\varepsilon > 0$

$$I_2(\delta) \leq \frac{\varepsilon}{\delta p'} \int_0^t \int_{A_\delta} |\phi[\nabla u - K(b(u))\mathbf{e}] - \phi[\nabla \hat{u} - K(b(\hat{u}))\mathbf{e}]|^{p'} + \frac{C}{\delta \varepsilon p} \int_0^t \int_{A_\delta} |K[b(u)] - K[b(\hat{u})]|^p.$$

From (2.5) and (2.1) we obtain

$$I_2(\delta) \leq \frac{\varepsilon}{p'} I_1(\delta) + \frac{C}{\varepsilon p} \delta^{p\gamma - 1}.$$

Hence, if $p\gamma > 1$, we have

$$\lim_{\delta \to 0} [I_1(\delta) + I_2(\delta)] \geq 0. \tag{2.8}$$

In the case when $p\gamma = 1$, we obtain the same result because we integrate on a set whose measure goes to 0. From (2.1), (2.7), and (2.8), letting $\delta \to 0$, we have

$$\begin{aligned} \int_\Omega \max\{b(u(t)) - b(\hat{u}(t)), 0\} &= \int_0^t \int_{\{u > \hat{u}\}} [b(u) - b(\hat{u})]_t \\ &\leq C^* \int_0^t \int_\Omega \max\{b(u) - b(\hat{u}), 0\}. \end{aligned} \tag{2.9}$$

From Gronwall's lemma we deduce that $b(u) \leq b(\hat{u})$. Using again (2.9) we also obtain that $b(u) = b(\hat{u})$ in the set A_δ. So $I_2(\delta) = 0$ and (2.7) gives $I_1(\delta) \leq 0$. From (2.3) we obtain

$$\int_0^t \int_\Omega \frac{|\nabla \psi_\delta(u - \hat{u})|^2}{|\nabla u - K(b(u))\mathbf{e}|^{2-p} - |\nabla \hat{u} - K(b(\hat{u}))\mathbf{e}|^{2-p}} \leq 0.$$

Hence $\max(0, \min(u - \hat{u}, \delta))$ is constant and this implies $u \leq \hat{u}$ since it is true on $(0,T) \times \partial\Omega$.

(ii) Suppose first that $C^* = 0$. Using (2.9) and that $f_2 = \hat{f}_2$ we have

$$\int_\Omega [b[u(t,\cdot)] - b[\hat{u}(t,\cdot)]]_+ \leq \int_\Omega [b(u_0) - b(\hat{u}_0)]_+ + \int_0^t \int_\Omega \left| f_1(s,\cdot) - \hat{f}_1(s,\cdot) \right| \operatorname{sign}_+(u - \hat{u})\, ds.$$

Adding this inequality with the similar estimate obtained for $[b(u) - b(\hat{u})]_-$ the result follows. Suppose now that $C^* > 0$. Multiplying the equations by e^{-C^*t} (in the sense that we multiply the previous test functions by e^{-C^*t}) the integrand in $I_j(\delta)$ is also multiplied by e^{-C^*t} and we can apply (2.8). Hence, we have

$$\begin{aligned} &\int_0^t \int_\Omega e^{-C^*t} [b(u)_t - b(\hat{u}_t)] \operatorname{sign}_+(u - \hat{u})\, ds \\ &\quad \leq \int_0^t \int_\Omega C^* e^{-C^*t} [b[u(s,\cdot)] - b[\hat{u}(s,\cdot)]]_+ ds \\ &\quad + \int_0^t \int_\Omega e^{-C^*t} \left(f_1 - \hat{f}_1\right) \operatorname{sign}_+(u - \hat{u})\, ds. \end{aligned}$$

Define $v = e^{-C^*t}b(u)$ and $\hat{v} = e^{-C^*t}b(\hat{u})$. As $\operatorname{sign}_+(v-\hat{v}) = \operatorname{sign}_+(u-\hat{u})$ and $v_t = -C^*v + e^{-C^*t}b(u)_t$ we obtain

$$\int_0^t \int_\Omega [v_t - \hat{v}_t] \operatorname{sign}_+(v-\hat{v})\, ds \le \int_0^t e^{-C^*t} \left[\int_\Omega \left(f_1 - \hat{f}_1 \right) \operatorname{sign}_+(v-\hat{v}) \right] ds.$$

As before we obtain in conclusion that

$$\|v(t,\cdot) - \hat{v}(t,\cdot)\|_{L^1(\Omega)} \le \|v_0 - \hat{v}_0\|_{L^1(\Omega)} + \int_0^t e^{-C^*s} \left\| f_1(s,\cdot) - \hat{f}_1(s,\cdot) \right\|_{L^1(\Omega)} ds.$$

Coming back to $b(u)$ and $b(\hat{u})$, we get the final estimate. □

We shall end this section by studying the comparison of solutions of the stationary problem

$$-\operatorname{div} \phi(\nabla u_\infty - K(b(u_\infty))\mathbf{e}) + g(x, u_\infty) = f_\infty \quad \text{in } \Omega, \tag{2.10}$$

$$u_\infty = 0 \quad \text{on } \partial\Omega. \tag{2.11}$$

As a consequence, we shall prove the uniqueness of solutions of (2.10) and (2.11): a result which will be useful for the stabilization of bounded weak solutions of the model problem.

PROPOSITION 2. *Assume* (1.6) *and* (2.1) *and suppose that one of the following assumptions holds:*

(2.12) $g(\cdot,\eta)$ *is a strictly increasing function on* η,

(2.13) $g(\cdot,\eta) = \hat{g}(\cdot, b(\eta))$ *with* $\hat{g}(\cdot,s)$ *a strictly increasing function on* s,

(2.14) $g(\cdot,\eta)$ *is a nondecreasing function on* η *and we have one of the additional conditions:*

(a) $p = 2$ *and* $N \ge 2$ *or* K *is also a monotone function,*

(b) $K(b(\eta)) = \lambda\eta$ *for some* $\lambda \in \mathbb{R}$,

(c) $1 < p \le 2$ *and there exists a constant*

$$|\phi(\xi - K(b(\eta))\mathbf{e}) - \phi(\xi - K(b(\hat{\eta}))\mathbf{e})| \le |\eta - \hat{\eta}| \left(C + |\xi|^{p-1} + |\eta|^{p-1} + |\hat{\eta}|^{p-1} \right)$$

for any $\zeta \in \mathbb{R}^N$ *and* $\eta, \hat{\eta} \in \mathbb{R}$.

Let $f_\infty, \hat{f}_\infty \in L^1(\Omega) + W^{-1,p'}(\Omega)$ *such that* $f_\infty \le \hat{f}_\infty$ *on* Ω. *Then for any* $u_\infty, \hat{u}_\infty$ *bounded weak solutions of the associated problems* (2.10) *and* (2.11) *we have* $u_\infty \le \hat{u}_\infty$ *on* Ω. *Moreover, in any case, if* $f_\infty - \hat{f}_\infty \in L^1(\Omega)$ *then*

$$\|g(\cdot,u_\infty) - g(\cdot,\hat{u}_\infty)\|_{L^1(\Omega)} \le \left\| f_\infty - \hat{f}_\infty \right\|_{L^1(\Omega)}. \tag{2.15}$$

Proof. For the sake of the notation we drop the subindex ∞ in the data and solutions. Arguing as in the proof of Theorem 3 we have

$$I_1(\delta) + I_2(\delta) + \int_\Omega [g(\cdot,u) - g(\cdot,\hat{u})]\psi_\delta(u-\hat{u}) = \int_\Omega (f - \hat{f})\psi_\delta(u-\hat{u}) \le 0,$$

where

$$I_1(\delta) = \frac{1}{\delta}\int_{A_\delta} \{\phi[\nabla u - K(b(u))\mathbf{e}] - \phi[\nabla \hat{u} - K(b(\hat{u}))\mathbf{e}]$$
$$\cdot [\nabla u - K(b(u))\mathbf{e} - \nabla \hat{u} + K(b(\hat{u}))\mathbf{e}]\}$$
$$I_2(\delta) = \frac{1}{\delta}\int_{A_\delta} \{\phi[\nabla u - K(b(u))\mathbf{e}] - \phi[\nabla \hat{u} - K(b(\hat{u}))\mathbf{e}] \cdot \mathbf{e}[K(b(u)) - K(b(\hat{u}))]\},$$

and $A_\delta = \{0 < u - \hat{u} < \delta\}$.

(i) Assume first $1 < p \leq 2$. From (2.1) we have (as in the proof of (2.8)) that

$$\lim_{\delta \to 0} [I_1(\delta) + I_2(\delta)] \geq 0 \tag{2.16}$$

and so

$$\int_\Omega (g(\cdot, u) - g(\cdot, \hat{u}))\, \mathrm{sign}_+(u - \hat{u}) \leq \int_\Omega \left(f - \hat{f}\right) \mathrm{sign}_+(u - \hat{u}), \tag{2.17}$$

where $\mathrm{sign}_+(r) = 0$ if $r \leq 0$ and $\mathrm{sign}_+(r) = 1$ if $r > 0$. If (2.12) is satisfied we have that $\mathrm{sign}_+(u - \hat{u}) = \mathrm{sign}_+(g(\cdot, u) - g(\cdot, \hat{u}))$ and the conclusion is clear. Now suppose that (2.13) is verified. From (2.17) and the fact that $\mathrm{sign}_+(b(u) - b(\hat{u})) \leq \mathrm{sign}_+(u - \hat{u})$ we have

$$\int_\Omega (\hat{g}(\cdot, b(u)) - \hat{g}(\cdot, b(\hat{u})))\, \mathrm{sign}_+(b(u) - b(\hat{u})) \leq 0.$$

As $\hat{g}$ is strictly increasing we conclude that $b(u) \leq b(\hat{u})$. In particular $b(u) = b(\hat{u})$ on the set A_δ. Then $I_1(\delta) \leq 0$ and similarly to the evolution case the inequality $I_1(\delta) \leq 0$ and (2.3) imply that $u \leq \hat{u}$.

(ii) Assume now that $p > 2$. The proof of (2.17) is the following: using (2.3) we have

$$\frac{C}{\delta}\int_{A_\delta} |\nabla(u - \hat{u})|^p + \int_\Omega (g(\cdot, u) - g(\cdot, \hat{u}))\, \psi_\delta(u - \hat{u})$$
$$\leq \frac{1}{\delta}\int_{A_\delta} \{\phi(\nabla \hat{u} - K(b(\hat{u})\mathbf{e})) - \phi(\nabla \hat{u} - K(b(u)\mathbf{e}))\} \tag{2.18}$$
$$\cdot \{\nabla(u - \hat{u})\} = I_3(\delta).$$

Using the Young inequality and the inequality of the Remark 3, we have

$$|I_3(\delta)| \leq \frac{\varepsilon}{\delta}\int_{A_\delta} |\nabla(u - \hat{u})|^p + \frac{C(\varepsilon)}{\delta}\int_{A_\delta} |K(b(\hat{u})) - K(b(u))|^{p'}$$
$$\times \left(1 + \|K(b(u))\|^p_{L^\infty(\Omega)} + \|K(b(\hat{u}))\|^p_{L^\infty(\Omega)} + |\nabla \hat{u}|^p\right) \tag{2.19}$$
$$\leq \frac{\varepsilon}{\delta}\int_{A_\delta} |\nabla(u - \hat{u})|^p + \frac{C(\varepsilon)}{\delta}\, \delta^{p'\gamma}\int_{A_\delta} H$$

for some $H \in L^1(\Omega)$. As $p'\gamma - 1 \geq 0$, letting $\delta \to 0$ we obtain (2.17).

(iii) Assume now (2.14), i.e., $g(\cdot, \eta)$ is merely a nondecreasing function on η. Again we only have to prove that $u \leq \hat{u}$ because the inequality (2.15) is then a consequence

of (2.17). For the special case $p = 2$ the uniqueness of the solution of (2.10), (2.11) was obtained in Carrillo and Chipot [18] under the assumption (2.1) (notice then that $\gamma \geq 1/2$) and $N \geq 2$ (see Theorem 6 of [18], [32], and [19]) or $N = 1$ and $K(b(\eta))$ a merely *continuous* monotone function (see Theorem 3, (ii) of [18]). Some easy modifications of the proofs lead to the comparison $u \leq \hat{u}$. In case (b), without loss of generality we can assume that $\mathbf{e} = \mathbf{e}_1$ where $\mathbf{e}_1$ is the first term of a orthonormal base of $\mathbb{R}^N$. Then we have

$$\phi(\nabla u - \lambda u \mathbf{e}_1) = e^{\lambda(p-1)x_1} \phi\left(\nabla\left(u e^{-\lambda x_1}\right)\right).$$

Using $e^{-\lambda x_1}(u - \hat{u})_+$ as test function we obtain

$$\int_{\{u>\hat{u}\}} \left[\phi\left(\nabla\left(u e^{-\lambda x_1}\right)\right) - \phi\left(\nabla\left(\hat{u} e^{-\lambda x_1}\right)\right)\right] \cdot \left[\nabla\left(u e^{-\lambda x_1}\right) - \nabla\left(\hat{u} e^{-\lambda x_1}\right)\right] e^{\lambda(p-1)x_1} \leq 0.$$

Applying (2.3) we conclude that $u \leq \hat{u}$. Assume finally the conditions of case (c). Let $\psi_\delta(u - \hat{u})$ the same test function of the proof of Theorem 3. Then, if we denote by C^* the constant in (2.3) we have

$$\begin{aligned}
&\frac{C^*}{\delta} \int_{A_\delta} \frac{|\nabla(u - \hat{u})|^2}{|\nabla u|^{2-p} + |\nabla \hat{u}|^{2-p}} \\
&\quad \leq \int_\Omega \left[\phi(\nabla u - K(b(u))\,\mathbf{e}) - \phi(\nabla \hat{u} - K(b(u))\,\mathbf{e})\right] \cdot \left[\nabla \psi_\delta(u - \hat{u})\right] \\
&\quad \leq \frac{1}{\delta} \int_{A_\delta} \left[\phi(\nabla \hat{u} - K(b(\hat{u}))\,\mathbf{e}) - \phi(\nabla \hat{u} - K(b(u))\,\mathbf{e})\right] \cdot \nabla(u - \hat{u}) \\
&\quad \leq \frac{1}{\delta} \int_{A_\delta} |u - \hat{u}| \left(C + |\nabla u|^{p-1} + |u|^{p-1} + |\hat{u}|^{p-1}\right) |\nabla(u - \hat{u})|.
\end{aligned}$$

As in Boccardo, Gallouët, and Murat [16] we notice that for any $\tau > 0$

$$\begin{aligned}
|u - \hat{u}|\,|\nabla u|^{p-1}\,|\nabla(u - \hat{u})| &\leq \frac{C^*}{\tau}\left(|\nabla u|^{p-2} + |\nabla \hat{u}|^{p-2}\right)|\nabla(u - \hat{u})|^2 \\
&\quad + C(\tau)\,|u - \hat{u}|^2\,|\nabla u|^p \\
C\,|u - \hat{u}|\,|\nabla(u - \hat{u})| &\leq \frac{C^*}{\tau}\left(|\nabla u|^{p-2} + |\nabla \hat{u}|^{p-2}\right)|\nabla(u - \hat{u})|^2 \\
&\quad + C(\tau)\,|u - \hat{u}|^2\,|\nabla u|^{2-p} \\
|u - \hat{u}|\left(|u|^{p-1} + |\hat{u}|^{p-1}\right)|\nabla(u - \hat{u})| &\leq \frac{C^*}{\tau}\left(|\nabla u|^{p-2} + |\nabla \hat{u}|^{p-2}\right)|\nabla(u - \hat{u})|^2 \\
&\quad + C(\tau)\,|u - \hat{u}|^2\,|\nabla u|^{2-p}.
\end{aligned}$$

In consequence, taking τ large enough, we have that

$$\int_{A_\delta} \frac{|\nabla(u - \hat{u})|^2}{|\nabla u|^{2-p} + |\nabla \hat{u}|^{2-p}} \leq C \int_{A_\delta} |u - \hat{u}|^2 H_1 \leq C\delta^2 \int_{A_\delta} H_1$$

with

$$H_1 = |\nabla u|^p + |\nabla u|^{2-p} + |\nabla \hat{u}|^{2-p}.$$

Introducing the function

$$H_2 = |\nabla u|^{2-p} + |\nabla \hat{u}|^{2-p}$$

is not difficult to show (see [16]) that the condition $1 < p \leq 2$ implies that $H_1, H_2 \in L^1(\Omega)$. Then, by Cauchy–Schwarz we get

$$\delta \int_\Omega |\nabla \psi_\delta (u - \hat{u})| = \int_{A_\delta} |\nabla (u - \hat{u})| \leq \left(C\delta^2 \int_{A_\delta} H_1 \right)^{1/2} \left(\int_{A_\delta} H_2 \right)^{1/2}.$$

Finally, using the Poincarè inequality, for any fixed $\mu, \mu > \delta$, we obtain that

$$\begin{aligned} \operatorname{meas} \{u - \hat{u} \geq \mu\} \leq \int_\Omega |\psi_\delta (u - \hat{u})| &\leq C \int_\Omega |\nabla \psi_\delta (u - \hat{u})| \\ &\leq C \left(\int_{A_\delta} H_1 \right)^{1/2} \left(\int_{A_\delta} H_2 \right)^{1/2}. \end{aligned}$$

Letting $\delta \to 0$ we get the conclusion since $\int_{A_\delta} H_i \to 0$ as $\delta \to 0$, for $i = 1, 2$. □

Remark 4. When $N = 1$ the assumption (2.1) in Theorems 3 and 4 can be generalized to the mere assumption that $K(b(\eta))$ be a continuous function of η. This is an easy adaptation of a result due to Benilan [9] (see also Wolanski [55]). We mention the papers Kalashnikov [39], Ishii [36], and Yin Jingxue [57], [58] where the authors prove the uniqueness of the solution of different special cases of the model equation on $\Omega = \mathbb{R}^N$ and without the regularity condition (2.6). The comparison properties of some special bounded solutions without condition (2.6) will be given in the next section.

We also point out that when $g(\cdot, \eta) = \lambda\eta$ with $\lambda > 0$, the estimate (2.15) proves that the abstract operator associated to $\mathcal{A}$ is an accretive operator in $L^1(\Omega)$ (see Benilan [8], [9] and Crandall [21] for the theory and applications of this class of operators). The proof of the case (c) of Proposition 2 is inspired in Boccardo, Gallouët, and Murat [16] (his result cannot be directly applied because their assumption (1) is not satisfied in our case). Finally we remark that a revision of the proof of part (c) shows that the conclusion still holds if the assumed inequality is verified merely for any $\xi \in R_\infty \equiv \{\zeta = \nabla u_\infty(x), \text{ for some } x \in \Omega\}$ and any $\eta, \hat{\eta} \in [-M, M]$ with $M = \max \{\|u_\infty\|, \|\hat{u}_\infty\|_\infty\}$. In particular it holds if we assume $u_\infty \in W^{1,\infty}(\Omega)$ [or $\hat{u}_\infty \in W^{1,\infty}(\Omega)$], $K(b(\cdot))$ locally Lipschitz continuous and $\xi - K(b(\eta))\mathbf{e} \neq \mathbf{0}$ for any $\xi \in R_\infty$ and $\eta \in [-M, M]$. We point out that in the case of Model 1 of the Introduction the function K is usually taken as a regular function such that $K(s) > c > 0$ for any $s \in \mathbb{R}$ and some $c > 0$ (see Bear [7], p. 492).

3. Existence of bounded weak solutions for the model problem. In order to obtain the existence of bounded weak solutions of the model problem (0.1) we shall need to assume some additional conditions on K, g, and f besides the already explicit ones in §2. So, the structure assumption (1.3) will require to assume that

$$\begin{cases} K \text{ is a continuous real function such that} \\ |K(b(\eta))| \leq C \left(1 + |\eta|^\lambda\right) \text{ for all } \eta \in \mathbb{R} \text{ and some } \lambda \in [0, p^*/p]. \end{cases} \tag{3.1}$$

Moreover the boundedness condition of the solution under investigation will be obtained by assuming that $g(\cdot, u)$ satisfies (1.7), (2.2), and also

$$g(x, \eta)\, \eta \geq 0 \quad \text{for any } \eta \in \mathbb{R} \quad \text{and a.e. } x \in \Omega. \tag{3.2}$$

Concerning the right-hand side term, we have

(3.3)

$$\begin{cases} f \in L^1((0,T)\times\Omega) + L^p\left(0,T : W^{-1,p'}(\Omega)\right) \text{ and } |f(t,x)| \le \bar{f}(x) := \operatorname{div} \mathbf{c}(x), \\ \text{for some } \mathbf{c} \in (L^q(\Omega))^N \text{ with } q > N/(p^\# - 1) \\ \text{if } p^\# \le N, q = \max\left(p', (p^\#)'\right) \text{ if } p^\# > N, \\ \text{where } p^\# = \max(p,2). \end{cases}$$

We shall obtain the existence of a bounded weak solution of (0.1) by using comparison arguments and suitable super- and subsolutions of the associated stationary problem. The concrete statement requires previously the next result.

LEMMA 3. *Let $\overline{f}$ be given by* (3.3). *Assume that g satisfies* (1.7) *and* (3.2). *Then there exists $\underline{u}, \overline{u} \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$, with $\underline{u} \le 0 \le \overline{u}$ satisfying*

$$-\operatorname{div} \phi(\nabla \underline{u} - K(b(\underline{u}))\mathbf{e}) + g(x,\underline{u}) = -\bar{f}(x) \quad \text{in } \Omega, \tag{3.4}$$

and

$$-\operatorname{div} \phi(\nabla \bar{u} - K(b(\bar{u}))\mathbf{e}) + g(x,\bar{u}) = \bar{f}(x) \quad \text{in } \Omega, \tag{3.5}$$

where again $\phi(\xi) = |\xi|^{p-2}\xi$ and $p > 1$.

We postpone the proof to later and state our existence result on bounded weak solutions of (0.1).

THEOREM 4. *Assume that the hypothesis* (1.6), (1.7), (2.1), (2.2), and (3.1)–(3.3) *are satisfied. Suppose $u_0 \in L^\infty(\Omega)$ be such that*

$$\underline{u}(x) \le u_0(x) \le \bar{u}(x) \quad \text{a.e. } x \in \Omega. \tag{3.6}$$

Then there exists a bounded weak solution u of problem (0.1).

Proof of Lemma 3. Define $\mathbf{a} : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ by

$$\mathbf{a}(\eta,\xi) = \phi(\xi - K(b(\eta))\mathbf{e}) + \phi(K(b(\eta))\mathbf{e}).$$

From (3.1) we deduce that

$$|\mathbf{a}(\eta,\xi)| \le C\left(|\eta|^{p^*/p} + |\xi|^{p-1}\right)$$

and from (2.3)

$$C|\xi|^p \le (\mathbf{a}(\eta,\xi)\cdot\xi)^{\alpha/2}\left(|\eta|^{p^*} + |\xi|^p\right)^{1-\alpha/2}$$

with $\alpha = 2$ if $p \ge 2$ and $\alpha = p$ if $1 < p \le 2$. Equation (3.5) can be equivalently written as

$$-\operatorname{div} \mathbf{a}(\bar{u}, \nabla\bar{u}) + g(x,\bar{u}) = \operatorname{div}\left(\mathbf{c}(x) + \mathbf{h}(u)\right),$$

with

$$\mathbf{h}(u) = \phi(K(b(\bar{u}))\mathbf{e}).$$

Assumption (3.1) implies that

$$|\mathbf{h}(\eta)| \le C(1 + |\eta|^\mu) \quad \text{with } \mu \in [0, p^*/p']$$

and so the existence of $\overline{u} \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ satisfying (3.5) is assured (if $p \geq 2$) by Theorem 2 of Boccardo and Giachetti [17]. The case $1 < p < 2$ can be obtained by obvious modifications of the mentioned work. This function $\overline{u}$ also satisfies that $\overline{u} \geq 0$ on Ω as we deduce by standard methods once we are assuming (3.2) and div $\mathbf{c}(x) \geq 0$. The existence of $\underline{u}$ is proved analogously. □

Proof of Theorem 4. Let $M > 0, M \geq \max\left(\|\overline{u}\|_{L^\infty(\Omega)}, \|u\|_{L^\infty(\Omega)}\right), \varepsilon > 0$ and $k, j, m, n \in \mathbb{N}^*$. We consider the regularized equation:

(3.7)

$$b_m(u)_t - \varepsilon\Delta u - \operatorname{div}\phi\left[\nabla u - K_j\left(b_m(u)\right)\mathbf{e}\right] + g_n(x,u) = f_{k,1}(t,x)\,\theta(u) + f_{k,2}(t,x),$$

where

$$\begin{cases} b_m(\eta) = \frac{1}{m}\eta + \bar{b}_m(\eta) \text{ with } \bar{b}_m \text{ the Yosida approximation of } b \text{ (it is well known} \\ \text{that } \bar{b}_m \text{ is a Lipschitz nondecreasing function such that } \left|\bar{b}_m\right| \leq |b| \text{ and} \\ \bar{b}_m \to b; \text{ see, e.g., Benilan [8], [9]);} \end{cases}$$

$$\begin{cases} K_j \in C^\infty(\mathbb{R}) \text{ satisfies } \|K_j\|_{L^\infty} \leq \hat{K}, \text{ where} \\ \hat{K} = \sup_{s\in[-2M,2M]} |K[b_1(s)]| \text{ and } K_j \to K\mathbf{1}_{[b_1(-2M),b_1(2M)]} \text{ as } j \to +\infty; \end{cases}$$

$$\begin{cases} g_n \in C^\infty(\Omega \times \mathbb{R}) \text{ satisfies (uniformly on } l) \text{ (1.7), (2.2), and} \\ g_n(x,\eta) \to g(x,\eta) \text{ in } L^1(\Omega) \text{ for any fixed } \eta \text{ and in } \mathbb{R} \text{ for a.e. } x \in \Omega \text{ as } n \to \infty; \end{cases}$$

$$\begin{cases} f_k \in C^\infty\left((0,T)\times\bar{\Omega}\right) \text{ satisfies (uniformly on } n) \text{ the inequality (3.3),} \\ f_k = f_{k,1} + f_{k,2} \text{ and } f_{k,1} \to f_1 \text{ in } L^1((0,T)\times\Omega), \\ f_{k,2} \to f_2 \text{ in } L^{p'}\left((0,T) : W^{-1,p'}(\Omega)\right) \quad \text{as } k \to \infty; \end{cases}$$

$$\begin{cases} \theta \text{ is a truncation function satisfying } \theta \in C^\infty(\mathbb{R}), \quad 0 \leq \theta \leq 1, \\ \theta(\eta) = 1 \text{ for } |\eta| \leq M \text{ and } \theta(\eta) = 0 \quad \text{for } |\eta| \geq 2M. \end{cases}$$

We also consider the regularized stationary equation

$$-\varepsilon\Delta u - \operatorname{div}\phi\left[\nabla u - K_j\left[b_m(u)\right]\mathbf{e}\right] + g_n(x,u) = \bar{f}(x) \text{ in } \Omega. \tag{3.8}$$

Applying Lemma 3 we obtain a function $\overline{u}_{\varepsilon,j,m,n} \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ satisfying (3.8). Analogously we get the existence of the associated function $u_{\varepsilon,j,m,n}$. Finally we regularize the initial condition by considering $u_{0,q} \in C_0^\infty(\Omega)$ such that $\underline{u}_{\varepsilon,j,m,n} \leq u_{0,q} \leq \overline{u}_{\varepsilon,j,m,n}$ and $u_{0,q} \to u_0$ in $L^\infty(\Omega)$ as $q \to \infty$. Equation (3.7) is uniformly parabolic and so by well-known results (see, e.g., Ladyzenskaya, Solonnikov, and Uraltceva [42, Chapt. V]) there exists a unique classical solution $U = u_{\varepsilon,m,j,n,k}$ of (3.7) satisfying

$$\begin{cases} U = 0 & \text{on } (0,T)\times\partial\Omega, \\ b_m(U(0,x)) = b_m\left(u_{0,q}(x)\right) & \text{in } \Omega. \end{cases} \tag{3.9}$$

In order to study the convergence of $u_{\varepsilon,m,j,k,n}$ we need the following result.

LEMMA 4. *The solution U of* (3.7) *and* (3.9) *is bounded in $L^p(0,T : W_0^{1,p}(\Omega))$ and this bound does not depend on ε, j, k, m, n.*

Proof of Lemma 4. We use again the notation $U = u$. Multiplying (3.7) by u, we have

$$\begin{aligned}
&\int_0^T \int_\Omega b_m(u)_t u + \varepsilon \int_0^T \int_\Omega |\nabla u|^2 + \int_0^T \int_\Omega |\nabla u - K_j [b_m(u)]\,\mathbf{e}|^p \\
&\quad = -\int_0^T \int_\Omega \phi\,[\nabla u - K_j [b_m(u)]\,\mathbf{e}] \cdot K_j [b_m(u)]\,\mathbf{e} - \int_0^T \int_\Omega g_n(x,u)u + \int_0^T \int_\Omega f_k(t,x)u \\
&\quad \le \frac{1}{p}\int_0^T \int_\Omega |\nabla u - K_j [b_m(u)]\,\mathbf{e}|^p + \frac{1}{p}\int_0^T \int_\Omega |K_j (b_m(u))\,\mathbf{e}|^p \\
&\qquad + \int_0^T \int_\Omega f_{k,1}(t,x)\theta(u)u + \int_0^T \langle f_{k,2}(t,x), u\rangle\,.
\end{aligned}$$

But

$$\int_0^T \int_\Omega f_{k,1}\theta(u)u \le M\,\|f_{k,1}\|_{L^1((0,T)\times\Omega)}$$

and using Young's inequality there exists a constant $C > 0$ such that

$$\begin{aligned}
\int_0^T \langle f_{k,2}, u\rangle &\le \|f_{k,2}\|_{L^{p'}\left(0,T:W^{-1,p'}(\Omega)\right)} \left[\int_0^T \int_\Omega |\nabla u|^p\right]^{1/p} \\
&\le C\,\|f_{k,2}\|^{p'}_{L^{p'}\left(0,T:W^{-1,p'}(\Omega)\right)} \\
&\quad + \frac{1}{2}\left(\int_0^T \int_\Omega |\nabla u - K_j [b_m(u)]\,\mathbf{e}|^p + \int_0^T \int_\Omega |K_j [b_m(u)]\,\mathbf{e}|^p\right).
\end{aligned}$$

So, we obtain

$$\int_\Omega B_m [u(T)] + \int_0^T \int_\Omega |\nabla u - K_j [b_m(u)]\,\mathbf{e}|^p \le M^*(T),$$

for some $M^*(T) > 0$. Hence the result. □

End of the proof of Theorem 4. As $b_m(U)_t \in L^1((0,T)\times\Omega)$ we can apply Theorem 3 and conclude the inequality $\underline{u}_{\varepsilon,j,m,n}(x) \le U(t,x) \le \overline{u}_{\varepsilon,j,m,n}(x)$ for $(t,x) \in (0,T)\times\Omega$. Moreover, a careful revision of the proof of Theorem 2 of [17] allows to check that $\|\overline{u}_{\varepsilon,j,m,n}\|_{L^\infty(\Omega)}$ is bounded by a constant independent of ε, j, m, and n. Using this fact, Lemma 4, and proceeding as in Theorem 1, we can pass to the limit as $\varepsilon \to 0$ and $m, j, k, n \to +\infty$, obtaining that $\overline{u}_{\varepsilon,j,m,n} \to \overline{u}, \underline{u}_{\varepsilon,j,m,n} \to \underline{u}$ at least weakly* in $L^\infty(\Omega)$, weakly in $W_0^{1,p}(\Omega)$, and also that $U \to u$ weakly in $L^p(0,T : W_0^{1,p}(\Omega))$ with u as a bounded weak solution of (0.1), satisfying

$$\underline{u}(x) \le u(t,x) \le \bar{u}(x) \quad \text{for } (t,x) \in (0,T)\times\Omega. \qquad \square \tag{3.10}$$

Remark 5. When b is assumed to be *strictly increasing* the existence of a bounded weak solution of (0.1) can be obtained for *any* $u_0 \in L^\infty(\Omega)$ (i.e., not necessarily satisfying (3.6)) if we suppose $f \in L^1(0,T : L^\infty(\Omega))$. Indeed, in that case we can repeat the proof of Theorem 3 but replacing $\overline{u}(x)$ by the supersolution

$$\bar{u}(t,x) = b^{-1}\left(\|u_0\|_{L^\infty(\Omega)} + \int_0^t \|f(s,\cdot)\|_{L^\infty(\Omega)}\,ds\right).$$

The process followed in the proof of Theorem 4 is useful to obtain general comparison results. Indeed, Theorem 3 assumes the regularity condition (2.6) which is very hard to check in some cases. We shall show in §4 that this condition is verified by any bounded weak solution of (0.1) if we additionally suppose that b is a locally Lipschitz function. Nevertheless we have the following result.

COROLLARY 1. *Assume the same hypotheses on b, K, and g given in Theorem 4. Let $(f, u_0), (\hat{f}, \hat{u}_0)$ be a couple of data satisfying* (3.3), (3.6), *and the analogous versions for $\hat{f}$ and $\hat{u}_0$. Assume also that $f \leq \hat{f}$ and $u_0 \leq \hat{u}_0$. Then there exist u and $\hat{u}$ weak solutions of the associated problems* (0.1) *such that $u \leq \hat{u}$ in $(0,T) \times \Omega$.*

Proof. Let $U = u_{\varepsilon,m,j,n,k}$ and $\hat{U} = \hat{u}_{\varepsilon,m,j,n,k}$ be the classical solutions obtained by the process described in the proof of Theorem 4 associated to the regularization of the data $f_k, \hat{f}_k, u_{0,q}$ and $\hat{u}_{0,q}$. Without loss of generality we can assume that $f_k \leq \hat{f}_k$ and $u_{0,q} \leq \hat{u}_{0,q}$. Then, as $b_m(U)_t, b_m(\hat{U})_t \in L^1((0,T) \times \Omega)$ we can apply Theorem 3 and obtain $U \leq \hat{U}$. Finally, the conclusion follows by passing to the limit as $\varepsilon \to 0$ and $m, j, n, k \to +\infty$. □

Remark 6. As far as we know the existence of solutions for the model problem has not be treated in the literature. Nevertheless there are many papers which obtain the existence of solutions for some similar problems. We mention explicitly the important work by Alt and Lukhaus [3] and their generalization made in Kaçur [37], [38]. Another point of view is presented in Blanchard and Francfort [13], [14]. Other related works are due to Bermudez, Durany, and Saguez [10], Bernis [11], Esteban and Vazquez [31], Simondon [52], Tsutsumi [53] and Xu [56] (see also the references in the mentioned papers). The comparison between solutions which are limits of sequences of more regular solutions is already an old argument (see Benilan [8], Bamberger [5], [6], and Blanchard and Francfort [14]).

4. Stabilization results for the model problem. Theorem 1 reduces the stabilization of bounded weak solutions to the study of conditions (1.16) and (1.17). We shall start this section by showing that the comparison principle (Corollary 1) and the uniqueness of solutions of the stationary problem (1.14) allows reduction of the stabilization property to the study of condition (1.16) for solutions that are monotone in time.

PROPOSITION 3. *Assume the hypotheses* (1.6) *on b,* (2.1) *and* (3.1) *on K, and* (1.7) *and* (2.2) *on g. Assume also that the stationary problem* (2.10) *and* (2.11) *has a unique bounded weak solution. Let f and f_∞ satisfy* (1.8), (1.13), *and* (3.3), *and assume that there exists $f_+(t,x), f_-(t,x)$ satisfying* (1.8) *with f_+ (respectively, f_-) monotone nonincreasing in t (respectively, nondecreasing) and such that*

$$-\bar{f}(x) \leq f_-(t,x) \leq f(t,x) \leq f_+(t,x) \leq \bar{f}(x) \quad \text{in } (0,\infty) \times \Omega, \tag{4.1}$$

($\bar{f}$ given in (3.3)) *and also satisfying*

$$\lim_{t\to\infty} f_+(t,\cdot) = \lim_{t\to\infty} f_-(t,\cdot) = f_\infty(\cdot) \quad \text{in } L^1(\Omega) + W^{-1,p'}(\Omega). \tag{4.2}$$

Let u, u_+ and u_- be the bounded weak solutions of (0.1) *associated to the data (f, u_0), $(f_+, \overline{u})$, and $(f_-, \overline{u})$, respectively, assured by Corollary* 1 *(with $\overline{u}, \underline{u}$ given in Lemma* 3). *Then if u_+, u_- satisfies* (1.16) *for any $u_\infty \in \omega(u)$ we deduce that u_∞ is a bounded solution of the stationary problem and in fact*

$$u(t,\cdot) \to u_\infty \quad \text{in } L^r(\Omega), \text{ as } t \to \infty, \text{ for any } r \in [1,\infty).$$

Proof. We shall follows closely an argument already used in Kröner and Rodrigues [41] (Theorem 6). First of all we point out that the assumptions made on g imply the condition (3.2) and then by Corollary 1 and the proof of Theorem 1 we deduce the existence of the mentioned bounded weak solutions u, u_+, and u_-. Moreover, we have

$$\underline{u}(x) \leq u_-(t,x) \leq u(t,x) \leq u_+(t,x) \leq \bar{u}(x) \qquad (t,x) \in (0,\infty) \times \Omega. \tag{4.3}$$

By comparison on the related regularized solutions and using (3.4) and (3.5) it is easy to see that $u_+(t,\cdot)$ (respectively, $u_-(t,\cdot)$) is monotone nonincreasing in t (respectively, nondecreasing). Then there exists $u_{\infty,+}(x), u_{\infty,-}(x)$ such that

$$u_+(t,\cdot) \to u_{\infty,+}, u_-(t,\cdot) \to u_{\infty,-} \text{ in } L^r(\Omega), \text{ for any } r \in [1,\infty) \text{ as } t \to \infty. \tag{4.4}$$

From (4.3) we deduce that

$$u_{\infty,-}(x) \leq u_\infty(x) \leq u_{\infty,+}(x) \quad \text{in } \Omega. \tag{4.5}$$

Now as u_+, u_- satisfies (1.16) and (1.17) (due to (4.4)) then Theorem 1 shows that $u_{\infty,+}$ and $u_{\infty,-}$ are bounded weak solutions of the same associated stationary problem (i.e., (1.14) with $\mathbf{A}$ given by (1.5); recall (4.2)). Finally, from the uniqueness of the bounded weak solution of the stationary problem and (4.5) we deduce that $u_{\infty,-} = u_\infty = u_{\infty,+}$ and so we have the conclusion. □

The important assumption (1.16) will be obtained in the two following results.

THEOREM 5. *Assume*

$$1 < p \leq 2, \tag{4.6}$$

g satisfies (1.7), (2.2), (3.2), *and f verifies* (3.3) *and*

(4.7)

$$\begin{cases} f \in L^\infty\left(0,\infty : L^1(\Omega) + W^{-1,p'}(\Omega)\right) \cap W^{1,1}_{\text{loc}}\left(0,\infty : L^1(\Omega) + W^{-1,p'}(\Omega)\right) \text{ and} \\ \displaystyle\int_t^{t+1} \int_\Omega \left\| \frac{\partial f}{\partial t} \right\|_{L^1(\Omega)+W^{-1,p'}(\Omega)} \leq C, \text{ for any } t > 0 \text{ and some } C \text{ independent on } t. \end{cases}$$

Let u_0 satisfy (3.6) *and also*

$$u_0 \in W_0^{1,p}(\Omega). \tag{4.8}$$

Finally, assume one of the following set of hypothesis:

$$\text{(A)} \qquad \begin{cases} \text{(4.9)} & b \text{ is a nondecreasing locally Lipschitz function,} \\ \text{(4.10)} & K \text{ is a locally Lipschitz function satisfying (3.1),} \end{cases}$$

or

$$\text{(B)} \qquad \begin{cases} \text{(4.11)} & b^{-1} \text{ is a nondecreasing locally Lipschitz function,} \\ \text{(4.12)} & K(b(\cdot)) \text{ is a locally Lipschitz function satisfying (3.1) holds.} \end{cases}$$

Then if u is the bounded weak solution given in Theorem 4 we have $u \in L^\infty(0,\infty : W_0^{1,p}(\Omega))$. Moreover, for any $t > 0$ and some $C > 0$ independent of t we have

$$\int_t^{t+1} \int_\Omega |b(u)_t| \leq C \tag{4.13}$$

when (A) *is satisfied, and*

$$\int_t^{t+1} \int_\Omega |u_t| \leq C \tag{4.14}$$

if (B) *holds.*

Proof. Assume that case (A) holds. Multiplying (0.1) by u, we have for any τ, σ satisfying $\tau > \sigma \geq \tau - 1 > 0$

$$\int_\sigma^\tau \int_\Omega |\nabla u - K[b(u)]\mathbf{e}|^p \leq C. \tag{4.15}$$

This is obtained by an easy adaptation of the proof of Lemma 4. Let us define

$$E(t) = \int_\Omega |\nabla u(t) - K[b(u(t))]\mathbf{e}|^p .$$

Assume that u_t is regular enough (otherwise we first work with the approximate solution $u_{\epsilon,m,j,k}$ and then pass to the limit). Taking $v = u_t$ in (1.11) we get

(4.16)

$$\int_\sigma^\tau \int_\Omega b(u)_t u_t + E(\tau) - E(\sigma) + J + \int_\Omega G(\cdot, u(\tau)) - \int_\Omega G(\cdot, u(\sigma)) = \int_\sigma^\tau \int_\Omega f u_t,$$

where $G(x, \cdot)$ is the primitive of $g(x, \cdot)$ and

$$J = \int_\sigma^\tau \int_\Omega \phi \left[\nabla u - K\left[b\left(u\right)\right]\mathbf{e}\right] \frac{dK}{ds}[b(u)]b(u)_t.$$

Then, using (4.10) and (4.6) and Young's inequality

$$\begin{aligned} J &\leq \varepsilon \int_\sigma^\tau \int_\Omega |b(u)_t|^2 + C\left(\varepsilon\right) \int_\sigma^\tau \int_\Omega |\nabla u - K[b(u)]\mathbf{e}|^{2(p-1)} \\ &\leq \varepsilon \int_\sigma^\tau \int_\Omega |b(u)_t|^2 + C\left(\varepsilon\right) \left[\left(\tau - \sigma\right)|\Omega|\right]^{1/q'} \left\{ \int_\sigma^\tau \int_\Omega |\nabla u - K(b(u))\mathbf{e}|^p \right\}^{1/q} \end{aligned}$$

with $q = p/(2(p-1))$. From (4.15) we obtain

$$J \leq \varepsilon \int_\sigma^\tau \int_\Omega |b(u)_t|^2 + C'\left(\varepsilon\right).$$

Moreover, by (4.7) and (3.10) if $f = f_1 + f_2$ with $f_1 \in L^\infty(0, \infty : L^1(\Omega))$ and $f_2 \in L^\infty(0, \infty : W^{-1,p'}(\Omega))$ we have

$$\begin{aligned} \int_\sigma^\tau \int_\Omega f_1 u_t &= \int_\Omega f_1(\tau, \cdot)u(\tau) - \int_\Omega f_1(\sigma, \cdot)u(\sigma) - \int_\sigma^\tau \int_\Omega \frac{\partial f_1}{\partial t} \\ &\leq C_1 + \left(\left(\tau - \sigma\right)|\Omega|\right)^{1/2} M \left(\int_\sigma^\tau \int_\Omega \left| \frac{\partial f_1}{\partial t} \right|^2 \right)^{1/2} \leq C_2. \end{aligned}$$

The term $\int_\sigma^\tau \langle f_2, u_t \rangle$ is treated in an analogous way to the proof of Lemma 4. From (4.9) we have

$$\frac{1}{L}\int_\sigma^\tau \int_\Omega |b(u)_t|^2 \leq \int_\sigma^\tau \int_\Omega b(u)_t u_t$$

for some $L = L(M) > 0$. Then, we obtain (for ϵ small enough)

$$\frac{1}{2L}\int_\sigma^\tau \int_\Omega |b(u)_t|^2 + E(\tau) - E(\sigma) \leq C. \tag{4.17}$$

In case (B) we use (4.11) and conclude

$$\frac{1}{L}\int_\sigma^\tau \int_\Omega (u_t)^2 \leq \int_\sigma^\tau \int_\Omega b(u)_t u_t.$$

Moreover, by (4.10)

$$\begin{aligned}\int_\sigma^\tau \int_\Omega |\nabla u - K[b(u)]\mathbf{e}|^{p-1} \left|\frac{d}{ds}(K \circ b)(u)\right| |u_t| \\ \leq \frac{1}{4L}\int_\sigma^\tau \int_\Omega (u_t)^2 + C\int_\sigma^\tau \int_\Omega |\nabla u - K(b(u))\mathbf{e}|^{2(p-1)}\end{aligned}$$

and as in the case (A) we arrive to

$$\frac{1}{2L}\int_\sigma^\tau \int_\Omega (u_t)^2 + E(\tau) - E(\sigma) \leq C. \tag{4.18}$$

The result follows from the following well-known result. □

LEMMA 5 (Nakao [49]). *Let $\varphi(t) \geq 0$ be a locally bounded function satisfying*

$$\varphi(t+1) \leq C\,[\varphi(t) - \varphi(t+1)] + \rho(t) \quad \textit{for } t > 0,$$

where C is a positive constant and $\rho > 0$ for large t. Then as $t \to +\infty$ one has:

$$\varphi(t) = 0(1)\ [\textit{respectively, } o(1)]\ \textit{if } \rho(t) = 0(1)\ [\textit{respectively}, o(1)].$$ □

When $p > 2$ we shall prove that condition (1.16) holds at least for the super- and subsolutions u_+ and u_-.

THEOREM 6. *Assume*

$$p \geq 2 \tag{4.19}$$

and suppose the same hypothesis than in Proposition 3 but with $g(\cdot, u)$ merely a non-decreasing function satisfying (1.7). *Then $u_+, u_- \in L^\infty(0, \infty : W_0^{1,p}(\Omega))$.*

Proof. From the proof of Proposition 3 we know that $u_+(t, \cdot)$ is monotone and nonincreasing in t and then $b(u_+(t, \cdot))$ satisfies this same property. Taking as test function $v = \bar{u} - u_+(t, \cdot)$ in the conditions (1.11) for u_+ and (1.15) for $\bar{u}$ we obtain

$$\begin{aligned}I_1(t) &= \int_\Omega \left[\phi(\nabla \bar{u} - K\left(b\left(\bar{u}\right)\right)\mathbf{e}) - \phi(\nabla u_+\left(t, \cdot\right) - K\left(b\left(u_+\left(t, \cdot\right)\right)\right)\mathbf{e})\right] \cdot (\nabla \bar{u} - \nabla u_+\left(t, \cdot\right)) \\ &\leq \int_\Omega \left(\bar{f} - f_+\left(t, \cdot\right)\right)\left(\bar{u} - u_+\left(t, \cdot\right)\right),\end{aligned}$$

where we have used that $v(t,\cdot) \geq 0$ for almost every $t > 0$ (see (4.3)). As in the proof of Theorem 3 we write

$$I_1(t) = I_2(t) + I_3(t)$$

with

$$\begin{aligned} I_2(t) = &\int_\Omega \{\phi[\nabla \bar{u} - K(b(\bar{u}))\,\mathbf{e}] - \phi[\nabla u_+(t,\cdot) - K(b(u_+(t,\cdot)))\,\mathbf{e}]\} \\ &\cdot \{\nabla \bar{u} - K(b(\bar{u}))\,\mathbf{e} - \nabla u_+(t,\cdot) + K(b(u_+(t,\cdot)))\,\mathbf{e}\} \\ I_3(t) = &\int_\Omega \{\phi[\nabla \bar{u} - K(b(\bar{u}))\,\mathbf{e}] - \phi[\nabla u_+(t,\cdot) - K(b(u_+(t,\cdot)))\,\mathbf{e}]\} \\ &\cdot \mathbf{e}\,[K(b(\bar{u})) - K(b(u_+(t,\cdot)))]\,. \end{aligned}$$

By Young's inequality we have that for any $\varepsilon > 0$

$$\begin{aligned} I_3(t) \leq &\frac{\varepsilon}{p'} \int |\phi[\nabla \bar{u} - K(b(\bar{u}))\,\mathbf{e}] - \phi[\nabla u_+(t,\cdot) - K(b(u_+(t,\cdot)))\,\mathbf{e}]|^{p'} \\ &+ \frac{C}{\varepsilon p} \int_\Omega |K(b(\bar{u})) - K(b(u_+(t,\cdot)))|^p\,. \end{aligned}$$

Using (2.5) and (4.3) we have that

$$I_3(t) \leq \frac{\varepsilon}{p'}\, I_2(t) + C_2$$

for some $C_2 > 0$ independent of t. From (2.3), (4.2), and (4.3) we get

$$\int_\Omega |\nabla \bar{u} - \nabla u_+(t,\cdot) - (K(b(\bar{u})) - K(b(u_+(t,\cdot)))\,\mathbf{e})|^p \leq C_3$$

for some $C_3 > 0$ independent of t. Finally using again that $u_+ \in L^\infty((0,\infty)\times\Omega)$ and that $\bar{u} \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ the conclusion follows for u_+. The proof for u_- is analogous. □

COROLLARY 2. *Assume the conditions of Proposition* 3 *and also* (4.7), (4.8) *and* [(4.9), (4.10)] *or* [(4.11), (4.12)] *if* $1 < p \leq 2$. *Then if* u *is the bounded weak solution of* (0.1) *associated to the data* (f, u_0) *assured by Corollary* 1, *for any* $u_\infty \in \omega(u)$ *we have that* u_∞ *is a bounded weak solution of the stationary problem and in fact* $u(t,\cdot) \to u_\infty$ *in* $L^r(\Omega)$, *as* $t \to \infty$, *for any* $r \in [1,\infty)$.

Our last result proves the condition (1.21) for the special case of

$$K(b(s)) = \lambda s \quad \text{for some } \lambda \in \mathbb{R} \text{ and any } s \geq 0. \tag{4.20}$$

We point out that similar properties to (1.21) have been proved in the literature when the elliptic operator $\mathcal{A}$ is assumed to be the gradient (or subdifferential) of some potential functional (see, e.g., Langlais and Phillips [43], Tsutsumi [53] and El Hachimi and de Thelin [29], [30]) but we also remark that when $K \not\equiv 0$ the associated elliptic operator $\mathcal{A}$ does not satisfy this structure condition.

THEOREM 7. *Assume that* g *satisfies* (1.7), (2.2), (3.2), *and* f *verifies* (3.3) *and* (4.7). *Let* u_0 *satisfying* (3.6) *and* (4.8). *Assume* (4.20) *and that* b *satisfies* (4.9) *or* (4.11). *Then if* u *is the bounded weak solution given in Corollary* 1, u *satisfies that* $u \in L^\infty(0,\infty : W_0^{1,p}(\Omega))$. *Moreover we have that*

$$b(u)_t \in L^2((0,\infty)\times\Omega) \quad \textit{if } b \textit{ satisfies } (4.9) \tag{4.21}$$

and

$$u_t \in L^2((0,\infty)\times\Omega) \quad \textit{if (4.11) holds.} \tag{4.22}$$

Proof. Assume that b satisfies (4.9). Without loss of generality we can assume $\mathbf{e} = \mathbf{e_1}$ (the first term of the orthonormal base of $\mathbb{R}^N$); otherwise, it is enough to make a change of base on $\mathbb{R}^N$. Multiplying by $e^{-\lambda x_1} u_t$ we have (assuming u_t is regular) that for any $T > 0$ we have

$$\int_0^T \int_\Omega b(u)_t u_t e^{-\lambda x_1} + \frac{1}{p}\int_0^T \int_\Omega e^{\lambda(p-1)x_1}\frac{\partial}{\partial t}|\nabla e^{-\lambda x_1} u|^p$$

$$+ \int_0^T \int_\Omega e^{-\lambda x_1}\frac{\partial}{\partial t} G(\cdot, u) = \int_0^T \int_\Omega e^{-\lambda x_1} f u_t.$$

The same kind of arguments of the proof of Theorem 5 leads to the conclusion

$$\int_0^T \int_\Omega |b(u)_t|^2 + E(T) - E(0) \le C$$

with C independent of T and then (4.21) follows. The proof of (4.22) is similar. □

COROLLARY 3. *Assume the hypothesis of Theorem 7 and also b strictly increasing if* (4.9) *holds. Then if u is the bounded weak solution of* (0.1) *assured by Corollary 1 we have that* $\omega(u) \neq \phi$ *and any* $u_\infty \in \omega(u)$ *is a bounded weak solution of the stationary problem. Moreover, there exists* $\tilde{t}_n \to +\infty$ *such that* $u(\tilde{t}_n, \cdot) \to u_\infty$ *strongly in* $W_0^{1,p}(\Omega)$.

Proof. Taking $k(s) = b(s)$ if (4.9) holds and $k(s) = s$ if b satisfies (4.11), from Proposition 1 and (2.3) we have that Theorem 2 can be applied, leading to the conclusion. □

Acknowledgments. The authors thank the referees for several observations on a preliminary version of the paper. They also thank L. Boccardo for pointing out reference [17] to the first author.

REFERENCES

[1] N. AHMED AND D.K. SUNADA, *Nonlinear flow in porous media,* J. Hydraulics Div. Proc. Amer. Soc. Civil Engrg., 95 (1969), pp. 1847–1857.

[2] N.D. ALIKAKOS AND P.W. BATES, *Stabilization of solutions for a class of degenerate equations in divergence form in one space dimension,* J. Differential Equations, 73 (1988), pp. 363–393.

[3] H.W. ALT AND S. LUCKHAUS, *Quasilinear Elliptic Parabolic Differential Equations,* Math. Z, 183 (1983), pp. 311–341.

[4] M. ARTOLA, *Sur une classe de problémes paraboliques quasilinaires,* Boll. Un. Mat. Ital., 5-B (1986), pp. 51–70.

[5] A. BAMBERGER, *Etude d'une èquation doublement non linèaire,* J. Funct. Anal., 24 (1977), pp. 148–155.

[6] ———, *Etude d'une èquation doublement non linèaire,* Rapport du Centre de Mathematiques Appliquées, Ecole Polytechnique, 1977. (Extended version of [5].)

[7] J. BEAR, *Dynamics of Fluids in Porous Media.* Elsevier, New York, 1972.

[8] PH. BENILAN, *Equations d'èvolution dans un espace de Banach quelconque et applications,* thesis, Univ. d'Orsay, Orsay, France, 1972.

[9] ———, *Evolution equations and accretive operators,* Lecture Notes, Univ. of Kentucky, Lexington, KY, 1981.

[10] A. BERMUDEZ, J. DURANY, AND C. SAGUEZ, *An existence theorem for an implicit nonlinear evolution equation,* Collect. Math., 35 (1984), pp. 19–34.

[11] F. BERNIS, *Existence results for double nonlinear higher order parabolic equations on unbounded domains,* Math. Ann., 279 (1988), pp. 373–394.

[12] J.G. BERRYMAN AND C.J. HOLLAND, *Stability of the separable solution for fast diffusion,* Arch. Rational Mech. Anal., 74 (1980), pp. 379–388.

[13] D. BLANCHARD AND G. FRANCFORT, *Study of a double nonlinear heat equation with no growth assumptions on the parabolic term,* SIAM J. Math. Anal., 19 (1988), pp. 1032–1056.

[14] ———, *A few results on degenerate parabolic equations,* Ann. Scuola Norm. Sup. Pisa Cl. Sci., 18 (1991), pp. 213–279.

[15] L. BOCCARDO, J.I. DIAZ, D. GIACHETTI, AND F. MURAT, *Existence and regularity of renormalized solutions for some elliptic problems involving derivatives of nonlinear terms,* J. Differential Equations, to appear.

[16] L. BOCCARDO, TH. GALLOUËT, AND F. MURAT, *Unicité de la solution de certaines équations elliptiques nonlinèaires,* C.R. Acad. Sci. Paris, 315 (1992), pp. 1159–1164.

[17] L. BOCCARDO AND D. GIACHETTI, *Existence results via regularity for some nonlinear elliptic problems,* Comm. Partial Differential Equations, 14 (1989), pp. 663–680.

[18] J. CARRILLO AND M. CHIPOT, *On some nonlinear elliptic equations involving derivatives of the nonlinearity,* Proc. Roy. Soc. Edinburgh Ser. A, 100 (1985), pp. 281–294.

[19] M. CHIPOT AND G. MICHAILLE, *Uniqueness results and monotonicity properties for strongly nonlinear elliptic variational inequalities,* Ann. Scuola Norm. Sup. Pisa, Cl. Sci., (1989), pp. 137–166.

[20] M. CHIPOT AND J.F. RODRIGUES, *Comparison and stability of solutions to a class of quasilinear parabolic problems,* Proc. Royal Soc. of Edinburgh Ser. A, 110 (1988), pp. 275–285.

[21] M.G. CRANDALL, *Nonlinear semigroups and evolution governed by accretive operators,* in Nonlinear Functional Analysis and Its Applications, F.E. Browder, ed., Proc. of Symposia in Pure Math., Vol. 45 (1986), pp. 305–338.

[22] J.I. DIAZ, *Nonlinear pde's and free boundaries, Vol. 1, Elliptic Equations,* Research Notes in Math. 106, Pitman, London, 1985.

[23] ———, *Nonlinear pde's and free boundaries, Vol. 2, Parabolic and Hyperbolic Equations,* in preparation.

[24] J.I. DIAZ AND M.A. HERRERO, *Estimates on the support of the solution of some nonlinear elliptic and parabolic problems,* Proc. Royal Soc. Edinburgh Ser., 89 (1981), pp. 249–258.

[25] J.I. DIAZ AND R. KERSNER, *On a nonlinear degenerate parabolic equation in infiltration or evaporation,* J. Differential Equations, 69 (1987), pp. 368–403.

[26] J.I. DIAZ AND A. LIÑÀN, *Tiempo de descarga en oleoductos o gaseoductos largos: Modelizaciòn y estudio de una ecuación parabólica doblemente no lineal,* in Actas de la Reunión Matemática en Honor a A. Dou, J.I. Diaz and J.M. Vegas, eds., Univ. Complutense, Madrid (1989), pp. 95–120.

[27] J.I. DIAZ AND L. VERON, in preparation.

[28] C. J. VAN DUIJN AND D. HILHORST, *On a doubly nonlinear equation in hydrology,* Nonlinear Anal. T.M.A.A., 11 (1987), pp. 305–333.

[29] A. EL HACHIMI AND F. DE THELIN, *Supersolutions and stabilization of the solutions of the equation* $\partial u/\partial t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = f(x,u)$, Nonlinear Anal. TMA, 12 (1988), pp. 1385–1398.

[30] ———, *Supersolutions and stabilization of the solutions of the equation* $\partial u/\partial u - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = f(x,u)$, Part. II, Publ. Mat., 35 (1981), pp. 347–362.

[31] J.R. ESTEBAN AND J.L. VAZQUEZ, *Homogeneous diffusion in* $\mathbb{R}$ *with power-like nonlinear diffusivity,* Arch. Rational Mech. Anal., 103 (1988), pp. 39–80.

[32] G. GAGNEUX AND F. GUERFI, *Approximations de la fonction de Heaviside et résultats d' unicité pour une classe de problèmes quasi-linéaires elliptiques-paraboliques,* Rev. Mat. Univ. Complut. Madrid, 3 (1990), pp. 59–87.

[33] B.H. GILDING, *The soil-moisture zone in a physically-based hydrologic model,* Advances in Water Resources, 6 (1983), pp. 36–43.

[34] ———, *Improved theory for a nonlinear degenerate parabolic equation,* Ann. Scuola Norm. Sup. Pisa, Cl. Sci. 14 (1989), pp. 165–224.

[35] A.A. HANNOURA AND F.B.J. BARENDS, *Non Darcy flow: a state of the art,* in Flow and Transport in Porous Media, A. Verruijt and F.B.J. Barends, eds., (1982), pp. 37–51.

[36] H. ISHII, *Asymptotic stability and blowing up of solutions of some nonlinear equations,* J. Differential Equations, 26 (1977), pp. 291–319.

[37] J. KAČUR, *On a solution of degenerate elliptic-parabolic systems in Orlicz-Sobolev spaces* I., Math. Z., 203 (1990), pp. 153–171.

[38] ———, *On a solution of degenerate elliptic-parabolic systems in Orlicz-Sobolev spaces* II. Math. Z., 203 (1990), pp. 569–579.

[39] A.S. KALASHNIKOV, *Some problems of the qualitative theory of nonlinear degenerate second-order parabolic equations,* Russian Math. Surveys, 42 (1987), pp. 169–222.

[40] S. KICHENASSAMY AND J. SMOLLER, *On the existence of radial solutions of quasilinear elliptic equations,* Nonlinearity, 3 (1990), pp. 677–694.

[41] D. KRÖNER AND J.F. RODRIGUES, *Global behaviour for bounded solutions of a porous media equation of elliptic parabolic type,* J. Math. Pures Appl., 64 (1985), pp. 105–120.

[42] O.A. LADYZHENSKAYA, V.A. SOLONNIKOV, AND N.N. URALTCEVA, *Linear and Quasi-Linear Equations of Parabolic Type,* Trans. Amer. Math. Soc., Providence, RI, 1968.

[43] M. LANGLAIS AND D. PHILLIPS, *Stabilization of solutions of nonlinear and degenerate evolution equations,* Nonlinear Anal. TMA, 9 (1985), pp. 321–333.

[44] L.S. LEIBENSON, *General problem of the movement of a compressible fluid in a porous medium,* Izv. Akad. Navk. SSSR, Geography and Geophysics, 9 (1945), pp. 7–10. (In Russian.)

[45] A. LIÑÀN, *Line packing and surge attenuation in long pipelines,* unpublished work.

[46] J.L. LIONS, *Quelques mèthodes de rèsolution de problèmes aux limites non linèaires,* Dunod, Paris, 1969.

[47] L.K. MARTINSON AND K.B. PAVLOV, *Unsteady shear flows of a conducting fluid with a rheological power law,* Magnit. Gidrodinamika, 2 (1971), pp. 30–58. (In Russian.)

[48] H. MATANO, *Existence of nontrivial unstable sets for equilibriums of strongly order-preserving systems,* J. Fac. Sci. Univ. Tokyo, Sec. 1A, 30 (1984), pp. 645–673.

[49] M. NAKAO, *A difference inequality and its application to nonlinear evolution equations,* J. Math. Soc. Japan, 30 (1978), pp. 747–762.

[50] A.S. SHAPIRO, *Compressible Fluid Flow, Vol.* II, Ronald Press, New York, 1954.

[51] J. SIMON, *Régularité de la solution d'un problème aux limites non linéaire,* Ann. Fac. Sci. Toulouse Math. (5), 3 (1981), pp. 247–274.

[52] F. SIMONDON, *Etude de l'èquation* $\partial_t b(u) - \operatorname{div} a(b(u), \nabla u) = 0$, Publ. Mat., Univ. Besançon, France, 1982.

[53] M. TSUTSUMI, *On solutions of some doubly nonlinear degenerate parabolic equations with absorption,* J. Math. Anal. Appl., 60 (1987), pp. 543–549.

[54] R.E. VOLKER, *Nonlinear flow in porous media by finite elements,* J. Hydraulics Div. Proc. Amer. Soc. Civil. Eng., 95 (1969), pp. 2093–2114.

[55] N.I. WOLANSKI, *Flow through a porous column,* J. Math. Anal. Appl., 109 (1985), pp. 140–159.

[56] X. XU, *Existence and Convergence Theorems for Doubly Nonlinear Partial Differential Equations of Elliptic-Parabolic Type,* J. Math. Anal. Appl., 150 (1990), pp. 205–223.

[57] J. YIN, *On a class of quasilinear parabolic equations of second order with double-degeneracy,* J. Partial Differential Equations, 3 (1990), pp. 49–64.

[58] ———, *Solutions with compact support for nonlinear diffusion equations,* Nonlinear Anal. TMA, 19 (1992), pp. 309–321.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1112–1121, July 1994

004

ON A GLOBALLY EXISTING SOLUTION TO THE INVISCID BURGERS EQUATION WITH A NONLOCAL TERM*

KAZUO ITO†

Abstract. It is shown that the inviscid Burgers equation with a nonlocal nonlinear term admits smooth global solutions for certain initial data which are smooth and nondecreasing. This result corresponds to a similar situation in the classical inviscid Burgers equation and is complementary to a result in a paper by R. Gardner [*SIAM J. Math. Anal.*, 18 (1987), pp. 172–183].

Key words. conservation laws, integro-differential equation, Cauchy problem

AMS subject classifications. primary 35Q35; secondary 35F25, 35L60, 35L65

1. Introduction and main results.

We consider the initial value problem

$$u_t + a\left(\frac{u^2}{2}\right)_x + b\left(\int_0^\infty u(x+\beta s,t)u_x(x+s,t)ds\right)_x = 0 \quad \text{in } \boldsymbol{R}\times(0,\infty), \tag{1}$$

$$u(x,0) = u_0(x) \quad \text{in } \boldsymbol{R} \tag{2}$$

where a, b, and β are real parameters with the constraint

$$a \neq b, \qquad \beta > 1.$$

Note that (1) involves a nonlocal term. Majda and Rosales proposed (1) as the equation governing the planar detonation front of reacting gases ([4]; see also [3] and [5]). When $a - b = 1$, Gardner proved in [2] that (1)–(2) has local solutions with Sobolev space data. He also showed that a smooth solution forms shocks in finite time when u_0, a, and β satisfy the following conditions:

(i) $u_0 \in C_c^\infty(\boldsymbol{R})$ and supp $u_0 \subset (-\infty, 0]$,
(ii) there exists $y < 0$ such that $u_0''(x) > 0$ for any $x \in (y, 0)$,
(iii) $0 < a < 1$, $\beta > 1$,
(iv) $1 - (1-a)(\beta - \beta^{-1})/2 > 0$.

Remark that from (i) and (ii) $u_0'(x) < 0$ for any $x \in (y, 0)$.

By the way, we know that the inviscid Burgers equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0 \quad \text{in } \boldsymbol{R}\times(0,\infty),$$

$$u(x,0) = u_0(x) \in C^1(\boldsymbol{R})$$

has a global classical solution if $u_0'(x) \geq 0$ for any $x \in \boldsymbol{R}$ while any C^1-solution forms shocks in finite time if $u_0'(x) < 0$ for some $x \in \boldsymbol{R}$.

From the above facts, we expect that (1)–(2) have a global solution if, roughly speaking, $u_0''(x) \leq 0$ for any $x \in \boldsymbol{R}$ (accordingly, $u_0'(x) \geq 0$ for any $x \in \boldsymbol{R}$ if $u_0'(\infty) = 0$) and a and b satisfy some conditions. We will show that this is in fact the case.

To state our results, we make some arrangements for (1). Now we assume that u has the second derivative with respect to x and $u, u_x \to 0$ as $x \to +\infty$. We put

* Received by the editors July 27, 1992; accepted for publication May 5, 1993.
† Department of Applied Science, Faculty of Engineering, Kyushu University, 36 Fukuoka, 812 Japan.

$v(x,t) = (a-b)u(x,t)$ and substitute into (1) and (2). Differentiating under integral sign and integrating by parts, we are led to

$$v_t + vv_x - b(a-b)^{-1}(\beta - 1)I(v_x, v_x) = 0, \tag{3}$$

$$v(x,0) = (a-b)u_0(x), \tag{4}$$

where

$$I(f,g) = \int_0^\infty f(x+\beta s)g(x+s)ds.$$

Hence, the equation we study from now on is

$$u_t + uu_x + \gamma I(u_x, u_x) = 0 \quad \text{in } \boldsymbol{R} \times (0,\infty), \tag{5}$$

$$u(x,0) = u_0(x) \quad \text{in } \boldsymbol{R}. \tag{6}$$

Our main result is the following.

THEOREM 1.1 (global existence). (i) *Let $\gamma < 0$. Suppose u_0 satisfies the following conditions*:

$$u_0 \in W^{2,1}(\rho,\infty) \quad \textit{for any} \quad \rho \in \boldsymbol{R}$$

and

$$u_0(x) < 0, \quad u_0'(x) > 0, \quad u_0''(x) < 0, \quad \textit{almost all } x \in (-\infty, x_0), \tag{7}$$

$$u_0(x) \equiv 0, \quad x \in [x_0, \infty) \tag{8}$$

for some $x_0 \in \boldsymbol{R}$. Then there exists a solution u of (5) *and* (6) *on $\boldsymbol{R} \times [0,\infty)$ such that*

$$u \in C([0,\infty);\ W^{1,1}(\rho,\infty)) \cap C^1([0,\infty);\ L^1(\rho,\infty)), \tag{9}$$

for any $\rho \in \boldsymbol{R}$.

(ii) *Suppose further that*

$$u_0''' \in W^{1,2}(\rho,\infty) \quad \textit{for any} \quad \rho \in \boldsymbol{R}.$$

Then the solution u also satisfies

$$u_{xx} \in C([0,\infty);\ L^1_{\mathrm{loc}}(\rho,\infty)),$$

$$u_{xxx} \in C([0,\infty);\ L^2(\rho,\infty)) \cap L^\infty(0,\infty;\ W^{1,2}(\rho,\infty))$$

for any $\rho \in \boldsymbol{R}$, and such a u is unique. In particular, $u \in C^2(\boldsymbol{R} \times [0,\infty))$.

Remark. When $a-b=1$ in (1), the assumptions in Theorem 1.1 are complementary to those of [2, Thm. 2.1]. In fact, $\gamma < 0$ means $a > 1$ (see (3)).

Theorem 1.1 is a consequence of a local existence theorem and a priori estimates given below. To state them, we prepare several notations.

We denote by $|u|_{W^{m,p}(\rho,\infty)}$ the $W^{m,p}(\rho,\infty)$-norm of $u=u(x)$. Let $-\infty \le \rho < 0$, $0 < T \le \infty$ and $0 \le L < \infty$. We put

$$L_0 = 1 + \sup_{x\in[\rho,\infty)} u_0(x), \tag{10}$$

$$u_L(x,t) = u(x+Lt,t), \tag{11}$$

and

$$\Omega(\rho,T,L) = \{(x,t) \in \boldsymbol{R}\times[0,\infty);\ t\in[0,T],\ x\in[\rho+Lt,\infty)\}. \tag{12}$$

Now we show a local existence theorem which improves that of [2, Thm. 1.1].

PROPOSITION 1.2 (local existence). (i) *Suppose that*

$$u_0 \in W^{2,1}(\rho,\infty) \quad \textit{for some } \rho\in\boldsymbol{R}.$$

Then there exists $T_0 = T_0(|u_0|_{W^{2,1}(\rho,\infty)}) > 0$ *and a solution* u *of* (5) *and* (6) *in* $\Omega(\rho,T_0,L_0)$ *such that*

$$u(x,t) \le L_0, \qquad (x,t)\in\Omega(\rho,T_0,L_0), \tag{13}$$

$$u_{L_0} \in C([0,T_0];\ W^{1,1}(\rho,\infty)) \cap C^1([0,T_0];\ L^1(\rho,\infty)). \tag{14}$$

(ii) *Furthermore, assume that*

$$u_0''' \in W^{1,2}(\rho,\infty)$$

for the same ρ *as in* (i). *Then the solution* u *also satisfies*

$$\partial_x^2 u_{L_0} \in C([0,T_0];\ L^1_{\mathrm{loc}}[\rho,\infty)),$$

$$\partial_x^3 u_{L_0} \in C([0,T_0];\ L^2(\rho,\infty)) \cap L^\infty(0,T_0;\ W^{1,2}(\rho,\infty)),$$

and such a u *is unique. In particular,* $u\in C^2(\Omega(\rho,T_0,L_0))$.

Remark. Our assumption on data is weaker than that of [2], for we do not require the decay of data at $x=-\infty$.

PROPOSITION 1.3 (a priori estimates). *Let* $\gamma<0$ *and* u_0 *satisfy* (7) *and* (8). *Let* u *be a solution of* (5) *and* (6) *such that*

$$u(x,t) \le L \quad \textit{for}\ \ (x,t)\in\Omega(\rho,T,L), \quad u\in C^2(\Omega(\rho,T,L)),$$

$$u_L \in L^\infty(0,T;\ W^{2,1}(\rho,\infty)), \qquad \partial_x^3 u_L \in L^\infty(0,T;\ W^{1,2}(\rho,\infty))$$

for some $T\in(0,\infty)$ *and* $L\in(0,\infty)$. *Then*

$$|\partial_x^j u_L(t)|_{L^1(\rho,\infty)} \le |u_0^{(j)}|_{L^1(\rho,\infty)}, \quad t\in[0,T], \quad j=0,1,2, \tag{15}$$

and

$$u<0, \quad u_x>0, \quad u_{xx}<0, \quad (x,t)\in\Omega(\rho,T,L)\cap\{x<x_0\}, \tag{16}$$

$$u\equiv 0, \qquad (x,t)\in\Omega(\rho,T,L)\cap\{x\ge x_0\}. \tag{17}$$

2. Proof of Proposition 1.2. First we prove (i). We construct a solution of (5) and (6) in $\Omega(\rho, T_0, L_0)$ using the following iteration scheme:

$$u^0(x,t) = u_0(x), \qquad (x,t) \in \Omega(\rho, T_0, L_0),$$

$$\begin{aligned} &u_t^{k+1} + u^k u_x^{k+1} + \gamma I(u_x^k, u_x^k) = 0 \quad \text{in } \Omega(\rho, T_0, L_0), \quad k \geq 1, \\ &u^{k+1}(x,0) = u_0(x) \quad \text{in } [\rho, \infty). \end{aligned} \tag{18}$$

For simplicity, we donote by $|u|_{m,p}$ the $W^{m,p}(\rho,\infty)$-norm of u and by $\|u\|_{m,p,T}$ the $L^\infty(0,T;\ W^{m,p}(\rho,\infty))$-norm of u.

Now we claim

$$u^k(x,t) \leq L_0, \qquad (x,t) \in \Omega(\rho, T_0, L_0), \tag{19}$$

$$\|u_{L_0}^k\|_{2,1,T_0} \leq M, \tag{20}$$

$$u_{L_0}^k \in C([0,T_0];\ W^{1,1}(\rho,\infty)) \tag{21}$$

for $k = 0, 1, 2, \ldots$, where

$$M = 2|u_0|_{2,1}, \tag{22}$$

$$T_0 = \min\left\{\frac{1}{|\gamma|M^2}, \frac{1}{4\Gamma_1 M}, \frac{1}{M}\log\frac{4}{3}\right\}, \tag{23}$$

and

$$\Gamma_1 = |\gamma|\left(\frac{3}{\beta - 1} + 2\right) + 1.$$

Indeed, (19), (20), and (21) can be shown by induction on k. Here we only show (20). It is obvious for $k = 0$. Suppose (20) holds for k. Let $X^k(\sigma; x, t)$ be the characteristic curve in $\Omega(\rho, T_0, L_0)$ passing x at $\sigma = t$, that is,

$$\begin{aligned} &\frac{\partial X^k}{\partial \sigma}(\sigma; x, t) = u^k(X^k(\sigma; x, t), \sigma), \\ &X^k(t; x, t) = x. \end{aligned} \tag{24}$$

Integrating (18) along the characteristic curve, we get

$$u^{k+1}(x,t) = u_0(X^k(0; x, t)) + \int_0^t f^k(X^k(\sigma; x, t), \sigma)\, d\sigma, \tag{25}$$

where

$$f^k = -\gamma I(u_x^k, u_x^k).$$

Note that (24) means

$$X_x^k(\sigma; x, t) = \exp\left(\int_t^\sigma u_x^k(X^k(\tau; x, t), \tau)\, d\tau\right). \tag{26}$$

Differentiating (25) and utilizing (26), we obtain

$$u_x^{k+1}(x,t) = u_0'(X^k(0;x,t))X_x^k(0;x,t) + \int_0^t f_x^k(X^k(\sigma;x,t),\sigma)X_x^k(\sigma;x,t)d\sigma, \tag{27}$$

and

$$u_{xx}^{k+1}(x,t) = u_0''(X^k(0;x,t))X_x^k(0;x,t)^2 + \int_0^t f_{xx}^k(X^k(\sigma;x,t),\sigma)X_x^k(\sigma;x,t)^2 d\sigma - \int_0^t u_{xx}^k \cdot u_x^{k+1}(X^k(\sigma;x,t),\sigma)X_x^k(\sigma;x,t)^2 d\sigma. \tag{28}$$

Simple computations show the estimates

$$|I(u,v)|_1 \le (\beta-1)^{-1}|u|_1|v|_1, \tag{29}$$

$$|\partial_x I(u,v)|_1 \le (\beta-1)^{-1}(|u_x|_1|v|_1 + |u|_1|v_x|_1), \tag{30}$$

$$|\partial_x^2 I(u,v)|_1 \le 2|u_x|_1|v_x|_1. \tag{31}$$

Evaluating (25), (27), and (28) in the $L^1(\rho,\infty)$-norm and using (29)–(31), we obtain

$$\|u_{L_0}^{k+1}\|_{2,1,T_0} \le (|u_0|_{2,1} + \Gamma_1 M^2 T_0)e^{MT_0} \le M, \tag{32}$$

where the second inequality is due to (22) and (23). This shows (20).

For the convergence of the scheme, set

$$v^k = u^{k+1} - u^k, \qquad k = 0,1,2,\ldots.$$

Then

$$\begin{aligned} v_t^k + u^k v_x^k &= g^k \quad \text{in } \Omega(\rho,T_0,L_0), \\ v^k(x,0) &\equiv 0 \quad \text{in } [\rho,\infty), \end{aligned} \tag{33}$$

where

$$g^k = -u_x^k v^{k-1} + f^k - f^{k-1}.$$

Hence, we obtain as in (25)

$$v^k(x,t) = \int_0^t g^k(X^k(\sigma;x,t),\sigma)d\sigma \tag{34}$$

for $(x,t) \in \Omega(\rho,T_0,L_0)$. Using (20), we have

$$|v_{L_0}^k(t)|_1 \le \int_0^t \Gamma_2 M e^{MT_0}|v_{L_0}^{k-1}(\sigma)|_1 d\sigma \tag{35}$$

for $t \in [0,T_0]$, where

$$\Gamma_2 = |\gamma\beta^{-1} - 1| + |\gamma|(1 + (\beta^{-1} + \beta)(\beta-1)^{-1}).$$

Thus

$$|v^k_{L_0}(t)|_1 \le 2M\frac{(\Gamma_2 M e^{MT_0})^k}{k!}, \qquad k = 0, 1, 2, \dots$$

for $t \in [0, T_0]$. Therefore, $\{u^k_{L_0}\}_{k=0}^{\infty}$ is a Cauchy sequence in $L^\infty(0, T_0;\ L^1(\rho, \infty))$. Furthermore, by the Gagliardo–Nirenberg inequality $|f_x|_1 \le C|f|_1^{1/2}|f_{xx}|_1^{1/2}$ and uniform boundedness of $\{\partial_x^2 u^k_{L_0}\}_{k=0}^{\infty}$ in $L^\infty(0, T_0;\ L^1(\rho, \infty))$, we see that $\{\partial_x u^k_{L_0}\}_{k=0}^{\infty}$ is a Cauchy sequence in $L^\infty(0, T_0;\ L^1(\rho, \infty))$. There exists $u_{L_0} \in C([0, T_0];\ W^{1,1}(\rho, \infty))$ such that

$$u^k_{L_0} \to u_{L_0} \quad \text{in } L^\infty(0, T_0; W^{1,1}(\rho, \infty)) \quad \text{as } k \to \infty.$$

Letting $k \to \infty$ in (18), we see that $u(x, t) = u_{L_0}(x - L_0 t, t)$ is the solution we are seeking.

The proof of (ii) proceeds as before. The main task is to show that $\{u^k_{xxx}\}$ and $\{u^k_{xxxx}\}$ are bounded in $L^\infty(0, T; L^2(\rho, \infty))$ for some positive constant T. We omit the details.

3. Proof of Proposition 1.3. Let

$$\Omega^+(\rho, T, L; x_0) = \Omega(\rho, T, L) \cap \{x \ge x_0\},$$

$$\Omega^-(\rho, T, L; x_0) = \Omega(\rho, T, L) \cap \{x < x_0\}.$$

First we prove

$$u_{xx}(x, t) < 0 \quad \text{in } \Omega^-(\rho, T, L; x_0), \tag{36}$$

$$u(x, t) \equiv 0 \quad \text{in } \Omega^+(\rho, T, L; x_0). \tag{37}$$

The claim (37) is a direct consequence of [2, Cor. 1.2] and this gives (17). In order to prove (36), we utilize the idea of [2, Thm. 2.2 (ii)]. We indicate the proof briefly. Consider the following equation:

$$\begin{aligned} u^\delta_t + u^\delta u^\delta_x + \gamma I(u^\delta_x, u^\delta_x) &= -\delta h(x) \quad \text{in } \Omega(\rho, T, L), \\ u^\delta(x, 0) &= u_0(x) \qquad \text{in } [\rho, \infty), \end{aligned} \tag{38}$$

where δ is a positive small parameter and we choose h so that h is in C^4 and each derivative is bounded, and so that $h(x) \equiv 0$ for $x \ge x_0$ and $h''(x) = \varphi(x)(x_0 - x)^2$ for $x < x_0$, where $\varphi \in C^\infty$ is positive and $\varphi \equiv 1$ near x_0. Let $T' > 0$ be the lifespan of the solution u^δ. From Proposition 1.2 we see that T' is determined uniformly in small δ.

We claim that there exists $x_1(\delta)$ such that

$$u^\delta_{xx} < 0 \tag{39}$$

for $(x, t) \in \{x_1(\delta) \le x < x_0\} \cap \Omega^-(\rho, T', L; x_0)$.

In order to show this, it is necessary to control the nonlocal term. Remark $u^\delta(x, t) \equiv 0$ for $x \ge x_0$. From the uniform boundedness of $\{u^\delta_{xxx}\}$ and $\{u^\delta_{xxxx}\}$ and from the Gagliardo–Nirenberg inequality $|f|_\infty \le C|f|_2^{1/2}|f_x|_2^{1/2}$, we have

$$|u^\delta_{xx}(x, t)| = \left|u^\delta_{xx}(x_0, t) + \int_{x_0}^{x} u^\delta_{xxx}(y, t)dy\right| \le c(x_0 - x).$$

This gives

$$|I(u^\delta_{xx}, u^\delta_{xx})(x,t)| \le c(x_0 - x)^3.$$

Hence

$$|\gamma\beta^{-1}(\beta-1)^2 I(u^\delta_x, u^\delta_x)(x,t)| < \delta h''(x) \tag{40}$$

for x in some small interval $[x_1(\delta), x_0)$ and $t \in [0, T']$. Next let $X^\delta(\sigma; x, t)$ be a characteristic curve passing x at $\sigma = t$, that is,

$$\frac{\partial X^\delta}{\partial \sigma}(\sigma; x, t) = u^\delta(X^\delta(\sigma; x, t), \sigma), \qquad X^\delta(t; x, t) = x.$$

Remark X^δ_x is positive (see (26)). Differentiating (38) twice with respect to x, integrating by parts in the nonlocal term, and integrating along the characteristic curve, we obtain

$$\begin{aligned} u^\delta_{xx}(x,t) = {} & u_0''(X^\delta(0;x,t))X^\delta_x(0;x,t)^{3-\gamma-\gamma/\beta} \\ & + \int_0^t \{\gamma\beta^{-1}(\beta-1)^2 I(u^\delta_{xx}, u^\delta_{xx}) - \delta h''\}(X^\delta(\sigma;x,t),\sigma) \\ & \cdot X^\delta_x(\sigma;x,t)^{3-\gamma-\gamma/\beta} d\sigma. \end{aligned} \tag{41}$$

It follows from (7) and (40) that the right-hand side of (41) is negative. Thus we obtain (39).

On the other hand, since $u_0''(x) < 0$ for $x \in [\rho, x_0)$, and since $u \in C^2(\Omega(\rho, T, L))$, we have that

$$u^\delta_{xx}(x,t) < 0 \quad \text{in } \Omega^-(\rho, t_1(\delta), L; x_1(\delta)) \tag{42}$$

for some $t_1(\delta) \in (0, T']$. Thus from (39) and (42),

$$u^\delta_{xx}(x,t) < 0 \quad \text{in } \Omega^-(\rho, t_1(\delta), L; x_0). \tag{43}$$

We can extend this fact to

$$u^\delta_{xx}(x,t) < 0 \quad \text{in } \Omega^-(\rho, T', L; x_0). \tag{44}$$

In fact, assume $u^\delta_{xx} = 0$ for some point in $\Omega^-(\rho, T', L; x_0) \setminus \Omega^-(\rho, t_1(\delta), L; x_0)$. Let $(\hat{x}(\delta), \hat{t}(\delta))$ be a point with the smallest t-coordinate such that $u^\delta_{xx}(x,t) = 0$. Substitute $(x,t) = (\hat{x}(\delta), \hat{t}(\delta))$ in (41). Then we obtain a contradiction from minimality of $\hat{t}(\delta)$ and the hypothesis of initial data and parameters γ and β. Therefore, we have (44).

As in the proof of Proposition 1.2, we can show that $\{u^\delta_L\}$ is a Cauchy sequence in $L^\infty(0, T'; W^{2,1}(\rho, \infty))$. Thus

$$u^\delta \to u \quad \text{in } L^\infty(0, T';\ W^{2,1}(\rho, \infty)) \quad \text{as } \delta \to 0,$$

which implies

$$u_{xx}(x,t) \le 0 \quad \text{in } \Omega^-(\rho, T', L; x_0). \tag{45}$$

If $u_{xx}(x,t) = 0$ holds at some point of $\Omega^-(\rho, T', L; x_0)$, then we obtain a contradiction from the above method. Therefore,

$$u_{xx}(x,t) < 0 \quad \text{in } \Omega^-(\rho, T', L; x_0). \tag{46}$$

We continue the proof, replacing $u_0(x)$ by $u(x,T')$ as initial data; then we get (36).

Now we show

$$u_0(x) < u(x,t) < 0, \tag{47}$$

$$0 < u_x(x,t) < \frac{u_0'(x)}{1 + (1 - \gamma\beta^{-1})u_0'(x)t}, \tag{48}$$

for $(x,t) \in \Omega^-(\rho, T, L; x_0)$. The second inequality of (47) and the first inequality of (48) are consequences of (36) and (37). Since $u < 0$, $u_x > 0$, it follows from (5) that

$$u_t > 0 \quad \text{in } \Omega^-(\rho, T, L; x_0), \tag{49}$$

which shows the first inequality of (47). For the second inequality of (48), we differentiate (5). After an integration by parts, we have

$$u_{xt} + uu_{xx} + (1 - \gamma\beta^{-1})u_x^2 + \gamma(1 - \beta^{-1})I(u_x, u_{xx}) = 0.$$

Since $u < 0$, $u_x > 0$, $u_{xx} < 0$, $\gamma < 0$, and $\beta > 1$, we obtain

$$u_{xt} < -(1 - \gamma\beta^{-1})u_x^2,$$

which gives the second inequality of (48). Finally we show (15). From (37) and (49) we have

$$\partial_t u_L \geq 0 \quad \text{in } [\rho, \infty) \times [0, T]. \tag{50}$$

Since $u_L \leq 0$, it follows that $|u_L|_1 = -\int_\rho^\infty u_L dx$, and from (50) we obtain

$$\partial_t |u_L(t)|_1 \leq 0, \qquad t \in [0, T]. \tag{51}$$

This shows (15) in $j = 0$. Also from (47) and (48) we compute

$$|\partial_x u_L(t)|_1 = \int_\rho^\infty \partial_x u_L(x,t)dx = -u_L(\rho, t) \leq |u_0|_\infty \leq |u_0'|_1.$$

Similarly, we can prove boundedness of the L^1-norm of $\partial_x^2 u_L$. The proof is completed.

4. Proof of Theorem 1.1. We first prove (ii). Mollifying data, we then prove (i).

(ii) Note $L_0 = 1$. First we construct the solution on $\Omega(-n, n, 1)$ for any natural number n. From Proposition 1.2 (ii), for some T_1 there exists a solution u on $\Omega(-n, T_1, 1)$ satisfying the assumptions of Proposition 1.3. So u satisfies a priori estimates, that is,

$$|\partial_x^j u(t)|_{L^1(\rho,\infty)} \leq |u_0^{(j)}|_{L^1(\rho,\infty)}$$

for $t \in [0, T_1]$ and $j = 0, 1, 2$. Also ,

$$u < 0, \quad u_x > 0, \quad u_{xx} < 0 \quad \text{in } \Omega(-n, T_1, 1) \cap \{x < x_0\},$$

$$u \equiv 0 \quad \text{in } \Omega(-n, T_1, 1) \cap \{x \geq x_0\}.$$

Therefore, $u(\cdot, T_1)$ satisfies a similar condition to u_0. If $T_1 < n$, we solve the equation by Proposition 1.2 (ii) with T_1 as initial time and with $u(\cdot, T_1)$ as initial data. We also denote the solution by u and denote the time of existence by T_2. Then, from a priori estimates, we can take $T_2 = T_1$ (see (23)) and u exists on $\Omega(-n, 2T_1, 1)$. If $2T_1 < n$, then we proceed as above. Thus there exists a solution u on $\Omega(-n, n, 1)$. We denote this function by u_n.

Now we define

$$u(x,t) = u_n(x,t) \quad \text{for } (x,t) \in \Omega(-n, n, 1).$$

This definition is well defined by [2, Cor. 1.2]. Then $u(x,t)$ is the solution on $\boldsymbol{R} \times [0, \infty)$ we have been looking for.

(i) Let $J_\varepsilon *$ be the Friedrichs mollifier. For the properties of $J_\varepsilon *$, see [1, Lemma 2.18]. We put

$$u_0^\varepsilon(x) = (J_\varepsilon * u_0)(x) \quad \text{for } x \in \boldsymbol{R}, \quad \varepsilon \in (0,1).$$

Then

$$u_0^\varepsilon \in W^{2,1}(\rho, \infty), \quad (u_0^\varepsilon)''' \in W^{1,2}(\rho, \infty) \tag{52}$$

for any $\rho \in (-\infty, 0)$. In particular,

$$|(u_0^\varepsilon)^{(j)}|_{L^1(\rho,\infty)} \leq |u_0^{(j)}|_{L^1(\rho-1,\infty)} \tag{53}$$

for $j = 0, 1, 2$, $\varepsilon \in (0,1)$, $\rho \in (-\infty, 0)$. Also,

$$\begin{array}{lll} u_0^\varepsilon < 0, \quad (u_0^\varepsilon)' > 0, \quad (u_0^\varepsilon)'' < 0 & \text{in} & (-\infty, x_0 + \varepsilon), \\ u_0^\varepsilon \equiv 0 & \text{in} & [x_0 + \varepsilon, \infty). \end{array} \tag{54}$$

Since u_0^ε satisfies (52) and (54), or the assumptions of data in (ii), there exists a solution u^ε on $\boldsymbol{R} \times [0, \infty)$. Then from Proposition 1.3 and (53),

$$|\partial_x^j \bar{u}^\varepsilon(t)|_{L^1(\rho,\infty)} \leq |u_0^{(j)}|_{L^1(\rho-1,\infty)} \tag{55}$$

for $j = 0, 1, 2$, where

$$\bar{u}^\varepsilon(x,t) = u^\varepsilon(x+t, t),$$

and

$$\begin{array}{lll} u^\varepsilon < 0, \quad u_x^\varepsilon > 0, \quad u_{xx}^\varepsilon < 0 & \text{in} & (-\infty, x_0 + \varepsilon) \times [0, \infty), \\ u^\varepsilon \equiv 0 & \text{in} & [x_0 + \varepsilon, \infty) \times [0, \infty). \end{array} \tag{56}$$

We claim that $\{\bar{u}^\varepsilon\}_{\varepsilon \in (0,1)}$ is a Cauchy sequence in $L^\infty(0, T;\ W^{1,1}(\rho, \infty))$ for any $\rho \in \boldsymbol{R}$, $T > 0$.

In order to show this, let $X^\varepsilon(\sigma; x, t)$ be the characteristic curve passing x at $\sigma = t$, that is,

$$\frac{\partial X^\varepsilon}{\partial \sigma}(\sigma; x, t) = u^\varepsilon(X^\varepsilon(\sigma; x, t), \sigma), \qquad X^\varepsilon(t; x, t) = x.$$

Put $U^{\varepsilon,\delta} = u^\varepsilon - u^\delta$ and $U_0^{\varepsilon,\delta} = u_0^\varepsilon - u_0^\delta$, then

$$U_t^{\varepsilon,\delta} + u^\varepsilon U_x^{\varepsilon,\delta} + u_x^\delta U^{\varepsilon,\delta} + \gamma I(U_x^{\varepsilon,\delta}, u_x^\varepsilon) + \gamma I(u_x^\delta, U_x^{\varepsilon,\delta}) = 0,$$
$$U^{\varepsilon,\delta}(x,0) = U_0^{\varepsilon,\delta}(x). \tag{57}$$

Integrating (57) along the characteristic curve and estimating in the $W^{1,1}(\rho,\infty)$-norm as in (32), it follows from (55) that

$$\|\bar{U}^{\varepsilon,\delta}\|_{1,1,T} \le |U_0^{\varepsilon,\delta}|_{1,1} \exp(|u_0''|_{L^1(\rho-1,\infty)} T(1 + \Gamma \exp |u_0''|_{L^1(\rho-1,\infty)}))$$

for some positive constant Γ depending on γ and β. Thus we have

$$\|\bar{U}^{\varepsilon,\delta}\|_{1,1,T} \to 0 \quad \text{as } \varepsilon,\ \delta \downarrow 0,$$

since $|U_0^{\varepsilon,\delta}|_{1,1} \to 0$ as $\varepsilon, \delta \downarrow 0$. This shows our claim.

We conclude that there exists $\bar{u} \in L^\infty(0,T;\ W^{1,1}(\rho,\infty))$ such that

$$\bar{u}^\varepsilon \to \bar{u} \quad \text{in } L^\infty(0,T;\ W^{1,1}(\rho,\infty))$$

as $\varepsilon \to 0$ for any $\rho \in \boldsymbol{R}$, $T > 0$. Then we easily see that $u(x,t) = \bar{u}(x-t,t)$ is a solution of (5) and (6) on $\boldsymbol{R} \times [0,\infty)$.

Acknowledgments. The author is grateful to Professor Atsushi Yoshikawa for having suggested the present problem to him.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] R. GARDNER, *Solutions of a nonlocal conservation law arising in combustion theory*, SIAM J. Math. Anal., 18 (1987), pp. 172–183.
[3] A. MAJDA, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Appl. Math. Science Series 53, Springer-Verlag, New York, 1984.
[4] A. MAJDA AND R. ROSALES, *A theory for spontaneous Mach-stem formation in reacting shock fronts. I, The basic perturbation analysis*, SIAM J. Appl. Math., 43 (1983), pp. 1310–1334.
[5] ———, *Resonantly interacting weakly nonlinear hyperbolic waves. I. A single space variable*, Stud. Appl. Math., 71 (1984), pp. 149–179.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1122–1127, July 1994

005

A VARIATIONAL PROBLEM FOR HARMONIC FUNCTIONS IN RING-SHAPED DOMAINS WITH PARTIALLY FREE BOUNDARY*

ANDREA COLESANTI†

Abstract. This paper considers two subsets Ω_0 and Ω of $\mathbb{R}^n$, $n = 2$ or $n = 3$, and two continuous real-valued functions g_0 and g defined on $\partial\Omega$ and $\partial\Omega_0$, respectively. The position of Ω is allowed to vary inside Ω_0 and the author looks for the minimum of the Dirichlet intergral of the function u, which is harmonic in $(\Omega_0 \setminus \Omega)$ and verifies the following boundary conditions: $u = g_0$ on $\partial\Omega_0$, $u = g$ on $\partial\Omega$. Under certain hypotheses on the regularity of $\partial\Omega_0$ and $\partial\Omega$, and on g_0 and g, an existence theorem is proved for the minimizing position of Ω; it is shown through an example that the solution of the considered problem is not unique in general.

Key words. harmonic functions, Dirichlet integral, partially free boundary

AMS subject classifications. 31C05, 35J05, 35R35

1. Introduction.

We study a minimum-value problem for the Dirichlet integral of harmonic functions in ring-shaped domains whose boundaries have given shape and free position.

Let Ω_0 and Ω be two bounded, connected subsets of $\mathbb{R}^n$, $n = 2$ or $n = 3$, with Ω_0 open, Ω closed, and $\Omega \subset \Omega_0$, and let g_0 and g be two given continuous functions defined on $\partial\Omega_0$ and $\partial\Omega$, respectively. Let m be a rigid motion of $\mathbb{R}^n$; we denote by M the set of all m such that $\Omega_m = m(\Omega) \subset \Omega_0$, and we define $g_m = g \circ m^{-1}$, i.e., g_m is the datum g moved from $\partial\Omega$ to $\partial\Omega_m$. As m varies in M, namely, as Ω moves inside Ω_0, we want to minimize the Dirichlet integral of u_m, which is the unique solution of the following boundary-value problem:

$$(1) \qquad \begin{cases} \Delta u_m = 0 & \text{in } \Omega_0 \setminus \Omega_m; \\ u_m = g_0 & \text{on } \partial\Omega_0, \quad u_m = g_m \quad \text{on } \partial\Omega_m. \end{cases}$$

This problem may be regarded as a generalization of the following: to minimize the capacity of an electric capacitor bounded by two plates $\partial\Omega_0$ and $\partial\Omega$, as the inner plate changes its position inside the outer one.

Two problems, in some sense analogous to the one of the present paper, are treated in [5] by Weinberger and Serrin, and in [1] by Alt and Caffarelli.

We assume that the following hypotheses hold.

(H0) $\exists G \in C^2(\Omega_0 \setminus \Omega) \cap C(\overline{\Omega_0 \setminus \Omega})$ such that

$$\int_{\Omega_0 \setminus \Omega} \|\nabla G\|^2 dx < +\infty ;$$

$$G = g_0 \quad \text{on } \partial\Omega_0, \qquad G = g \quad \text{on } \partial\Omega.$$

(H1) $\sup\{g_0(P) : P \in \partial\Omega_0\} < \inf\{g(P) : P \in \partial\Omega\}$.

(H2) $\Omega_0 \setminus \Omega$ satisfies the outer sphere property at each point of its boundary, uniformly in the radius of the sphere.

Our main result is the following.

* Received by the editors March 29, 1993; accepted for publication (in revised form) June 16, 1993.

† Istituto di Analisi Globale e Applicazioni, Via di S.Marta 13/a, 50139 Firenze, Italy.

THEOREM 1. *If* (H0), (H1), *and* (H2) *hold, then there exists* $m_0 \in M$ *such that*

$$\int_{\Omega_0 \setminus \Omega_{m_0}} \|\nabla u_{m_0}\|^2 \, dx \leq \int_{\Omega_0 \setminus \Omega_m} \|\nabla u_m\|^2 \, dx \quad \forall m \in M.$$

We remark that hypothesis (H2) guarantees the existence of a solution for problem (1) and, together with (H1), allows us (see Lemma 1 in §2) to infer that along any minimizing sequence $\{m_k\}$ in M:

$$\liminf_{k \to \infty} dist\ (\partial\Omega_{m_k}, \partial\Omega_0) > 0\,.$$

The proof of Theorem 1 is given in §2. In §3 there are some examples; in particular the first one proves that if we drop hypothesis (H1), the conclusion of Theorem 1 may fail, while the second one shows that the considered problem may have, in general, more than one solution. Furthermore we make some short remarks about possible generalizations of the considered problem in dimension greater than three.

2. Proof of Theorem 1. We prove Theorem 1 in dimension three; the two-dimensional case is analogous. We need the following.

LEMMA 1. *Let* $\{m_k\}$ *be a sequence in* M *such that*

$$\lim_{k \to \infty} m_k = m;$$

and

$$\partial\Omega_m \cap \partial\Omega_0 \neq \emptyset.$$

Then, if d_k *denotes the Dirichlet integral of* u_{m_k}, *we have*

$$\lim_{k \to \infty} d_k = +\infty.$$

Let P be a point of $\partial\Omega_0 \cap \partial\Omega_m$. Let S be a sphere of radius $R > 0$, externally tangent to $\partial\Omega_0$ in P, and π be the tangent plane to S in P. We choose a coordinate system such that P coincides with the origin and $\pi \equiv \{y = 0\}$, where (x, y), $x \in \mathbb{R}^2$, denotes the generic point in $\mathbb{R}^3$. Let $\psi(x)$ and $\phi_k(x)$, for k sufficiently large, be continuous, local representations of $\partial\Omega_0$ and $\partial\Omega_{m_k}$, respectively, defined in a neighborhood B of P in π. As $\Omega_{m_k} \subset \Omega_0$, we assume without loss of generality that

$$\psi(x) < \phi_k(x) \quad \forall x \in B.$$

We now integrate $\|\nabla u_{m_k}\|^2$ over a region smaller than $(\Omega_0 \setminus \Omega_{m_k})$ to get a lower estimate for d_k:

$$d_k \geq \int_B \int_\psi^{\phi_k} \|\nabla u_{m_k}\|^2 dxdy \geq \int_B \int_\psi^{\phi_k} \left(\frac{\partial u_{m_k}}{\partial y}\right)^2 dxdy. \tag{2}$$

As (H1) holds a positive constant N exists, such that

$$u_{m_k}(x, \phi_k(x)) - u_{m_k}(x, \psi(x)) > N \quad \forall x \in B.$$

Hence, by the Hölder inequality and (2), we have

$$d_k \geq N^2 \int_B \frac{1}{\phi_k(x) - \psi(x)} dx.$$

Lastly we observe that for any k, there exists a ball of radius R (we can choose the initial R small enough), internally tangent to $\partial\Omega_{m_k}$ in $m_k(m^{-1}(P))$, whose center $C_k \equiv (x_k, y_k)$ converges to $C \equiv (0, R)$. For some positive ρ we then have

$$\phi_k(x) \leq y_k - \sqrt{R^2 - \|x - x_k\|^2} \quad \text{in } \{\|x\| < \rho\}; \tag{3}$$

$$\psi(x) \geq -\left[R - \sqrt{R^2 - \|x\|^2}\right] \quad \text{in } \{\|x\| < \rho\}. \tag{4}$$

By substituting (3) and (4) in (2), and letting k tend to infinity, we have

$$\lim_{k\to\infty} d_k \geq \frac{N^2}{2} \int_{\|x\|<\rho} \frac{1}{R - \sqrt{R^2 - \|x\|^2}}\, dx.$$

The right-hand side of this inequality is an improper integral whose value is $+\infty$, so the conclusion of the lemma follows.

Proof of Theorem 1. Let μ be the infimum of the Dirichlet integral of u_m as m varies in M; by hypothesis (H0) $\mu < \infty$. Let $\{m_k\}$ be a minimizing sequence:

$$\mu = \lim_{k\to\infty} \int_{\Omega_0 \setminus \Omega_{m_k}} \|\nabla u_{m_k}\|^2\, dx.$$

As Ω_0 and Ω are bounded and as a rigid motion of $\mathbb{R}^3$ depends essentially on six real parameters, M may be regarded as a bounded subset of $\mathbb{R}^6$. This implies that, up to subsequences, m_k converges to a limit m. By Lemma 1, it follows that

$$\partial\Omega_m \cap \partial\Omega_0 = \emptyset;$$

that is, m is still in M. For simplicity, we write u instead of u_m and u_k instead of u_{m_k}.The proof is divided into two steps: first we prove that the sequence $\{u_k\}$ converges uniformly to u in every compact subset of $(\Omega_0 \setminus \Omega_m)$; then we easily deduce that

$$\lim_{k\to\infty} \int_{\Omega_0 \setminus \Omega_{m_k}} \|\nabla u_{m_k}\|^2\, dx \geq \int_{\Omega_0 \setminus \Omega_m} \|\nabla u\|^2\, dx; \tag{5}$$

so that the conclusion of Theorem 1 follows.

Let P be a point on $\partial\Omega_{m_k}$ and S be a sphere with center Q and radius $R > 0$, externally tangent to $\partial(\Omega_0 \setminus \Omega_{m_k})$ in P. As (H1) holds, R may be chosen sufficiently small so that it is independent of P, and it is obviously independent of k. We define

$$h(x) = \frac{1}{R} - \frac{1}{\|x - Q\|}.$$

h is a barrier function for $(\Omega_0 \setminus \Omega_{m_k})$ in P. By the continuity of g_{m_k} and the positivity of h far from P, for any given ε a positive constant $C'_\varepsilon > 0$ exists such that

$$|g_{m_k}(x) - g_{m_k}(P)| < \varepsilon + C'_\varepsilon\, h(x) \quad \forall x \in \partial\Omega_{m_k}.$$

Furthermore, as m is in M, we may assume that

$$dist\big(\partial\Omega_{m_k}, \partial\Omega_0\big) \geq \eta > 0 \quad \forall k;$$

and this implies that for some positive C''_ε,

$$|g_0(x) - g_{m_k}(P)| < \varepsilon + C''_\varepsilon \; h(x) \quad \forall x \in \partial\Omega_0.$$

If $C_\varepsilon = \max(C'_\varepsilon, C''_\varepsilon)$ we have by the maximum principle

$$|u_k(x) - u_k(P)| < \varepsilon + C_\varepsilon h(x) \quad \forall x \in \Omega_0 \setminus \Omega_{m_k}.$$

It is easy to see that C_ε depends only on the modulus of continuity of g, on g_0, and on the quantities R and η, i.e., it is independent of P and k. Thus we have proved that for any positive ε there exists $\delta > 0$, such that

$$\|x - y\| < \delta\,, x \in \Omega_0 \setminus \Omega_{m_k}\,, y \in \partial\Omega_{m_k} \;\Rightarrow\; |u_k(x) - g_{m_k}(y)| < \varepsilon\,;$$

and exactly the same argument as for functions u_k may be repeated for u. On the other hand, as Ω_{m_k} tends to Ω_m, an index k_0 exists such that

$$\|m_k(P) - m(P)\| < \delta \quad \forall P \in \partial\Omega \quad \forall k \geq k_0.$$

By these inequalities and by the maximum principle, it follows that $u - u_k$ is bounded by 2ε in absolute value in $(\Omega_0 \setminus (\Omega_{m_k} \cup \Omega_m))$, as $k > k_0$. This proves that the sequence $\{u_k\}$ converges uniformly to u in every compact subset of $(\Omega_0 \setminus \Omega_m)$ and as the considered functions are all harmonic, the uniform convergence holds for their derivatives also.

Now let σ be a fixed positive number and T be a compact subset of $(\Omega_0 \setminus \Omega_m)$ such that

$$\int_T \|\nabla u\|^2 \, dx > \int_{\Omega_0 \setminus \Omega_m} \|\nabla u\|^2 \, dx - \sigma\,.$$

From a certain k' on we have $T \subset (\Omega_0 \setminus \Omega_{m_k})$, so by the uniform convergence and the definition of μ

$$\mu \geq \lim_{k\to\infty} \int_T \|\nabla u_k\|^2 \, dx = \int_T \|\nabla u\|^2 \, dx \geq \int_{\Omega_0 \setminus \Omega_m} \|\nabla u_m\|^2 \, dx - \sigma\,.$$

As σ is arbitrary, this proves (5), i.e., the theorem.

3. Remarks and examples.

Remark 1. Up to now the outer sphere hypothesis (H2) has been used, explicitly or implicitly, only to prove the following three facts:

(i) the boundary-value problem (1) has a solution;

(ii) at each point of $\partial\big(\Omega_0 \setminus \Omega_m\big)$ a barrier function exists like the one defined in the proof of Theorem 1;

(iii) inequalities (3) and (4) hold in the proof of Lemma 1.

One can easily check that these conditions are still verified in dimension two, if we replace (H2) with the weaker hypothesis of outer cone property.

Remark 2. One may try to generalize the results of this paper for the problem stated in the Introduction, in $\mathbb{R}^n$ with $n > 3$. On the other hand, for n large, the outer sphere property is not sufficient in general, to prove that the Dirichlet integral of u_m is unbounded when $\Omega_m \cap \Omega_0 \neq \emptyset$, and then that the minimizing position of Ω is "strictly inside" Ω_0. Hence to prove an existence theorem like Theorem 1, one would be forced to enlarge the set of admissible positions of Ω, and consequently to consider even weak solutions for the boundary-value problem (1). This kind of study is not the purpose of this paper.

Example 1. This example proves that the minimum of the Dirichlet integral of u_m may be attained for an m such that

$$\partial\Omega_0 \cap \partial\Omega_m \neq \emptyset,$$

if hypothesis (H1) is dropped.

Let Ω_0 and Ω be as in Fig. 1; assume that along the straight-line segments $\overline{PQ}$ and $\overline{P'Q'}$, g and g_0 are equal and that they vary continuously between the values of 0 and 1. By the maximum principle, u_m is constant on each component of $\partial(\Omega_0 \setminus \Omega_m)$ so that its Dirichlet integral is zero and represents an absolute minimum. We observe that for any other position of m such that $\partial\Omega_0 \cap \partial\Omega_m = \emptyset$, the Dirichlet integral of u_m is strictly positive.

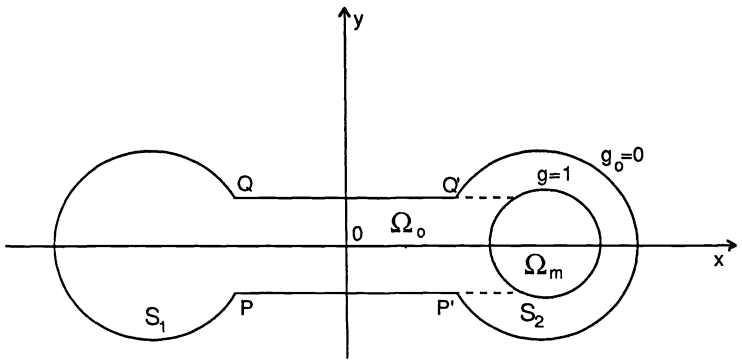


FIG. 1

Example 2. The solution of our minimum-value problem is not unique in general. Let Ω_0 and Ω be as in Fig. 2, with data $g_0 \equiv 0$ and $g \equiv 1$. Let m_0 be such that the Dirichlet integral of u_{m_0} is minimum. The length of the straight-line segment PQ is smaller than the diameter of Ω, so that Ω_{m_0} is not contained in the rectangle $PQP'Q'$. So Ω_{m_0} is in S_1 or in S_2, but then another minimizing position is obtained by reflecting Ω_{m_0} with respect to the y axis.

We observe that by slightly modifying the boundary of Ω_0, one may find another example such that there are at least two distinct minimizing positions, and they are not symmetric with respect to any direction (see [2]).

Example 3. Let Ω_0 and Ω be disks in the two-dimensional plane and $g_0 \equiv 0$, $g \equiv 1$. In this case using a symmetrization argument (see also [3, Thm. 2.31]) one can easily infer that the minimizing position m_0 is unique and it is such that Ω_0 and Ω_{m_0} are concentric. Furthermore, it can be proved (see [2]) that the Dirichlet integral of u_m is a decreasing function of the distance between the centers of Ω_0 and Ω_m.

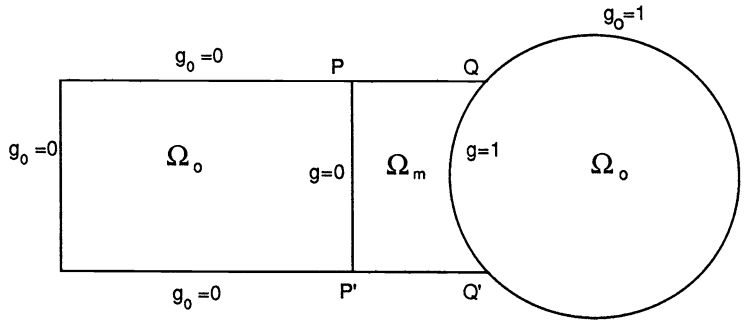


FIG. 2

Aknowledgments. The author wishes to thank Professor Carlo Pucci and Professor Marco Longinetti for their precious contributions to this work.

REFERENCES

[1] J. W. ALT AND L. A. CAFFARELLI, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105–114.

[2] A. COLESANTI, *Un problema di minimo per funzioni armoniche in anelli con frontiera parzialmente libera*, Pubbl. Ist. Anal. Glob. Appl., 33 (1992).

[3] B. KAWOHL, *Rearrangements and Convexity of Level Sets in* P.D.E., Springer-Verlag, New York, 1985.

[4] M. LONGINETTI, *Sulla convessitá delle linee di livello di funzioni armoniche*, Boll. U.M.I., (6) 2-A (1983), pp. 71–75.

[5] H. F. WEINBERGER AND J. B. SERRIN, *Optimal shapes for brittle beams under torsion*, in Complex Analysis and its Applications, Nauka, Moscow, 665 (1968), pp. 88–91.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1128–1156, July 1994

006

THE THERMISTOR PROBLEM: EXISTENCE, SMOOTHNESS, UNIQUENESS, BLOWUP*

S. N. ANTONTSEV† AND M. CHIPOT‡

Abstract. The goal of this paper is to study a nonlinear system modeling the heat diffusion produced by Joule effect in an electric conductor. Existence, uniqueness, smoothness, and blowup in particular are studied.

Key words. parabolic systems, existence, uniqueness, smoothness, blowup

AMS subject classifications. 35K20, 35K35, 35K45, 35K60

1. Introduction.

The heat produced in a conductor by an electric current leads to the so-called thermistor problem, i.e., to the system

$$u_t - \nabla \cdot (\kappa(u)\nabla u) = \sigma(u)|\nabla\varphi|^2 \quad \text{in } \Omega \times (0,T), \tag{1.1a}$$

$$u = 0 \quad \text{on } \Gamma \times (0,T), \quad u(\cdot,0) = u_0, \tag{1.1b}$$

$$\nabla \cdot (\sigma(u)\nabla\varphi) = 0 \quad \text{in } \Omega \times (0,T), \tag{1.1c}$$

$$\varphi = \varphi_0 \quad \text{on } \Gamma \times (0,T). \tag{1.1d}$$

We assume here that Ω is a smooth, bounded open set of $\mathbf{R}^n$, Γ denotes its boundary, T is some positive given number, φ is the electrical potential, u the temperature inside the conductor, $\kappa(u) > 0$ the thermal conductivity, and $\sigma(u) > 0$ the electrical conductivity. The physical situation is when $n = 3$ and Ω is the spatial domain occupied by the body that we consider and which is assumed to be a conductor of both heat and electricity. However, we will consider the general case $n \geq 1$.

If $\mathcal{I}$ denotes the current density and $\mathcal{Q}$ the vector of heat flow then the Ohm law and the Fourier law read, respectively,

$$\mathcal{I} = -\sigma(u)\nabla\varphi, \tag{1.2}$$

$$\mathcal{Q} = -\kappa(u)\nabla u. \tag{1.3}$$

Then equations (1.1a) and (1.1c) follow from the conservation laws

$$\nabla \cdot \mathcal{I} = 0, \qquad \rho c \frac{\partial u}{\partial t} + \nabla \cdot \mathcal{Q} = \mathcal{I} \cdot \mathcal{E}, \tag{1.4}$$

where $\mathcal{E}$ denotes the electric field, ρ the density of the conductor, c its heat capacity (see also [C.1], [C.P.], [H.R.S.], and [Ko]). We assume here that $\rho c \equiv 1$.

Remark 1.1. Due to (1.1c), (1.1a) also reads

$$u_t = \nabla \cdot (\kappa(u)\nabla u + \sigma(u)\varphi\nabla\varphi) \quad \text{in } \Omega \times (0,T).$$

*Received by the editors June 30, 1992; accepted for publication December 4, 1992.

†Universidad de Oviedo, Departamento de Matématicas, Calvo Sotelo, s/n 33007 Oviedo, Spain.

‡Université de Metz, Département de Mathématiques, Ile de Saulcy, 57045 Metz-Cedex 01, France.

The similarity with the two-phase filtration problem should be noticed. Indeed, if u is the concentration and φ the pressure, then the equations of two-phase filtration read

$$u_t = \nabla \cdot (\kappa(u)\nabla u + b(u)\nabla\varphi) \quad \text{in } \Omega \times (0,T),$$
$$\nabla \cdot (\sigma(u)\nabla\varphi) = 0 \quad \text{in } \Omega \times (0,T).$$

We refer the reader to [A.K.M.] for details.

Instead of (1.1b) we will also consider the boundary condition

$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma \times (0,T), \quad u(\cdot,0) = u_0, \tag{1.1b'}$$

where $\partial u/\partial n$ denotes the outward normal derivative of u.

The paper is divided as follows. In §2 we will show existence of a weak solution to (1.1). In §3 we will focus on the question of smoothness. In §4 we will analyze the dependence of the solution with respect to the data and derive uniqueness results. Finally, in the last section we will investigate the issue of global existence or blowup.

We will use standard notation for parabolic problems and we refer to [L.S.U.] for details.

2. Existence of a weak solution. Let V be a subspace of $H^1(\Omega)$ containing $H_0^1(\Omega)$, V' its dual (see, for instance, [B.L.], [D.L.], [J.L.L.], or [G.T.] for the definition and the properties of the Sobolev spaces). Recall first the following well-known result of the theory of linear parabolic equations (see [D.L.], [L.S.U.]).

Assume

$$u_0 \in L^2(\Omega),$$
$$\kappa \in L^\infty(\Omega \times (0,T)), \qquad 0 < \kappa_1 \le \kappa \le \kappa_2 < +\infty, \tag{2.1}$$

where κ_1, κ_2 are two positive constants.

THEOREM 2.1. *If $f \in L^2(0,T;V')$, there exists a unique u such that*

$$u \in L^2(0,T;V) \cap C([0,T]\,;L^2(\Omega)), \qquad u_t \in L^2(0,T;V'), \tag{2.2}$$

$$\left\langle \frac{d}{dt}u, v \right\rangle + \int_\Omega \kappa \nabla u \cdot \nabla v\, dx = \langle f, v\rangle \quad \text{a.e. } t \in (0,T), \quad \forall v \in V, \tag{2.3}$$

$$u(0) = u_0. \tag{2.4}$$

Moreover, we have the estimate

$$\begin{aligned} \frac{1}{2}|u(t)|_2^2 + \kappa_1 \int_0^t \|\nabla u(t)\|_2^2\, dt &\le \frac{1}{2}|u(t)|_2^2 + \int_0^t \int_\Omega \kappa |\nabla u|^2\, dx\, dt \\ &= \frac{1}{2}|u(0)|_2^2 + \int_0^t \langle f, u(t)\rangle\, dt \quad \text{a.e. } t \in (0,T). \end{aligned} \tag{2.5}$$

($\langle\cdot\rangle$ *is the duality bracket between* V', V, $|\cdot|_p$ *the usual* L^p *norm,* $|\nabla u|$ *the Euclidean norm of the gradient of* u.)

We will assume that

$$\varphi_0 \in L^\infty(0,T;H^1(\Omega) \cap L^\infty(\Omega)), \tag{2.6}$$

$$\kappa, \sigma \text{ continuous}, \quad 0 < \kappa_1 \le \kappa \le \kappa_2, \quad 0 < \sigma_1 \le \sigma \le \sigma_2, \tag{2.7}$$

where κ_i, σ_i are positive constants. Then we can prove the following.

THEOREM 2.2. *If* (2.1), (2.6), (2.7) *hold, then there exists a weak solution to* (1.1) *with the boundary conditions* (1.1b) *or* (1.1b′).

Proof. In the case (1.1b) V will be $H_0^1(\Omega)$ and V will be $H^1(\Omega)$ in the case (1.1b′). Choose $w \in L^2(0,T;L^2(\Omega))$; then for almost every $t \in (0,T)$ there exists a unique $\varphi(\cdot,t)$ solution to

$$\nabla \cdot (\sigma(w)\nabla\varphi) = 0 \quad \text{in } \Omega, \qquad \varphi = \varphi_0 \quad \text{on } \Gamma \times (0,T), \tag{2.8}$$

and we have the following.

LEMMA 2.1. *$\varphi \in L^\infty(0,T;H^1(\Omega) \cap L^\infty(\Omega))$ and for almost every t we have*

$$\int_\Omega |\nabla\varphi(x,t)|^2 \, dx \le C(\sigma_1, \sigma_2, \varphi_0), \tag{2.9}$$

where $C(\sigma_1, \sigma_2, \varphi_0)$ denotes a constant depending only on $\sigma_1, \sigma_2, \varphi_0$.

Proof. Assume that we have proved that φ is measurable in t; then from the maximum principle,

$$|\varphi|_\infty \le |\varphi_0|_\infty. \tag{2.10}$$

Moreover, by multiplying the first equation of (2.8) by $\varphi - \varphi_0 \in H_0^1(\Omega)$ we get

$$\int_\Omega \sigma(w)\nabla\varphi \cdot \nabla(\varphi - \varphi_0) = 0;$$

hence

$$\sigma_1 \int_\Omega |\nabla\varphi(x)|^2 dx \le \left| \int_\Omega \sigma(w)\nabla\varphi(x) \cdot \nabla\varphi_0 \, dx \right| \le \sigma_2 \int_\Omega |\nabla\varphi(x)| \, |\nabla\varphi_0(x)| \, dx,$$

which gives the result by the Cauchy–Schwarz inequality.

Let us postpone for the time being the proof of the measurability of φ.

Remark that by (1.1c) the right-hand side of (1.1a) can be written as

$$\sigma(u)|\nabla\varphi|^2 = \nabla \cdot (\sigma(u)\varphi\nabla\varphi). \tag{2.11}$$

It is clear then that

$$\langle \nabla \cdot (\sigma(w)\varphi\nabla\varphi), v \rangle = -\int_\Omega \sigma(w)\varphi\nabla\varphi \cdot \nabla v \, dx \quad \forall v \in V$$

defines an element f of $L^2(0,T;V')$. According to Theorem 2.1 there exists a unique u satisfying (2.2)–(2.4) with $\kappa = \kappa(w)$. Let us consider the map

$$w \to u = F(w). \tag{2.12}$$

This map carries $L^2(0,T;L^2(\Omega))$ into itself. Moreover, by (2.5) we have

$$\frac{1}{2}|u(t)|_2^2 + \kappa_1 \int_0^t \|\nabla u(t)\|_2^2 \, dt \le \frac{1}{2}|u(0)|_2^2 - \int_0^t \int_\Omega \sigma(w)\varphi\nabla\varphi \cdot \nabla u \, dx \, dt. \tag{2.13}$$

It follows, by Cauchy–Schwarz and Young's inequalities, that

$$\begin{aligned}\frac{1}{2}|u(t)|_2^2+\kappa_1\int_0^t\|\nabla u(t)\|_2^2\,dt &\le \frac{1}{2}|u(0)|_2^2+C\int_0^t\|\nabla\varphi(t)\|_2\,\|\nabla u(t)\|_2\,dt\\ &\le \frac{1}{2}|u(0)|_2^2+\frac{\kappa_1}{2}\int_0^t\|\nabla u(t)\|_2^2\,dt\\ &\quad+\frac{C^2}{2\kappa_1}\int_0^t\|\nabla\varphi(t)\|_2^2\,dt;\end{aligned}\tag{2.14}$$

hence

$$|u(t)|_2^2+\int_0^t\|\nabla u(t)\|_2^2\,dt\le C(u_0,T,\kappa_i,\sigma_i,\varphi_0)\,.\tag{2.15}$$

From (2.3) one easily deduces

$$|u_t|_{L^2(0,T;V')}\le C'(u_0,T,\kappa_i,\sigma_i,\varphi_0)\,.\tag{2.16}$$

(Note that f is bounded in $L^2(0,T;V')$ by (2.9), (2.10)). So, provided we take R large enough, $w\to u$ maps the ball B_R of center 0 and radius R in $L^2(0,T;L^2(\Omega))$ into itself. Moreover, since the space

$$\{u\in L^2(0,T;V)\,|\,u_t\in L^2(0,T;V')\}$$

is compactly imbedded in $L^2(0,T;L^2(\Omega))$, this ball will be carried into a relatively compact set by (2.15), (2.16). If we can show that this map is continuous it will be done by the Schauder fixed point theorem. So for that consider a sequence $w_n\in L^2(0,T;L^2(\Omega))$ such that

$$w_n\to w\quad\text{in } B_R.$$

Define as in (2.8), $\varphi_n, f_n=\nabla\cdot(\sigma(w_n)\varphi_n\nabla\varphi_n)$, and $u_n=F(w_n)$. We have to show that

$$u_n\to u=F(w)\quad\text{in } B_R.$$

For that, by subtracting the equation satisfied by u from the one satisfied by u_n, and taking $v=u_n-u$, we get, after integrating in t,

$$\begin{aligned}&\frac{1}{2}|(u_n-u)(t)|_2^2+\kappa_1\int_0^t\|\nabla(u_n-u)(t)\|_2^2\,dt\\ &\le\frac{1}{2}|(u_n-u)(t)|_2^2+\int_0^t\int_\Omega\kappa(w_n)|\nabla(u_n-u)|^2\,dx\,dt\\ &=\int_0^t\int_\Omega(\kappa(w)-\kappa(w_n))\nabla u\cdot\nabla(u_n-u)\,dx\,dt+\int_0^t\langle f_n-f,u_n-u\rangle\,dt\\ &=I_1+I_2.\end{aligned}\tag{2.17}$$

Set

$$I_3 = \frac{\kappa_1}{4} \int_0^t \|\nabla (u_n - u)(t)\|_2^2 \, dt.$$

Then using Young's inequality we get

$$\begin{aligned} |I_1| &= \left| \int_0^t \int_\Omega (\kappa(w) - \kappa(w_n)) \nabla u \cdot \nabla (u_n - u) \, dx \, dt \right| \\ &\le I_3 + \frac{1}{\kappa_1} \int_0^t \|(\kappa(w) - \kappa(w_n)) \nabla u\|_2^2 \, dt, \\ |I_2| &= \left| \int_0^t \int_\Omega \sigma(w_n) \varphi_n \nabla \varphi_n - \sigma(w) \varphi \nabla \varphi \cdot \nabla (u_n - u) \, dx \, dt \right| \\ &\le I_3 + \frac{1}{\kappa_1} \int_0^t \|\sigma(w_n) \varphi_n \nabla \varphi_n - \sigma(w) \varphi \nabla \varphi\|_2^2 \, dt. \end{aligned}$$

Thus, taking into account the definition of I_3, we obtain

$$\begin{aligned} &|(u_n - u)(t)|_2^2 + \int_0^t \|\nabla (u_n - u)(t)\|_2^2 \, dt \\ &\le \frac{1}{\kappa_1} \left[\min\left(\frac{1}{2}, \frac{\kappa_1}{2}\right) \right]^{-1} \left\{ \int_0^T \|(\kappa(w) - \kappa(w_n)) \nabla u\|_2^2 \, dt \right. \\ &\qquad \left. + \int_0^T \|\sigma(w_n) \varphi_n \nabla \varphi_n - \sigma(w) \varphi \nabla \varphi\|_2^2 \, dt \right\}. \end{aligned} \tag{2.18}$$

Since u_n is in a relatively compact set of B_R it is enough to show that u is the only limit point for u_n. Let u' be such a limit point, i.e.,

$$u' = \lim_{n_k \to \infty} u_{n_k} \quad \text{in } B_R;$$

assuming that we have extracted another sequence of n_k that we still denote by n_k we can assume

$$w_{n_k} \to w \quad \text{a.e. in } \Omega \times (0,T). \tag{2.19}$$

Then, since $|\nabla u|^2 \in L^1(\Omega \times (0,T))$ and by (2.19), $|\kappa(w) - \kappa(w_{n_k})|^2 \to 0$ almost everywhere by the Lebesgue theorem we get

$$\int_0^T \|(\kappa(w) - \kappa(w_{n_k})) \nabla u\|_2^2 \, dt = \int_0^T \int_\Omega |(\kappa(w) - \kappa(w_{n_k}))|^2 |\nabla u|^2 \, dx \, dt \to 0.$$

Next, for $n = n_k$ the second integral in the right-hand side of (2.18) reads

$$\begin{aligned} &\int_0^T \|\sigma(w_n) \varphi_n \nabla \varphi_n - \sigma(w) \varphi \nabla \varphi\|_2^2 \, dt \\ &\le \int_0^T \|\sigma(w_n) \varphi_n \nabla \varphi_n - \sigma(w_n) \varphi_n \nabla \varphi\|_2^2 \, dt \\ &\quad + \int_0^T \|\sigma(w_n) \varphi_n \nabla \varphi - \sigma(w_n) \varphi \nabla \varphi\|_2^2 \, dt \\ &\quad + \int_0^T \|\sigma(w_n) \varphi_\nabla \varphi_- \sigma(w) \varphi \nabla \varphi\|_2^2 \, dt \\ &= I + II + III. \end{aligned}$$

Clearly,

$$I \leq C \int_0^T \int_\Omega |\nabla(\varphi_n - \varphi)|^2 \, dt \, dx,$$

$$II \leq C \int_0^T \int_\Omega |\varphi_n - \varphi|^2 |\nabla\varphi|^2 \, dt \, dx,$$

$$III \leq C \int_0^T \int_\Omega |\sigma(w_n) - \sigma(w)|^2 |\nabla\varphi|^2 \, dt \, dx.$$

By (2.9), (2.19), and from the Lebesgue theorem we can obtain $III \to 0$. Next, φ_n satisfies

$$\nabla \cdot (\sigma(w_n) \nabla\varphi_n) = 0, \qquad \varphi_n = \varphi_0 \quad \text{on } \Gamma.$$

Hence,

$$\int_\Omega \sigma(w_n) \nabla\varphi_n \cdot \nabla(\varphi_n - \varphi) \, dx = \int_\Omega \sigma(w) \nabla\varphi \cdot \nabla(\varphi_n - \varphi) \, dx$$

and

$$\int_\Omega \sigma(w_n) |\nabla(\varphi_n - \varphi)|^2 \, dx = \int_\Omega (\sigma(w) - \sigma(w_n)) \nabla\varphi \cdot \nabla(\varphi_n - \varphi) \, dx,$$

which implies

$$\int_\Omega |\nabla(\varphi_n - \varphi)|^2 \, dx \leq C \int_\Omega |\sigma(w) - \sigma(w_n)|^2 |\nabla\varphi|^2 \, dx. \tag{2.20}$$

Thus,

$$I \leq C \int_0^T \int_\Omega |\nabla(\varphi_n - \varphi)|^2 \, dx \leq C \int_0^T \int_\Omega |\sigma(w) - \sigma(w_n)|^2 |\nabla\varphi|^2 \, dx \to 0$$

as above for *III*. By the Poincaré inequality this implies

$$\int_0^T \int_\Omega |\varphi_n - \varphi|^2 \, dx \to 0,$$

and up to an extracted subsequence we can assume

$$\varphi_n - \varphi \to 0 \quad \text{a.e. on } \Omega \times (0,T);$$

then the Lebesgue convergence theorem gives $II \to 0$ and $u_n \to u = u'$ in $L^2(0,T; L^2(\Omega))$. This completes the proof.

Proof of the measurability of φ. We want to show that φ is measurable in t with values in $H^1(\Omega)$. First remark that if $w \in C([0,T] \times \bar{\Omega})$, then $\varphi \in C([0,T], H^1(\Omega))$. Indeed

$$\nabla \cdot (\sigma(w(t)) \nabla\varphi(t)) = \nabla \cdot (\sigma(w(t')) \nabla\varphi(t')) = 0.$$

Hence,

$$\int_\Omega \sigma(w(t)) |\nabla(\varphi(t) - \varphi(t'))|^2 \, dx = \int_\Omega \sigma(w(t')) - \sigma(w(t)) \nabla\varphi(t') \cdot \nabla(\varphi(t) - \varphi(t')) \, dx$$

and

$$\int_\Omega |\nabla(\varphi(t)-\varphi(t'))|^2\,dx \le C\int_\Omega |\sigma(w(t'))-\sigma(w(t))|^2\,|\nabla\varphi(t')|^2\,dx \to 0$$

when $t \to t'$ by the Lebesgue theorem. Now if $w \in L^2(0,T;L^2(\Omega))$, there exists w_n in $C([0,T]\times\bar{\Omega})$ such that $w_n \to w$ in $L^2(0,T;L^2(\Omega))$, and also almost everywhere on $\Omega\times[0,T]$. From (2.20) we deduce that

$$\int_\Omega |\nabla(\varphi_n-\varphi)|^2\,dx \to 0,$$

and thus since φ_n is measurable so does φ.

3. Smoothness of weak solutions. Existence of classical solutions. In this section we will assume that (2.7) holds and that

$$|\kappa|_{C^{1+\alpha}(\mathbf{R})},\ |\sigma|_{C^{1+\alpha}(\mathbf{R})} \le K, \qquad 0<\alpha<1, \tag{3.1}$$

where K is some constant. Recall that $C^{1+\alpha}(\mathbf{R})$ denotes the space of C^1 functions with derivatives Hölder continuous of order α, $|\cdot|_{C^{1+\alpha}(\mathbf{R})}$ the usual norm on this space. Ω_t will denote the set $\Omega_t=\Omega\times(0,t)$ and $|\cdot|_{q,r,\Omega_T}$ the usual norm on $L^r(0,T;L^q(\Omega))$ (see [L.S.U.]).

THEOREM 3.1. *Let $w=(u,\varphi)$ be any weak solution of the problem* (1.1) *with the boundary condition* (1.1b) *or* (1.1b′) *such that*

$$|\varphi|_{\infty,\Omega_T} + \|\nabla\varphi\|_{q,r,\Omega_T} \le M_0, \tag{3.2}$$

where (see [L.S.U.])

$$0<\chi<1, \quad q\in\left[\frac{n}{1-\chi},+\infty\right], \quad r\in\left[\frac{2}{1-\chi},+\infty\right], \quad \frac{2}{r}+\frac{n}{q}=1-\chi.$$

Then

$$w\in C^{2+\alpha,1+(\alpha/2)}(\Omega_T'), \qquad \bar{\Omega}_T'\subset\Omega_T$$

and

$$|w|_{C^{2+\alpha,1+(\alpha/2)}(\Omega_T')} \le C\left(M_0,\operatorname{dist}(\Omega_T\setminus\Omega_T'),\ |u|_{2,\Omega_T}\right). \tag{3.3}$$

If in addition to (3.1), (3.2) *we have*

$$|u_0|_{C^{2+\alpha}(\bar{\Omega})} + |\varphi_0|_{C^{2+\alpha,1+(\alpha/2)}(\Gamma_T)} = H < +\infty, \tag{3.4}$$

$$u_0|_\Gamma = 0 \quad \textit{for (1.1b)} \quad \textit{or} \quad \left.\frac{\partial u_0}{\partial n}\right|_\Gamma = 0 \quad \textit{for (1.1b′)}, \quad \Gamma_T=\Gamma\times(0,T),$$

then

$$|w|_{C^{2+\alpha,1+(\alpha/2)}(\bar{\Omega}_T)} \le C(M_0,H). \tag{3.5}$$

Proof. The ingredients are well known results of the linear theory of equations of elliptic or parabolic types (see [L.S.U.], [L.U.]). In the formulae below α will be a number between 0 and 1 that may differ from one formula to another.

Step 1. Consider u the solution to the equation

$$\frac{\partial u}{\partial t} = \nabla \cdot (\kappa(u)\nabla u + G), \qquad G = \sigma\varphi\nabla\varphi \in L^q(0,T;L^r(\Omega));$$

then we have

$$|u|_{C^{\alpha,\alpha/2}(\Omega'_T)} \leq C_1\left(M_0,\ \text{dist}\,(\Omega_T \setminus \Omega'_T),\ |u|_{2,\Omega_T}\right), \tag{3.6}$$

and in the case where (3.4) holds,

$$|u|_{C^{\alpha,\alpha/2}(\bar{\Omega}_T)} \leq \bar{C}_1(M_0,H) \tag{3.7}$$

for some $1 > \alpha = \alpha(q,r) > 0$ with $\alpha(q,r) \to 1$ when $(q,r) \to +\infty$.

Step 2. We have $\sigma(u(\cdot,t)) \in C^\alpha(\Omega)$. Then consider φ the solution to the elliptic equation

$$\nabla \cdot (\sigma(u(x,t))\nabla\varphi) = 0, \qquad \varphi|_\Gamma = \varphi_0.$$

Here t is some parameter and the estimates are not depending on t. We have

$$\sup_{t\leq T} |\varphi|_{C^{1+\alpha}(\Omega')} \leq C_2(C_1,\ \text{dist}\ (\Omega \setminus \Omega'), M_0), \tag{3.8}$$

respectively, in the case (3.4):

$$\sup_{t\leq T} |\varphi|_{C^{1+\alpha}(\bar{\Omega})} \leq \bar{C}_2\left(\bar{C}_1, H\right). \tag{3.9}$$

Step 3. From (3.8) and (3.9) we now have

$$G = \sigma\varphi\nabla\varphi \in L^\infty(0,T;L^p(\Omega')) \subset L^p(\Omega'_T), \tag{3.10}$$

respectively,

$$G = \sigma\varphi\nabla\varphi \in L^\infty(0,T;L^p(\Omega)) \subset L^p(\Omega_T) \tag{3.11}$$

for any $p, 1 < p < +\infty$, with

$$|G|_{p,\Omega'_T} \leq C_3(C_1,C_2) = C_3\left(M_0,\ \text{dist}\,(\Omega_T \setminus \Omega'_T),\ |u|_{2,\Omega_T},\ p\right) \quad \forall p > 1 \tag{3.12}$$

and in case (3.4):

$$|G|_{p,\Omega_T} \leq \bar{C}_3(C_1,C_2,p,H). \tag{3.13}$$

Moreover, (3.6), (3.7) are valid for any $0 < \alpha < 1$ if p is large enough. At the same time we have also

$$|\nabla u|_{C^{\alpha,\alpha/2}(\Omega'_T)} \leq C_4(C_3), \tag{3.14}$$

respectively,

$$|\nabla u|_{C^{\alpha,\alpha/2}(\Omega_T)} \leq \bar{C}_4\left(\bar{C}_3\right), \tag{3.15}$$

if p is large enough.

Step 4. Consider then the linear elliptic problem

$$\Delta\varphi = -\frac{\sigma'}{\sigma}\nabla u\cdot\nabla\varphi = g\in C^{\alpha}(\Omega), \quad \varphi|_{\Gamma}=\varphi_0.$$

From this equation we deduce

$$\sup_{t\le T}|\varphi|_{C^{2+\alpha}(\Omega')}\le C_5(C_4,M_0), \tag{3.16}$$

respectively,

$$\sup_{t\le T}|\varphi|_{C^{2+\alpha}(\bar{\Omega})}\le \bar{C}_5\left(\bar{C}_4,M_0\right). \tag{3.17}$$

We would like now to show that

$$\nabla\varphi,\varphi_t\in C^{\alpha,\alpha/2}(\Omega_T).$$

Recall that $\nabla\varphi\in C^{\alpha}(\Omega)$ by (3.8) and (3.9). Introduce the function

$$\varphi^{\tau}=\frac{\varphi(x,t+\tau)-\varphi(x,t)}{\tau^{\alpha}}\quad \forall\tau>0.$$

Then φ^{τ} is a solution to the following elliptic problem:

$$\nabla\cdot(\sigma(u(x,t+\tau))\nabla\varphi^{\tau}+\sigma^{\tau}\nabla\varphi(x,t))=0, \qquad \varphi^{\tau}|_{\Gamma}=\varphi_0^{\tau} \tag{3.18}$$

with an obvious notation for σ^{τ}. From (3.6), (3.7), (3.14), and (3.15) we have that

$$\sigma(u(\cdot,t))\in C^{\alpha}(\Omega), \qquad g(\cdot,t)=\sigma^{\tau}(u(\cdot,t))\nabla\varphi(\cdot,t)\in C^{\alpha}(\Omega),$$

and consequently,

$$\sup_{t\le T}|\nabla\varphi^{\tau}|_{C^{\alpha}(\Omega')}\le C_6(C_1,C_4), \tag{3.19}$$

or in the case of (3.4),

$$\sup_{t\le T}|\nabla\varphi^{\tau}|_{C^{\alpha}(\Omega)}\le \bar{C}_6\left(\bar{C}_1,\bar{C}_4\right). \tag{3.20}$$

Hence,

$$\nabla\varphi\in C^{\alpha,\alpha}(\Omega_T)\subset C^{\alpha,\alpha/2}(\Omega_T)$$

and from the equation in u:

$$u_t-\kappa(u)\Delta u=\sigma(u(x,t))|\nabla\varphi|^2+\kappa'(u)|\nabla u|^2; \tag{3.21}$$

we deduce

$$|u|_{C^{2+\alpha,1+(\alpha/2)}(\Omega'_T)}\le C_7(C_1,C_6), \tag{3.22}$$

respectively,

$$|u|_{C^{2+\alpha,1+(\alpha/2)}(\overline{\Omega_T})} \leq \bar{C}_7\left(\bar{C}_1, \bar{C}_6, H\right). \tag{3.23}$$

We are now able to prove that

$$\varphi_t \in C^{\alpha,\alpha/2}\left(\Omega'_T\right) \qquad \left(\text{respectively}, C^{\alpha,\alpha/2}\left(\overline{\Omega_T}\right)\right).$$

For this remark that

$$\nabla \cdot (\sigma \nabla \varphi_t + \sigma_t \nabla \varphi) = 0, \qquad \varphi_t|_\Gamma = \varphi_{0t}. \tag{3.24}$$

From this equation we derive

$$\nabla \varphi_t \in C^{\alpha}\left(\Omega'\right) \qquad \left(\text{respectively}, C^{\alpha}\left(\bar{\Omega}\right)\right) \tag{3.25}$$

and

$$\varphi_t \in C^{2+\alpha}\left(\Omega'\right) \qquad \left(\text{respectively}, C^{2+\alpha}\left(\bar{\Omega}\right)\right)$$

with

$$\sup_{t\leq T} |\varphi_t|_{C^{2+\alpha}(\Omega')} \leq C_8(C_6, C_7) \qquad \left(\text{respectively}, \sup_{t\leq T} |\varphi_t|_{C^{2+\alpha}(\bar{\Omega})} \leq \bar{C}_8(\bar{C}_6, \bar{C}_7)\right). \tag{3.26}$$

Next, we introduce the function

$$\varphi_t^\tau = \frac{\varphi_t(x, t+\tau) - \varphi_t(x, t)}{\tau^\alpha}.$$

For φ_t^τ we get the equation

$$\nabla \cdot \left(\sigma(x, t+\tau) \nabla \varphi_t^\tau + Q\right) = 0, \qquad \varphi_t^\tau|_\Gamma = \varphi_{0t}^\tau \tag{3.27}$$

with

$$Q = \sigma^\tau \nabla \varphi_t^\tau + \sigma_t^\tau \nabla \varphi(x, t+\tau) + \sigma_t \nabla \varphi^\tau.$$

We have

$$\sup_{t\leq T} |Q|_{C^{\alpha}(\Omega')} \leq C_9(C_6, C_7, C_8) \qquad \left(\text{respectively}, \sup_{t\leq T} |Q|_{C^{\alpha}(\bar{\Omega})} \leq \bar{C}_9(\bar{C}_6, \bar{C}_7, \bar{C}_8)\right)$$

from which it follows that

$$|\varphi_t^\tau|_{C^{1+\alpha}(\Omega')} \leq C_{10}(C_9) \quad \forall \tau > 0 \quad \left(\text{respectively}, |\varphi_t^\tau|_{C^{1+\alpha}(\bar{\Omega})} \leq \bar{C}_{10}(\bar{C}_9) \quad \forall \tau > 0\right)$$

or

$$|\nabla \varphi_t^\tau|_{C^{\alpha,\alpha/2}(\Omega'_T)} \leq C_{11}\left(C_{10}\right), \tag{3.28}$$

or in the case where (3.4) holds,

$$|\nabla \varphi_t^\tau|_{C^{\alpha,\alpha/2}(\Omega_T)} \leq \bar{C}_{11}\left(\bar{C}_{10}\right). \tag{3.29}$$

This completes the proof.

Remark 3.1. Recall that for any weak solution of the linear elliptic problem

$$\nabla \cdot (\sigma \nabla \varphi) = 0, \qquad \varphi|_\Gamma = \varphi_0 \tag{3.30}$$

we have

$$|\varphi|_\infty \leq |\varphi_0|_\infty, \qquad \|\nabla \varphi\|_{p,\Omega} \leq C(p) \|\nabla \varphi_0\|_{p,\Omega}.$$

Here $p = p(\tau), \tau = \sigma_1/(\sigma_2 - \sigma_1)$ is a given function such that

$$2 < p(\tau), \quad 0 < \tau < \infty, \quad p(\tau) \to +\infty \quad \text{when } \tau \to +\infty, \tag{3.31}$$

which is nondecreasing with τ (recall that $\sigma_1 \leq \sigma \leq \sigma_2$).

In the two-dimensional case, i.e., when $n = 2$ the assumptions of Theorem 3.2 are fulfilled for $r = \infty, q = p > 2 = n$. Thus, in this case any weak solution to (1.1) is a smooth classical solution in Ω_T (of course, if $\sigma \in C^{1+\alpha}$) and extends smoothly up to the boundary if $u_0 \in C^{2+\alpha}(\bar{\Omega}), \varphi_0 \in C^{2+\alpha,1+(\alpha/2)}(\Gamma)$.

For $n > 2$ the above argument is valid only if σ has a small oscillation in such a way that

$$n < p\left(\frac{\sigma_1}{\sigma_2 - \sigma_1}\right). \tag{3.32}$$

Remark 3.2. To complete Theorem 2.1, the situation regarding existence of a classical solution is the following:

(1) If $n = 2$ for arbitrary smooth σ and any t;

(2) If $n > 2$ for smooth σ with small oscillations and any t;

(3) If $n > 2$ for u_0 with a small oscillation and t small;

(4) If $n > 2$ for t small (σ, u_0 arbitrary) then (1.1) has a classical solution.

Situations (1) and (2) are clear. To show (3), assume that (2.7), (3.1), and (3.4) hold. Moreover, denote by M a small constant such that

$$-\frac{M}{4} \leq u_0(x) \leq \frac{M}{4} \tag{3.33}$$

and

$$n < p\left(\frac{\sigma_1^M}{\sigma_2^M - \sigma_1^M}\right) \tag{3.34}$$

with

$$\sigma_1^M = \min_{[-M,+M]} \sigma, \quad \sigma_2^M = \max_{[-M,+M]} \sigma, \quad \sigma_2^M - \sigma_1^M = \operatorname*{osc}_{[-M,+M]} \sigma,$$

$p(\tau)$ being the function of Remark 3.1, osc denoting the oscillation. Define then a function σ^M by

$$\sigma^M(\tau) = \begin{cases} \sigma(\tau) & \text{if } |\tau| \leq M/2, \\ \sigma(M) & \text{if } \tau \geq M, \\ \sigma(-M) & \text{if } \tau \leq -M, \end{cases} \tag{3.35}$$

and such that

$$\sigma^M \in C^{1+\alpha}, \quad \operatorname*{osc}_{\mathbf{R}} \sigma^M = \sigma_2^M - \sigma_1^M = \operatorname*{osc}_{[-M,+M]} \sigma.$$

Then it is clear that (1.1) corresponding to σ^M has a classical solution (u, φ) for all $t \leq T$. Then choose t_0 such that

$$|u(x,t) - u_0(x)| \leq \frac{M}{4} \quad \text{or} \quad -\frac{M}{2} \leq u(x,t) \leq \frac{M}{2} \quad \text{for } t \leq t_0.$$

We have for $t \leq t_0$,

$$\sigma^M (u(x,t)) = \sigma (u(x,t));$$

hence $u(x,t)$ is a classical solution to (1.1) for $t \leq t_0$.

To see (4), introduce the function

$$\sigma^\varepsilon (u,x) = \begin{cases} \sigma(u) & \text{if } |u - u_0(x)| \leq \varepsilon/2, \\ \sigma(u_0(x) - \varepsilon) & \text{if } u \leq u_0(x) - \varepsilon, \quad x \in \Omega, \\ \sigma(u_0(x) + \varepsilon) & \text{if } u_0(x) + \varepsilon \leq u, \end{cases}$$

which is defined for $x \in \Omega, \varepsilon/2 < |u - u_0(x)| < \varepsilon$ so that

$$\sigma^\varepsilon (u,x) \in C^{1+\alpha}(\mathbf{R} \times \Omega), \qquad \operatorname*{osc}_{\mathbf{R}\times\Omega} \sigma^\varepsilon (u,x) = \operatorname*{osc}_{[|u-u_0(x)|\leq\varepsilon, x\in\Omega]} \sigma^\varepsilon (u,x).$$

Clearly, $\sigma^\varepsilon(u,x) \to \sigma(u_0(x))$ when $\varepsilon \to 0$. We select ε small enough such that if

$$\lambda^\varepsilon = \frac{\sigma^\varepsilon (u,x)}{\sigma(u_0(x))}, \qquad \tau = \min_{\mathbf{R}\times\Omega} \frac{\lambda^\varepsilon}{(\max \lambda^\varepsilon - \min \lambda^\varepsilon)},$$

we have

$$n < p(\tau).$$

Consider now the problem (1.1), where $\sigma(u)$ is replaced by $\sigma^\varepsilon(u,x)$. The equation for φ reads

$$\nabla \cdot (\sigma^\varepsilon \nabla \varphi) = 0.$$

By Theorem 2.1, there exists a weak solution to problem (1.1) corresponding to $\sigma = \sigma^\varepsilon(u,x)$. Let us show that this solution is in fact classical. Introduce $v = \sigma(u_0(x))\varphi(x,t)$. Then, v satisfy

$$\nabla \cdot [\lambda^\varepsilon (\nabla v - v \nabla \ \ln \ \sigma(u_0(x)))] = 0.$$

Note that $\lambda^\varepsilon \to 1$ when $\varepsilon \to 0$. According to the fact that $n < p(\tau)$ and (2) we have

$$\nabla v \in L^\infty (0,T; L^p(\Omega)), \qquad p > n,$$

and thus

$$\nabla \varphi = \frac{1}{\sigma(u_0(x))} (\nabla v - \varphi \nabla \sigma(u_0(x))) \in L^\infty (0,T; L^p(\Omega)).$$

Then, by Theorem 3.1, we have

$$u \in C^{2+\alpha,1+(\alpha/2)}\left(\bar{\Omega}_T\right).$$

Hence

$$|u(x,t) - u_0(x,t)| \leq C(\varepsilon)\, t.$$

Selecting t such that $C(\varepsilon)t < \varepsilon/2$ we have $\sigma^\varepsilon(u,x) = \sigma(u)$, and thus the existence of a classical solution for small t is established.

Remark 3.3. So we have existence of a classical solution to (1.1) for small t. To extend this solution for all $t \leq T$ we need estimates for $t \leq T$. According to Theorem 3.1 the estimate (see (3.2))

$$\|\nabla\varphi\|_{q,r,\Omega_T} \leq M, \qquad \frac{2}{r} + \frac{n}{q} = 1 - \chi$$

is enough. We are now going to establish this estimate for

$$r = q = \frac{2+n}{1-\chi} > 2 + n.$$

Indeed we have the following.

THEOREM 3.2. *Let (u,φ) be a classical solution to the problem* (1.1) *and assume that*

$$0 < \sigma_1 \leq \sigma \leq \sigma_2 < +\infty, \qquad |\sigma'| \leq K \tag{3.36}$$

$$\sup_{0\leq t\leq T}\left(|\varphi_0|_{C^\alpha(\bar{\Omega})}; \|\nabla\varphi_0\|_{p,\Omega}\right) \leq M, \qquad p > 2; \tag{3.37}$$

then for $2s + 2 > n$ and any $T < +\infty$,

$$\|\nabla u\|_{2s+2,\Omega_T} + \|\nabla\varphi\|_{2s+2,\Omega_T} \leq C\left\{\|u_0\|^{(1)}_{2s+2,\Omega_T} + \|\varphi_0\|^{(1)}_{2s+2,\Omega_T}\right\}, \tag{3.38}$$

where $C = c(s,n,T,\Omega,p,K,\sigma_i,M)$, and

$$\|f\|^{(1)}_{k,\Omega_T} = |f|_{k,\Omega_T} + \|\nabla f\|_{k,\Omega_T}, \qquad |f|_{k,\Omega_T} = |f|_{k,k,\Omega_T}.$$

Proof. The proof goes through several steps. The scheme is the following.

Step 1. Considering t as a parameter we derive local estimates inside Ω for any $t \leq T$ for the solution to the problem

$$\nabla\cdot(\sigma(u)\nabla\varphi) = 0, \qquad \varphi|_\Gamma = \varphi_0. \tag{3.39}$$

Step 2. We derive local estimates for the solution u to the problem

$$u_t - \nabla\cdot(\kappa(u)\nabla u) = \nabla\cdot(\sigma(u)\varphi\nabla\varphi) = \sigma(u)|\nabla\varphi|^2, \tag{3.40}$$

where $(\sigma(u)\varphi\nabla\varphi) = \sigma(u)|\nabla\varphi|^2$ is considered as a given function of x and t.

Step 3. We deduce global estimates for (u,φ).

Let us first go through Step 1.

Step 1. Let us denote by $\xi_k, k = 1, \dots, m$ smooth functions such that

$$\sum_1^m \xi_k^{2s+2} = 1, \quad x \in \bar{\Omega}, \quad 0 \le \xi_k(x), \quad |\xi, \nabla\xi, \nabla^2\xi| \le 1,$$

the diameter of their support being smaller than some small number that we will choose later on.

First we have the following.

LEMMA 3.1. *Suppose that φ is a classical solution to*

$$\nabla \cdot (\sigma(u)\nabla\varphi) = 0, \quad x \in \Omega, \quad \varphi|_\Gamma = \varphi_0$$

and that $\xi_k(x)$ is a smooth function such that $(\varphi - \varphi_0)\xi_k(x)$ vanishes outside of the domain $\Omega_k \subset \bar{\Omega}$. Then, for any $2s + 2 > n$ and $t \le T$,

(3.41)

$$\int_\Omega |\nabla\varphi|^{2s+2} \cdot \xi_k^{2s+2}\, dx \le C\left\{\int_\Omega \left(\delta^{2s+2}|\nabla u|^{2s+2} + |\nabla\varphi_0|^{2s+2}\right) \cdot \xi_k^{2s+2}\, dx + I_1(k)\right\}.$$

In the above inequality $C = C(s, n, k), \delta = \operatorname{osc}_{\Omega_k}(\varphi - \varphi_0)$,

(3.42)

$$I_1(k) = \int_{\Omega_k} \left[\delta^{2s+2}|\nabla\xi_k|^{2s+2} + \delta^{s+1}|\varphi - \varphi_0|^{s+1}\left(|\nabla\xi_k|^{2s+2} + \xi_k^{s+1}|\nabla^2\xi_k|^{s+1}\right)\right] dx,$$

φ_0 being the function such that

$$\Delta\varphi_0 = 0, \qquad \varphi_0|_\Gamma = \varphi_0.$$

In particular, if

$$\xi_k \equiv 1,$$

then

$$\int_\Omega |\nabla\varphi|^{2s+2}\, dx \le C\left\{\int_\Omega |\nabla\varphi_0|^{2s+2} + |\nabla u|^{2s+2}\right\} dx, \tag{3.43}$$

where $C = C(s, n, k, |\varphi_0|_\infty)$.

Proof. The proof is similar to the one in [A.K.M., p. 254]. It is based on [L.S.U., p. 94, form. 5.8]:

(3.44)

$$\begin{aligned}\int_\Omega |\nabla(\varphi - \varphi_0)|^{2s+2} \cdot \xi_k^{2s+2}\, dx \\ \le 16\left(\operatorname*{osc}_{\Omega_k}(\varphi - \varphi_0)\right)^2 \left\{\int_\Omega c^2 |\nabla(\varphi - \varphi_0)|^{2s-2} |\nabla^2(\varphi - \varphi_0)|^2 \xi_k^{2s+2}\right. \\ \left. + |\nabla(\varphi - \varphi_0)|^{2s} \cdot |\nabla\xi_k|^2 \xi_k^{2s}(s+1)^2\, dx\right\}, \quad c^2 = n^2 + s^2.\end{aligned}$$

From this inequality by Young's inequality and local estimates of $\nabla^2\varphi$ in terms of $\Delta\varphi$ we deduce

$$\int_\Omega |\nabla\varphi|^{2s+2}\cdot\xi_k^{2s+2}\,dx \le C\left[\delta^{s-1}\int_\Omega |\Delta\varphi|^{s+1}\cdot\xi_k^{2s+2}\,dx + \int_\Omega |\nabla\varphi_0|^{2s+2}\cdot\xi_k^{2s+2}\,dx + I_1(k)\right]. \tag{3.45}$$

We now use the elliptic equation

$$\Delta\varphi = -\frac{\sigma'}{\sigma}\,\nabla u\cdot\nabla\varphi$$

to deduce

$$\int_{\Omega_k} |\Delta\varphi|^{2s+2}\cdot\xi_k^{2s+2}\,dx \le C'\int_{\Omega_k}\left[\frac{1}{2\varepsilon}|\nabla\varphi|^{2s+2} + \frac{\varepsilon}{2}|\nabla u|^{2s+2}\right]\cdot\xi_k^{2s+2}\,dx \tag{3.46}$$

for some constant C'. Substituting (3.46) in the right-hand side of (3.45) with $\varepsilon = CC'\delta^{s+1}$ we obtain (3.41).

Step 2. Next we have the following.

LEMMA 3.2. *Let $u(x,t)$ be a classical solution to*

$$u_t = \nabla\cdot(\kappa(u)\nabla u + \sigma\varphi\nabla\varphi),\quad u(0) = u_0,\quad u|_\Gamma = 0,\quad \text{or}\quad \left.\frac{\partial u}{\partial n}\right|_\Gamma = 0 \tag{3.47}$$

and ξ_k as in the preceding lemma. Then, for any $2s+2 > n$ and $t \le T$,

$$\int_0^t\int_\Omega |\nabla u|^{2s+2}\cdot\xi_k^{2s+2}\,dx\,d\tau \le C\left[\int_0^t\int_\Omega |\nabla\varphi|^{2s+2}\cdot\xi_k^{2s+2}\,dx\,d\tau + I_2(k,t)\right], \tag{3.48}$$

where $C = C(n,s,\Omega,T,\sigma_1,|\varphi|_\infty)$,

$$I_2(k,t) = \left(\|u_0(x)\,\xi_k\|^{(1)}_{2s+2,\Omega_k}\right)^{2s+2} + \int_0^t \|u\nabla\xi_k\|^{2s+2}_{2s+2,\Omega}\,d\tau + \int_0^t \|\nabla\theta_k\|^{2s+2}_{2s+2,\Omega_k}\,d\tau. \tag{3.49}$$

$|f|^{(1)}_{k,\Omega} = |f|_{k,\Omega} + \|\nabla f\|_{k,\Omega}$, $\theta_k(x,t)$ is the solution to the problem

$$\Delta\theta_k = -\nabla\xi_k\cdot(\kappa(u)\nabla u + \sigma\varphi\nabla\varphi),\quad x\in\Omega_k,\quad \theta_k|_{\Gamma_k} = 0. \tag{3.50}$$

In particular, if $\xi_k \equiv 1$, then

$$\int_0^t\int_\Omega |\nabla u|^{2s+2}\,dx\,d\tau \le C\left[\int_0^t\int_\Omega |\nabla\varphi|^{2s+2}\,dx\,d\tau + \left(\|u_0\|^{(1)}_{2s+2,\Omega}\right)^{2s+2}\right]. \tag{3.51}$$

Proof. Introduce

$$u^k(x,t) = u(x,t)\,\xi_k(x).$$

Then

$$\frac{\partial u^k}{\partial t} = \nabla \cdot (\kappa(u)\nabla u^k + \sigma\varphi\nabla\varphi\xi_k + G_k), \quad x \in \Omega_k, \quad u^k|_{\Gamma k} = 0,$$
$$u^k(0) = u_0\xi_k, \quad x \in \Omega_k, \quad G_k = -(\kappa(u)u\nabla\xi_k + \nabla\theta_k). \tag{3.52}$$

We then deduce (see [L.S.U., Thm. 8.1, 8, Thm. 10.1, 10, Chap. III] and [A.K.M., Thm. 1, p. 230])

$$\|\nabla u^k\|^q_{q,\Omega_t} \le C\left(\|G_k\|^q_{q,\Omega_t} + \|\nabla\varphi\xi_k\|^q_{q,\Omega_t} + \|u_0\xi_k\|^{(1)}_{q,\Omega}\right)$$
$$\le C\left[\|\nabla\varphi\xi_k\|^q_{q,\Omega_t} + I_2(k,t)\right].$$

Hence

$$\|\nabla u\xi_k\|^q_{q,\Omega_t} \le C\left(\|\nabla\varphi\xi_k\|^q_{q,\Omega_t} + I_2\right)$$

(this for any $1 < q < \infty, t \le T, C = C(\Omega_k, T, n, q)$). When $q = 2s+2$ we get (3.48), and the proof of Lemma 3.2 is complete.

Step 3. Substituting (3.41) into the right-hand side of (3.48) and choosing the domain Ω_k small enough in such a way that

$$C^2\delta^{2s+2} \le \tfrac{1}{2},$$

we obtain

$$\int_0^t \int_\Omega |\nabla u|^{2s+2}\,\xi_k^{2s+2}\,dx\,d\tau \le C\left[I_1(k) + I_2(kt)\right]. \tag{3.53}$$

From (3.42) we have, $(2s+2 > n)$,

$$|I_1(k)| \le C\|\varphi - \varphi_0\|^{2s+2}_{\infty,\Omega} \le 2C\|\varphi_0\|^{2s+2}_{\infty,\Omega} \le \tilde{C}\left(\|\varphi_0\|^{(1)}_{\infty,\Omega}\right)^{2s+2}. \tag{3.54}$$

From (3.49) we also get

$$|I_2(k,t)| \le C\left[\left(\|u_0\|^{(1)}_{2s+2,\Omega}\right)^{2s+2} + \|u\|^{2s+2}_{2s+2,\Omega_t} + \int_0^t \|\nabla\theta_k\|^{2s+2}_{2s+2,\Omega_k}\,d\tau\right]. \tag{3.55}$$

For the solution to the problem (3.50) we have the following representation formula:

$$\theta_k = P\left((\kappa(u)\nabla u + \sigma\varphi\nabla\varphi)\nabla\xi_k\,|x\right), \qquad \left(\kappa(u)\nabla u = \nabla\left(\int_0^u \kappa(s)\,ds\right)\right),$$

where

$$P(g|x) = \int_{\Omega_k} I(x-y)\,g(y)\,dy,$$

I being Green's function. Thus

$$\nabla\theta_k = -\int_{\Omega_k} \nabla I \cdot \Delta\xi_k\kappa(u)\,u\,dy - \int_{\Omega_k} \nabla^2 I \cdot \nabla\xi_k \int_0^{u(x,t)} \kappa(s)\,ds\,dy$$
$$+ \int_{\Omega_k} \nabla I \cdot \nabla\xi_k\sigma\varphi\nabla\varphi\,dy. \tag{3.56}$$

By the properties of the operator P (see [L.U.]) and (3.56) we deduce

(3.57)

$$\begin{aligned}\|\nabla\theta_k\|_{2s+2,\Omega_k} &\le C\left[\|u\|_{(2s+2)n/(2s+2+n),\Omega}+\|u\|_{2s+2,\Omega}+\|\nabla\varphi\|_{(2s+2)n/(2s+2+n),\Omega}\right]\\ &\le C\left[\|u\|_{2s+2,\Omega}+\|\nabla\varphi\|_{(2s+2)n/(2s+2+n),\Omega}\right].\end{aligned}$$

From (3.53), (3.54), (3.55), (3.57) we have

(3.58)

$$\begin{aligned}\int_0^t \|\nabla u\|_{2s+2,\Omega}^{2s+2}\,dt &\le C\left[\left(\|u_0\|_{2s+2,\Omega}^{(1)}\right)^{2s+2}+|u|_{2s+2,\Omega_t}^{2s+2}+\int_0^t\left(\|\varphi_0\|_{2s+2,\Omega}^{(1)}\right)^{2s+2}dt\right.\\ &\qquad\left.+\int_0^t\left(\|\nabla\varphi\|_{(2s+2)n/(2s+2+n),\Omega}\right)^{2s+2}dt\right]\\ &\equiv C\left[H(u,\varphi)+|u|_{2s+2,\Omega_t}^{2s+2}+\int_0^t\left(\|\nabla\varphi\|_{(2s+2)n/(2s+2+n),\Omega}\right)^{2s+2}dt\right]\\ &\equiv Q.\end{aligned}$$

From (3.43) we also get, for φ,

$$\|\nabla\varphi\|_{2s+2,\Omega_t}^{2s+2}+\|\nabla u\|_{2s+2,\Omega_t}^{2s+2}\le 2Q. \tag{3.59}$$

Moreover, we have

$$\begin{aligned}\|u\|_{2s+2,\Omega_t}^{2s+2} &\le \varepsilon\|\nabla u\|_{2s+2,\Omega_t}^{2s+2}+C_\varepsilon\int_0^t|u|_{2,\Omega}^{2s+2}\,dt\\ &\le \varepsilon\|\nabla u\|_{2s+2,\Omega_t}^{2s+2}+\tilde{C}_\varepsilon\left(\int_0^t|u_0|_{2,\Omega}^2+\int_0^\tau\|\nabla\varphi\|_{2,\Omega}^2\,d\tau\right)^{(2s+2)/2}\\ &\le \varepsilon\|\nabla u\|_{2s+2,\Omega_t}^{2s+2}+\tilde{C}_\varepsilon' H\end{aligned} \tag{3.60}$$

and

$$\begin{aligned}\int_0^t\|\nabla\varphi\|_{q,\Omega}^{2s+2}\,dt &\le \int_0^t\|\nabla\varphi\|_{p,\Omega}^{(2s+2)/q}\|\nabla\varphi\|_{(q-1)p/(p-1),\Omega}^{(2s+2)(q-1)/q}\,dt\\ &\le C\int_0^t\|\nabla\varphi_0\|_{p,\Omega}^{(2s+2)/q}\|\nabla\varphi\|_{2s+2,\Omega}^{(2s+2)(q-1)/q}\,dt\\ &\le \varepsilon\|\nabla\varphi\|_{2s+2,\Omega_t}^{2s+2}+C_\varepsilon H\end{aligned} \tag{3.61}$$

$(q=(2s+2)n/(n+2s+2), p>2)$. Combining (3.59)–(3.61) we obtain (3.58), and the Theorem (3.2) is proved.

Remark 3.4. The estimate (3.38) allows us to prove the existence of a solution to (1.1) in the space of (u,φ) such that

$$(\nabla u,\nabla\varphi)\in L^{2s+2}(0,T;L^{2s+2}(\Omega)),\qquad (u_t,\nabla^2 u,\nabla^2\varphi)\in L^{s+1}(0,T;L^{s+1}(\Omega)).$$

4. Dependence on the data and uniqueness results. In this section we will assume that (2.6) and (2.7) hold and that κ, σ are Lipschitz continuous, i.e., that for some constant K,

$$|\kappa(u_1) - \kappa(u_2)|, \qquad |\sigma(u_1) - \sigma(u_2)| \le K|u_1 - u_2| \quad \forall u_1, u_2 \in \mathbf{R}. \tag{4.1}$$

Then we have the following.

THEOREM 4.1. *Let $(u_i, \varphi_i), i = 1, 2$, two weak solutions to* (1.1) *with the boundary conditions* (1.1b) *or* (1.1b′) *corresponding to the data $(u_0^i, \varphi_0^i, \kappa^i, \sigma^i)$. Assume that* (2.1), (2.6), (2.7), *and* (4.1) *hold for $(u_0^i, \varphi_0^i, \kappa^i, \sigma^i), i = 1, 2$, and also that*

$$\nabla u_i, \nabla \varphi_i \in L^{2q/(q-n)}(0, T; L^q(\Omega)), \qquad q > n \vee 2, \quad i = 1, 2, \tag{4.2}$$

where $n \vee 2$ denotes the maximum of 2 and n. Then we have

$$\begin{aligned} &|w(t)|_2^2 + \int_0^t \|\nabla w(\tau)\|_2^2 \, d\tau + \int_0^t \|\nabla \varphi(\tau)\|_2^2 \, d\tau \\ &\quad \le C\left(|w_0|_2^2 + |\kappa|_\infty^2 + |\sigma|_\infty^2 + \int_0^t \|\nabla \varphi_0\|_2^2 \, d\tau\right) \quad \forall t \le T, \end{aligned} \tag{4.3}$$

where

$$w = u_1 - u_2, \quad \varphi = \varphi_1 - \varphi_2, \quad w_0 = u_0^1 - u_0^2, \quad \varphi_0 = \varphi_0^1 - \varphi_0^2,$$
$$\kappa = \kappa^1 - \kappa^2, \qquad \sigma = \sigma^1 - \sigma^2,$$
$$|\kappa|_\infty = \sup_{\tau \in \mathbf{R}} |\kappa(\tau)|, \ |\sigma|_\infty = \sup_{\tau \in \mathbf{R}} |\sigma(\tau)|, \ C = C\left(T, \|\nabla u_i\|_{q,2q/(q-n)}, \|\nabla \varphi_i\|_{q,2q/(q-n)}\right).$$

Proof. Subtracting the equation satisfied by u_2 from the one satisfied by u_1 we obtain

$$\begin{aligned} w_t &= \nabla \cdot (\kappa^1(u_1) \nabla u_1 - \kappa^2(u_2) \nabla u_2) + \sigma^1(u_1) |\nabla \varphi_1|^2 - \sigma^2(u_2) |\nabla \varphi_2|^2 \\ &= \nabla \cdot (\kappa^1(u_1) \nabla w) + \nabla \cdot (\kappa^1(u_1) - \kappa^1(u_2) \nabla u_2) \\ &\quad + \nabla \cdot (\kappa^1(u_2) - \kappa^2(u_2) \nabla u_2) + (\sigma^1(u_1) - \sigma^1(u_2)) |\nabla \varphi_1|^2 \\ &\quad + \sigma^1(u_2) \nabla \varphi \cdot (\nabla \varphi_1 + \nabla \varphi_2) + (\sigma^1(u_2) - \sigma^2(u_2)) |\nabla \varphi_2|^2. \end{aligned}$$

If we multiply by w and integrate over Ω we get

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} |w|_2^2 + \kappa_1 \int_\Omega |\nabla w|^2 \, dx &\le \frac{1}{2} \frac{d}{dt} |w|_2^2 + \int_\Omega \kappa^1(u_1) |\nabla w|^2 \, dx \\ &= \int_\Omega \kappa^1(u_2) - \kappa^1(u_1) \nabla u_2 \cdot \nabla w \, dx \\ &\quad + \int_\Omega \kappa^2(u_2) - \kappa^1(u_2) \nabla u_2 \cdot \nabla w \, dx \\ &\quad + \int_\Omega (\sigma^1(u_1) - \sigma^1(u_2)) |\nabla \varphi_1|^2 w \, dx \\ &\quad + \int_\Omega \sigma^1(u_2) \nabla \varphi \cdot (\nabla \varphi_1 + \nabla \varphi_2) w \, dx \\ &\quad + \int_\Omega (\sigma^1(u_2) - \sigma^2(u_2)) |\nabla \varphi_2|^2 w \, dx. \end{aligned} \tag{4.4}$$

Using (4.1) and Hölder's and Young's inequalities we easily see that

(4.5)

$$\begin{aligned}
\left|\int_\Omega \kappa^1(u_2)-\kappa^1(u_1)\,\nabla u_2\cdot\nabla w\,dx\right| &\le K\int_\Omega |\nabla u_2\,\|\nabla w\|\,w|\,dx\\
&\le K\,|\nabla u_2|_q\,\|\nabla w\|_2\,|w|_{2q/(q-2)}\\
&\le \varepsilon\,\|\nabla w\|_2^2+C_\varepsilon\,\|\nabla u_2\|_q^2\,|w|_{2q/(q-2)}^2,\\
\left|\int_\Omega \kappa^2(u_2)-\kappa^1(u_2)\,\nabla u_2\cdot\nabla w\,dx\right| &\le |\kappa|_\infty\,\|\nabla u_2\|_2\,\|\nabla w\|_2\\
&\le \varepsilon\,\|\nabla w\|_2^2+C_\varepsilon\,|\kappa|_\infty^2\,\|\nabla u_2\|_q^2,\\
\left|\int_\Omega (\sigma^1(u_1)-\sigma^1(u_2))\,|\nabla\varphi_1|^2\,w\,dx\right| &\le K\int_\Omega |\nabla\varphi_1|^2\,|w|^2\,dx\\
&\le K\,\|\nabla\varphi_1\|_q^2\,|w|_{2q/(q-2)}^2,\\
\left|\int_\Omega \sigma^1(u_2)\,\nabla\varphi\cdot(\nabla\varphi_1+\nabla\varphi_2)\,w\,dx\right| &\le \sigma_2\,\|\nabla\varphi\|_2\,\|\nabla(\varphi_1+\varphi_2)\|_q\,|w|_{2q/(q-2)}=I,\\
\left|\int_\Omega (\sigma^1(u_2)-\sigma^2(u_2))\,|\nabla\varphi_2|^2\,w\,dx\right| &\le \|\nabla\varphi_2\|_q^2\,|w|_{2q/(q-2)}\,|\sigma|_\infty\\
&\le C\,\|\nabla\varphi_2\|_q^2\left(|w|_{2q/(q-2)}^2+|\sigma|_\infty^2\right).
\end{aligned}$$

To estimate I, we need to estimate φ. So, we use the equation satisfied by φ_1 and φ_2 to get

$$\begin{aligned}
-\nabla\cdot(\sigma^1(u_1)\,\nabla\varphi) &= -\nabla\cdot(\sigma^1(u_2)-\sigma^1(u_1)\,\nabla\varphi_2)\\
&\quad -\nabla\cdot(\sigma^2(u_2)-\sigma^1(u_2)\,\nabla\varphi_2).
\end{aligned}\tag{4.6}$$

Multiplying this equation by $\varphi-\varphi_0$ and integrating over Ω leads to

(4.7)

$$\begin{aligned}
\int_\Omega \sigma^1(u_1)\,|\nabla\varphi|^2\,dx &= \int_\Omega \sigma^1(u_2)-\sigma^1(u_1)\,\nabla\varphi_2\cdot\nabla\varphi\,dx\\
&\quad+\int_\Omega \sigma^2(u_2)-\sigma^1(u_2)\,\nabla\varphi_2\cdot\nabla\varphi\,dx+\int_\Omega \sigma^1(u_1)\,\nabla\varphi\cdot\nabla\varphi_0\,dx\\
&\quad-\int_\Omega \sigma^1(u_2)-\sigma^1(u_1)\,\nabla\varphi_2\cdot\nabla\varphi_0\,dx\\
&\quad-\int_\Omega \sigma^2(u_2)-\sigma^1(u_2)\,\nabla\varphi_2\cdot\nabla\varphi_0\,dx.
\end{aligned}$$

From this equality it follows that

$$\begin{aligned}
\|\nabla\varphi\|_2^2 \le C\Big\{&\|\nabla\varphi\|_2\left(\|\nabla\varphi_2\|_q\,|w|_{2q/(q-2)}+|\sigma|_\infty\,\|\nabla\varphi_2\|_2+\|\nabla\varphi_0\|_2\right)\\
&+\|\nabla\varphi_2\|_q\,\|\nabla\varphi_0\|_2\,|w|_{2q/(q-2)}+\|\nabla\varphi_2\|_2\,\|\nabla\varphi_0\|_2\,|\sigma|_\infty\Big\},
\end{aligned}$$

and by Young's inequality,

$$\|\nabla\varphi\|_2^2 \le \varepsilon\,\|\nabla\varphi\|_2^2+C_\varepsilon\left(\|\nabla\varphi_2\|_q^2\,|w|_{2q/(q-2)}^2+|\sigma|_\infty^2\,\|\nabla\varphi_2\|_2^2+\|\nabla\varphi_0\|_2^2\right).\tag{4.8}$$

We thus obtain

$$\begin{aligned}\|\nabla\varphi\|_2 &\le C\left(\|\nabla\varphi_2\|_q\,|w|_{2q/(q-2)} + |\sigma|_\infty\,\|\nabla\varphi_2\|_2 + \|\nabla\varphi_0\|_2\right)\\ &\le C\left(\|\nabla\varphi_2\|_q\,|w|_{2q/(q-2)} + |\sigma|_\infty\,\|\nabla\varphi_2\|_q + \|\nabla\varphi_0\|_2\right),\end{aligned}$$

and so

$$\begin{aligned}I &\le C\left[\left\{\|\nabla\varphi_1\|_q^2 + \|\nabla\varphi_2\|_q^2\right\}|w|_{2q/(q-2)}^2\right.\\ &\qquad \left. + \|\nabla(\varphi_1+\varphi_2)\|_q\,|w|_{2q/(q-2)}\left\{|\sigma|_\infty\,\|\nabla\varphi_2\|_q + \|\nabla\varphi_0\|_2\right\}\right] \qquad (4.9)\\ &\le C\left[\left\{\|\nabla\varphi_1\|_q^2 + \|\nabla\varphi_2\|_q^2\right\}|w|_{2q/(q-2)}^2 + |\sigma|_\infty^2\,\|\nabla\varphi_2\|_q^2 + \|\nabla\varphi_0\|_2^2\right].\end{aligned}$$

Collecting (4.4), (4.5), and (4.9) and choosing $\varepsilon = \kappa_1/6$ in (4.5), we get

(4.10)

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}|w|_2^2 + \frac{2\kappa_1}{3}\|\nabla w\|_2^2 &\le C\left\{\|\nabla u_2\|_q^2 + \|\nabla\varphi_1\|_q^2 + \|\nabla\varphi_2\|_q^2\right\}|w|_{2q/(q-2)}^2\\ &\quad + C\left\{|\kappa|_\infty^2\,\|\nabla u_2\|_q^2 + |\sigma|_\infty^2\,\|\nabla\varphi_2\|_q^2 + \|\nabla\varphi_0\|_2^2\right\}.\end{aligned}$$

From the Gagliardo-Nirenberg interpolation inequality we have for some constant C,

$$|w|_{2q/(q-2)} \le C\,|w|_2^{1-(n/q)}\left(|w|_2^2 + \|\nabla w\|_2^2\right)^{n/2q} \qquad \forall w \in H^1(\Omega). \tag{4.11}$$

Hence (4.10) becomes

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}|w|_2^2 + \frac{2\kappa_1}{3}\|\nabla w\|_2^2 &\le C\left\{\|\nabla u_2\|_q^2 + \|\nabla\varphi_1\|_q^2 + \|\nabla\varphi_2\|_q^2\right\}|w|_2^{2(1-(n/q))}\\ &\quad\cdot\left(|w|_2^2 + \|\nabla w\|_2^2\right)^{n/q} \qquad (4.12)\\ &\quad + C\left\{|\kappa|_\infty^2\,\|\nabla u_2\|_q^2 + |\sigma|_\infty^2\,\|\nabla\varphi_2\|_q^2 + \|\nabla\varphi_0\|_2^2\right\}.\end{aligned}$$

Hence by applying the Young inequality

$$ab \le \varepsilon a^{q/n} + C_\varepsilon b^{q/(q-n)},$$

it follows that for any $\varepsilon > 0$,

(4.13)

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}|w|_2^2 + \frac{2\kappa_1}{3}\|\nabla w\|_2^2 &\le 3\varepsilon\left(|w|_2^2 + \|\nabla w\|_2^2\right)\\ &\quad + C_\varepsilon\left(\|\nabla u_2\|_q^{2q/(q-n)} + \|\nabla\varphi_1\|_q^{2q/(q-n)} + \|\nabla\varphi_2\|_q^{2q/(q-n)}\right)|w|_2^2\\ &\quad + C\left\{|\kappa|_\infty^2\,\|\nabla u_2\|_q^2 + |\sigma|_\infty^2\,\|\nabla\varphi_2\|_q^2 + \|\nabla\varphi_0\|_2^2\right\},\end{aligned}$$

where C_ε is some constant depending on ε. Hence, by choosing $3\varepsilon = \kappa_1/6$,

(4.14)

$$\frac{d}{dt}|w|_2^2 + \kappa_1 \|\nabla w\|_2^2 \leq C\left[\left(1 + \|\nabla u_2\|_q^{2q/(q-n)} + \|\nabla \varphi_1\|_q^{2q/(q-n)} + \|\nabla \varphi_2\|_q^{2q/(q-n)}\right)|w|_2^2 + |\kappa|_\infty^2 \|\nabla u_2\|_q^2 + |\sigma|_\infty^2 \|\nabla \varphi_2\|_q^2 + \|\nabla \varphi_0\|_2^2\right].$$

If we set

$$[w(t)] = |w|_2^2 + \kappa_1 \int_0^t \|\nabla w\|_2^2 \, d\tau, \tag{4.15}$$

$$H = C\left(1 + \|\nabla u_2\|_q^{2q/(q-n)} + \|\nabla \varphi_1\|_q^{2q/(q-n)} + \|\nabla \varphi_2\|_q^{2q/(q-n)}\right) \in L^1(0,T) \tag{4.16}$$

(see (4.2)), then (4.14) also reads

$$\frac{d}{dt}[w] - H[w] \leq |\kappa|_\infty^2 \|\nabla u_2\|_q^2 + |\sigma|_\infty^2 \|\nabla \varphi_2\|_q^2 + \|\nabla \varphi_0\|_2^2$$

or

$$\frac{d}{dt}\left(e^{-\int_0^t H(s)ds}[w]\right) \leq e^{-\int_0^t H(s)ds}\left\{|\kappa|_\infty^2 \|\nabla u_2\|_q^2 + |\sigma|_\infty^2 \|\nabla \varphi_2\|_q^2 + \|\nabla \varphi_0\|_2^2\right\}.$$

Hence, integrating between o and t,

$$\begin{aligned}[w] &\leq [w(0)] + e^{\int_0^t H(s)ds}\left(\int_0^t e^{-\int_0^t H(s)ds}\left\{|\kappa|_\infty^2 \|\nabla u_2\|_q^2 + |\sigma|_\infty^2 \|\nabla \varphi_2\|_q^2 + \|\nabla \varphi_0\|_2^2\right\} d\tau\right) \\ &= |w_0|_2^2 + C(T)\left(|\kappa|_\infty^2 \left\{\int_0^t \|\nabla u_2\|_q^{2q/(q-n)} \, d\tau\right\}^{(q-n)/q}\right. \\ &\qquad \left. + |\sigma|_\infty^2 \left\{\int_0^t \|\nabla \varphi_2\|_q^{2q/(q-n)} \, d\tau\right\}^{(q-n)/q} + \int_0^t \|\nabla \varphi_0\|_2^2 \, d\tau\right).\end{aligned}$$

So we have

$$|w|_2^2 + \int_0^t \|\nabla w\|_2^2 \, d\tau \leq C\left(|w_0|_2^2 + |\kappa|_\infty^2 + |\sigma|_\infty^2 + \int_0^t \|\nabla \varphi_0\|_2^2 \, d\tau\right). \tag{4.17}$$

To complete the estimate (4.3) we go back to (4.8), which implies by (4.11),

$$\begin{aligned}\|\nabla \varphi\|_2^2 \leq C\left(\|\nabla \varphi_2\|_q^2 |w|_2^{2(1-(n/q))}\left(|w|_2^2 + \|\nabla w\|_2^2\right)^{n/q}\right. \\ \left. + |\sigma|_\infty^2 \|\nabla \varphi_2\|_q^2 + \|\nabla \varphi_0\|_2^2\right).\end{aligned} \tag{4.18}$$

Integrating between zero and t and applying Hölder's inequality we arrive at

$$\begin{aligned}
&\int_0^t \|\nabla\varphi\|_2^2\, d\tau \\
&\quad \le C\left[\left\{\int_0^t \|\nabla\varphi_2\|_q^{2q/(q-n)}\, d\tau\right\}^{(q-n)/q}\left\{\int_0^t |w|_2^2 + \|\nabla w\|_2^2\, d\tau\right\}^{n/q}\sup_{\tau\le t}|w|_2^{2(1-(n/q))}\right. \\
&\qquad \left. + |\sigma|_\infty^2\left\{\int_0^t \|\nabla\varphi_2\|_2^{2q/(q-n)}\, d\tau\right\}^{(q-n)/q} + \int_0^t \|\nabla\varphi_0\|_2^2\, d\tau\right] \\
&\quad \le C\left(|w_0|_2^2 + |\kappa|_\infty^2 + |\sigma|_\infty^2 + \int_0^t \|\nabla\varphi_0\|_2^2\, d\tau\right)
\end{aligned}$$

by (4.17). This completes the proof of (4.3).

Remark 4.1. If $\kappa_i \equiv 1$, Theorem 4.1 holds when we just assume that

$$\nabla\varphi_i \in L^{2q/(q-n)}(0,T;L^q(\Omega)), \qquad q > n \vee 2$$

since in the second side of (4.4) the two first integrals disappear.

COROLLARY 4.1. *There exists at most one weak solution to* (1.1) *with the boundary conditions* (1.1b) *or* (1.1b′) *such that*

$$\nabla u, \nabla\varphi \in L^{2q/(q-n)}(0,T;L^q(\Omega)), \qquad q > n \vee 2, \tag{4.19}$$

where $n \vee 2$ *denotes the maximum of* 2 *and* n.

Proof. If $(u_i,\varphi_i), i = 1,2$, are two weak solutions to (1.1) with the boundary conditions (1.1b) or (1.1b′) and corresponding to the same initial and boundary data, then (4.3) reads

$$|w(t)|_2^2 + \int_0^t \|\nabla w(\tau)\|_2^2\, d\tau + \int_0^t \|\nabla\varphi(\tau)\|_2^2\, d\tau \le 0.$$

and the result follows (see also Remark 4.1).

THEOREM 4.2. *Assume that* (4.1) *holds and that there exists one weak solution* (u_1,φ_1) *to* (1.1) *with the boundary conditions* (1.1b) *or* (1.1b′) *such that*

$$\begin{aligned}
&\nabla u_1 \in L^{2q/(q-n)}(0,T;L^q(\Omega)), \quad \nabla\varphi_1 \in L^{4q/(q-n)}(0,T;L^q(\Omega)), \\
&\qquad q > n \vee 2, \quad \varphi_1 \text{ bounded},
\end{aligned} \tag{4.20}$$

where $n \vee 2$ *denotes the maximum of* 2 *and* n. *Then, every weak solution* (*or classical solution*) (u_2,φ_2) *to* (1.1), *which is such that* φ_2 *is bounded, agrees with it.*

Proof. If we set $w = u_1 - u_2, \varphi = \varphi_1 - \varphi_2$ we have

(4.21)

$$\begin{aligned}
w_t &= \nabla\cdot(\kappa(u_2)\nabla w + (\kappa(u_1)-\kappa(u_2))\nabla u_1) + \nabla\cdot(\sigma(u_1)\varphi_1\nabla\varphi_1 - \sigma(u_2)\varphi_2\nabla\varphi_2) \\
&= \nabla\cdot(\kappa(u_2)\nabla w + (\kappa(u_1)-\kappa(u_2))\nabla u_1) + \nabla\cdot((\sigma(u_1)-\sigma(u_2))\varphi_1\nabla\varphi_1 \\
&\qquad + \sigma(u_2)(\varphi_1-\varphi_2)\nabla\varphi_1 + \sigma(u_2)\varphi_2(\nabla\varphi_1 - \nabla\varphi_2)) \\
&= \nabla\cdot(\kappa(u_2)\nabla w + (\kappa(u_1)-\kappa(u_2))\nabla u_1) + \nabla\cdot((\sigma(u_1)-\sigma(u_2))\varphi_1\nabla\varphi_1 \\
&\qquad + \sigma(u_2)\varphi\nabla\varphi_1 + \sigma(u_2)\varphi_2\nabla\varphi).
\end{aligned}$$

If we multiply by w and integrate over Ω we get

$$\frac{1}{2}\frac{d}{dt}|w|_2^2 + \kappa_1 \|\nabla w\|_2^2 \le I_1 + I_2 + I_3 + I_4, \tag{4.22}$$

where

$$I_1 = -\int_\Omega (\sigma(u_1) - \sigma(u_2))\,\varphi_1 \nabla\varphi_1 \cdot \nabla w\, dx,$$

$$I_2 = -\int_\Omega \sigma(u_2)\,\varphi \nabla\varphi_1 \cdot \nabla w\, dx,$$

$$I_3 = -\int_\Omega \sigma(u_2)\,\varphi_2 \nabla\varphi \cdot \nabla w\, dx,$$

$$I_4 = -\int_\Omega (\kappa(u_1) - \kappa(u_2))\,\nabla u_1 \cdot \nabla w\, dx.$$

Since σ and the φ_i's are bounded we obtain, by Hölder's inequality,

$$|I_1| \le C\int_\Omega |\nabla\varphi_1 \|\nabla w\|\, w|\, dx \le C\|\nabla\varphi_1\|_q \|\nabla w\|_2 |w|_{2q/(q-2)}, \tag{4.23}$$

$$|I_2| \le C\int_\Omega |\varphi \|\nabla\varphi_1\|\, \nabla w|\, dx \le C\|\nabla\varphi_1\|_q \|\nabla w\|_2 |\varphi|_{2q/(q-2)},$$

$$|I_3| \le C\int_\Omega |\nabla\varphi \|\nabla w|\, dx \le C\|\nabla\varphi\|_2 \|\nabla w\|_2. \tag{4.24}$$

$$|I_4| \le C\int_\Omega |\nabla u_1 \|\nabla w\|\, w|\, dx \le C\|\nabla u_1\|_q \|\nabla w\|_2 |w|_{2q/(q-2)}. \tag{4.25}$$

Since $q > n$ from the Sobolev imbedding theorem, we get

$$|I_2| \le C\|\nabla\varphi_1\|_q \|\nabla w\|_2 \|\nabla\varphi\|_2. \tag{4.26}$$

Now from the equation satisfied by φ_1, φ_2 we have

$$0 = \nabla\cdot(\sigma(u_2)\nabla\varphi_2) = \nabla\cdot(\sigma(u_2)\nabla(\varphi_2 - \varphi_1)) + \nabla\cdot((\sigma(u_2) - \sigma(u_1))\nabla\varphi_1).$$

Multiplying by φ and integrating on Ω we obtain

$$\int_\Omega \sigma(u_2)|\nabla\varphi|^2\, dx = \int_\Omega (\sigma(u_1) - \sigma(u_2))\nabla\varphi_1 \cdot \nabla\varphi\, dx.$$

Hence by Hölder's inequality,

$$\|\nabla\varphi\|_2^2 \le C\int_\Omega |w\,\|\nabla\varphi_1\|\,\nabla\varphi|\, dx \le C\|\nabla\varphi\|_2 \|\nabla\varphi_1\|_q |w|_{2q/(q-2)}. \tag{4.27}$$

Collecting (4.22)–(4.27) we get

$$\frac{1}{2}\frac{d}{dt}|w|_2^2 + \kappa_1\|\nabla w\|_2^2 \le C\{\|\nabla\varphi_1\|_q\|\nabla w\|_2|w|_{2q(q-2)} + \|\nabla u_1\|_q\|\nabla w\|_2|w|_{2q/(q-2)} + \|\nabla\varphi_1\|_q^2\|\nabla w\|_2|w|_{2q/(q-2)}\}.$$

Applying Young's inequality we easily deduce that

$$\frac{1}{2}\frac{d}{dt}|w|_2^2 + \frac{\kappa_1}{2}\|\nabla w\|_2^2 \le C\{\|\nabla\varphi_1\|_q^2 + \|\nabla u_1\|_q^2 + \|\nabla\varphi_1\|_q^4\}|w|_{2q/(q-2)}^2.$$

By (4.11) and again applying Young's inequality we get

$$\frac{1}{2}\frac{d}{dt}|w|_2^2+\frac{\kappa_1}{2}\|\nabla w\|_2^2\leq C\{\|\nabla\varphi_1\|_q^2+\|\nabla u_1\|_q^2+\|\nabla\varphi_1\|_q^4\}|w|_2^{2(1-(n/q))}\cdot(|w|_2^2+\|\nabla w\|_2^2)^{n/q}$$
$$\leq \varepsilon(|w|_2^2+\|\nabla w\|_2^2)+C_\varepsilon\left\{\|\nabla\varphi_1\|_q^{2q/(q-n)}+\|\nabla u_1\|_q^{2q/(q-n)}\right.$$
$$\left.+\ \|\nabla\varphi_1\|_q^{4q/(q-n)}\right\}|w|_2^2.$$

Choosing $\varepsilon=\kappa_1/2$ we obtain

$$\frac{d}{dt}|w|_2^2\leq C\left\{1+\|\nabla u_1\|_q^{2q/(q-n)}+\|\nabla\varphi_1\|_q^{4q/(q-n)}\right\}|w|_2^2.$$

Since

$$1+\|\nabla u_1\|_q^{2q/(q-n)}+\|\nabla\varphi_1\|_q^{4q/(q-n)}\in L^1(0,T),$$

the result follows from the Gronwall inequality.

Remark 4.2. These results improve preceding results of [Ch.C.]. Note that (4.2) holds automatically when $n=1$; see [Ch.C.].

5. A blowup result. The results of this section improve and complete the results contained in [A.C.1]. Interesting results on blowup could also be found in [L].

Let us consider $(u(x,t),\varphi(x,t))$, a local solution to

$$\begin{aligned}
&u_t=\nabla\cdot(\kappa(u)\nabla u)+\sigma(u)|\nabla\varphi|^2, \quad x\in\Omega, \quad t>0,\\
&\partial u/\partial n=0, \quad x\in\Gamma, \quad t>0,\\
&u(x,0)=u_0(x), \qquad x\in\Omega,\\
&\nabla\cdot(\sigma(u)\nabla\varphi)=0, \quad x\in\Omega, \quad t>0,\\
&\varphi=\varphi_0, \quad x\in\Gamma, \quad t>0.
\end{aligned} \tag{5.1}$$

Let us assume that

$$u_0(x)\geq 0, \qquad x\in\Omega. \tag{5.2}$$

$$0<\kappa(s), \quad \sigma(s)<+\infty \quad \forall s\geq 0, \quad \sigma \text{ differentiable}, \quad \sigma'(s)\geq 0 \quad \forall s\geq 0, \tag{5.3}$$

$$\int_0^{+\infty}\frac{ds}{\sigma(s)}<+\infty. \tag{5.4}$$

If $d\gamma(x)$ is the superficial measure on Γ we remark that

$$\lambda\to\int_\Gamma|\varphi-\lambda|^2\,d\gamma(x)$$

achieves its minimum value for

$$\lambda=\bar{\varphi}=\frac{1}{|\Gamma|}\int_\Gamma\varphi\,d\gamma(x).$$

So, if we set

$$\varphi_\Omega = \frac{1}{|\Omega|}\int_\Omega \varphi\, dx,$$

we have, for some constant C,

$$\int_\Gamma |\varphi - \bar{\varphi}|^2\, d\gamma(x) \le \int_\Gamma |\varphi - \varphi_\Omega|^2\, d\gamma(x) \le C\int_\Omega |\nabla\varphi|^2\, dx \quad \forall \varphi \in H^1(\Omega).$$

Let us denote again by C the best constant such that

$$\int_\Gamma |\varphi - \bar{\varphi}|^2\, d\gamma(x) \le C\int_\Omega |\nabla\varphi|^2\, dx \quad \forall \varphi \in H^1(\Omega). \tag{5.5}$$

Then we can prove the following.

THEOREM 5.1. *Assume that*

$$\int_\Omega \int_{u_0(x)}^{+\infty} \frac{ds}{\sigma(s)}\, dx < \frac{1}{C}\int_0^{+\infty}\int_\Gamma |\varphi_0 - \bar{\varphi}_0|^2\, d\gamma(x)\, dt, \tag{5.6}$$

where

$$\bar{\varphi}_0 = \frac{1}{|\Gamma|}\int_\Gamma \varphi_0\, d\gamma(x),$$

then (5.1) *cannot have a smooth global solution.*

Proof. Let us assume that (5.1) has a smooth global solution. Define

$$Y(t) = \int_\Omega \left(\int_{u(x,t)}^{+\infty} \frac{ds}{\sigma(s)}\right) dx. \tag{5.7}$$

From (5.2) and the maximum principle (see [F.]) it is clear that

$$u(x,t) \ge 0, \quad x \in \Omega, \quad t > 0, \tag{5.8}$$

and thus $Y(t)$ makes sense and is nonnegative (see (5.4)).

Differentiating we obtain, using (5.1),

$$\begin{aligned} \frac{dY(t)}{dt} &= -\int_\Omega \frac{u_t}{\sigma(u)}\, dx \\ &= -\int_\Omega \frac{\nabla\cdot(\kappa(u)\nabla u) + \sigma(u)|\nabla\varphi|^2}{\sigma(u)}\, dx \\ &= -\int_\Omega \nabla\cdot(\kappa(u)\nabla u)\cdot\frac{1}{\sigma(u)}\, dx - \int_\Omega |\nabla\varphi|^2\, dx. \end{aligned} \tag{5.9}$$

Integrating by parts we have, since $\partial u/\partial n = 0$ on Γ and by (5.3),

$$-\int_\Omega \nabla\cdot(\kappa(u)\nabla u)\cdot\frac{1}{\sigma(u)}\, dx = -\int_\Omega \frac{\kappa(u)\sigma'(u)}{\sigma^2(u)}|\nabla u|^2\, dx \le 0. \tag{5.10}$$

Hence

$$\frac{dY(t)}{dt} \leq -\int_{\Omega} |\nabla\varphi|^2 \, dx, \tag{5.11}$$

from which it follows that

$$\frac{dY(t)}{dt} \leq -\frac{1}{C}\int_{\Gamma} |\varphi_0 - \bar{\varphi}_0|^2 \, d\gamma(x).$$

Integrating between zero and t we get

$$0 \leq Y(t) \leq T(0) - \frac{1}{C}\int_0^t \int_{\Gamma} |\varphi_0 - \bar{\varphi}_0|^2 \, d\gamma(x) \, dt,$$

which by (5.6) is impossible for t large.

Remark 5.1. In the case where

$$\varphi_0 = \varphi_0(x), \tag{5.12}$$

it is shown that (5.1) has a global solution if and only if

$$\varphi_0 = \text{ Const.}$$

Indeed, in this case (5.6) holds except when

$$\varphi_0 = \bar{\varphi}_0 = \text{ Const.}$$

A more convincing example showing the sharpness of (5.6) under the assumptions of Theorem 5.1 is the following. Consider $\Omega = (0,1)$. Then if φ is a function in $H^1(0,1)$,

$$\bar{\varphi} = \frac{1}{|\Gamma|}\int_{\Gamma} \varphi \, d\gamma(x) = \frac{1}{2}\{\varphi(0) + \varphi(1)\}.$$

Moreover,

$$\int_{\Gamma} |\varphi - \bar{\varphi}|^2 \, d\gamma(x) = |\varphi(0) - \bar{\varphi}|^2 + |\varphi(1) - \bar{\varphi}|^2 = \frac{|\varphi(0) - \varphi(1)|^2}{2}.$$

Now we have

$$|\varphi(0) - \varphi(1)| = \left|\int_0^1 \varphi'(s) \, ds\right| \leq \left\{\int_0^1 (\varphi'(s))^2 \, ds\right\}^{1/2},$$

which shows by squaring that

$$\frac{|\varphi(0) - \varphi(1)|^2}{2} = \int_{\Gamma} |\varphi - \bar{\varphi}|^2 \, d\gamma(x) \leq \frac{1}{2}\int_0^1 (\varphi'(s))^2 \, ds. \tag{5.13}$$

The constant $\frac{1}{2}$ in (5.13) is the best possible as it can be seen by taking

$$\varphi(s) = s.$$

Then consider the one-dimensional version of (5.1) with

$$u(0) = u_0 = \text{ Const.},$$

and look for a solution

$$u = u(t)$$

depending on t only. Set

$$\varphi_0(0,t) = A_0(t), \qquad \varphi_0(1,t) = A_1(t).$$

Then, clearly, the equation satisfied by φ leads to

$$\varphi(x,t) = A_0(t) + x(A_1(t) - A_0(t)),$$

and the equation in u becomes

$$u_t = \sigma(u)(A_1(t) - A_0(t))^2 \tag{5.14}$$

or

$$\int_{u_0}^{u} \frac{ds}{\sigma(s)} = \int_0^t (A_1(s) - A_0(s))^2 \, ds.$$

In the case we are considering, the failure of (5.6) reads

$$\int_{u_0}^{+\infty} \frac{ds}{\sigma(s)} \geq \int_0^{+\infty} (A_1(s) - A_0(s))^2 \, ds.$$

This implies that (5.14) has a global solution which is bounded when

$$\int_{u_0}^{+\infty} \frac{ds}{\sigma(s)} > \int_0^{+\infty} (A_1(s) - A_0(s))^2 \, ds,$$

and is unbounded otherwise.

Remark 5.2. In dimension 1 and when $\kappa \equiv 1$ it is possible to show that $u(x,t)$ blows up globally, i.e., if t^* denotes the blowup time then

$$u(x,t) \to +\infty \quad \text{a.e. } x \in \Omega \quad \text{when } t \to t^*.$$

Indeed, if for instance $\Omega = (0,1)$, then by integrating the equation

$$(\sigma(u)\varphi') = 0$$

we get

$$\sigma(u)\varphi' = C(t).$$

Hence

$$\varphi' = \frac{C(t)}{\sigma(u)}$$

with

$$\varphi(1,t) - \varphi(0,t) = C(t) \int_0^1 \frac{dx}{\sigma(u(x,t))}.$$

Setting $\lambda(t) = \varphi(1,t) - \varphi(0,t)$ the equation satisfied by u reads

$$u_t = u_{xx} + \frac{\lambda^2}{\sigma(u)} \left(\int_0^1 \frac{dx}{\sigma(u(x,t))} \right)^{-2}.$$

Differentiating in x we see that $v = u_x$ satisfies

$$v_t = v_{xx} - \lambda^2 \frac{\sigma'(u)}{\sigma(u)^2} \left(\int_0^1 \frac{dx}{\sigma(u(x,t))} \right)^{-2} v,$$

$$v(x,t) = 0, \quad x = 0,1, \quad v(x,0) = (u_0)_x.$$

Assuming that $(u_0)_x \in L^\infty(0,1)$ it follows from the maximum principle, recall that

$$\lambda^2 \frac{\sigma'(u)}{\sigma(u)^2} \left(\int_0^1 \frac{dx}{\sigma(u(x,t))} \right)^{-2} \geq 0,$$

that

$$|u_x|_\infty \leq |(u_0)_x|_\infty. \tag{5.15}$$

Hence

$$u(x,t) = \int_{x_0}^x u_x(x,t)\, dx + u(x_0,t).$$

If $u(x_0,t)$ blows up, then $u(x,t)$ blows up for any x since the integral is bounded thanks to (5.15).

Acknowledgments. This work was started when both authors were visiting the Institute for Mathematics and its Applications in Minneapolis. We thank this institution for its support.

The work of the first author was completed during a visit at the University of Metz. He would like to thank the Mathematics Department for its hospitality.

REFERENCES

[A.C.1] S. N. ANTONTSEV AND M. CHIPOT, *Some results on the thermistor problem,* Proceedings of the Conference Free-Boundary Problems in Continuum Mechanics, Novosibirsk, Russia, July 1991.

[A.C.2] ———, *Existence, stability, blow up of the solution for the thermistor problem,* Dokl. Russian Acad. Nauk, 324 (1992).

[A.K.M.] S. N. ANTONTSEV, A. V. KAZHIKHOV, AND V. N. MONAKHOV, *Boundary Value Problems in Mechanics of Nonhomogeneous Fluids,* Stud. Math. Appl., 22 (1990), North-Holland, Amsterdam.

[B.L.] A. BENSOUSSAN AND J. L. LIONS, *Applications des Inéquations Variationnelles en Contrôle Stochastique,* Dunod, Paris, 1978.

[C.F.1] X. CHEN AND A. FRIEDMAN, *The thermistor problem for conductivity which vanishes at large temperature,* Quart. Appl. Math., to appear.

[C.F.2] ———, *The thermistor problem with one-zero conductivity*, to appear.
[C.F.3] ———, *The thermistor problem with one-zero conductivity* II, preprint, to appear.
[C.D.K.] M. CHIPOT, J. I. DIAZ, AND R. KERSNER, *Existence and uniqueness results for the thermistor problem with temperature dependent conductivity*, to appear.
[Ch.C] M. CHIPOT AND G. CIMATTI, *A uniqueness result for the thermistor problem*, European J. Appl. Math., 2 (1991), pp. 97–103.
[C.1] G. CIMATTI, *Existence of weak solutions for the nonstationary problem of the Joule heating of a conductor*, preprint, Università di Pisa, Pisa, Italy, 1992.
[C.2] ———, *A bound for the temperature in the thermistor problem*, J. Appl. Math., 40 (1988), pp. 15–22.
[C.3] ———, *Remark on existence and uniqueness for the thermistor problem*, Quart. Appl. Math., 47 (1989), pp. 117–121.
[C.P.] G. CIMATTI AND G. PRODI, *Existence results for a nonlinear elliptic system modelling a temperature dependent electrical resistor*, Ann. Mat. Pura Appl., 152 (1989), pp. 227–236.
[D.L.] R. DAUTRAY AND J. L. LIONS, *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*, Masson, Paris, 1988.
[F.] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[G.T.] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1985.
[H.R.S.] S. D. HOWISON, J. F. RODRIGUES, AND M. SHILLOR, *Stationary solution to the thermistor problem*, J. Math. Anal. Appl., to appear.
[Ko.] F. KOHLRAUSCH, *Über den stationären Temperatur-zustand eines electrisch geheizten Leiters*, Ann. Phys., 1 (1900), pp. 132–158.
[L] A. LACEY, to appear.
[L.S.U.] O. A. LADYŽENSKAJA, V. A. SOLONIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.
[L.U.] O. A. LADYŽENSKAJA AND N. N. URAL'CEVA, *Equations aux Dérivées Partielles de Type Elliptique*, Dunod, Paris, 1968.
[J.L.L.] J. L. LIONS, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[P.W.] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[X.] X. XU, *A Stefan-like problem arising from the electrical heating of a conductor with conductivity vanishing at finite temperature*, preprint, to appear.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1157–1166, July 1994

007

EXISTENCE AND UNIQUENESS OF THE C^α SOLUTION FOR THE THERMISTOR PROBLEM WITH MIXED BOUNDARY VALUE *

GUANGWEI YUAN† AND ZUHAN LIU‡

Abstract. The thermistor problem is modeled as a coupled system of nonlinear PDEs with a quadratic growth on the gradient of one of the unknowns. The existence and uniqueness of C^α-solution for this system with mixed boundary conditions is established.

Key words. Hölder estimate, mixed boundary value, thermistor

AMS subject classifications. 35D05, 35K65

1. Introduction.

Let $\Omega \subset \mathcal{R}^N (N \geq 1)$ be a $C^{2+\alpha}$-bounded domain. In the physical situation that we have in mind, Ω represents an electric solid which is also conductor of heat. If $\varphi = \varphi(x,t), u = u(x,t)$ denote, respectively, the potential and temperature inside Ω, then we consider the following problem (P):

$$\nabla \cdot (\sigma(u) \nabla \varphi) = 0 \quad \text{in } \Omega_T \equiv \Omega \times (0,T), \tag{1.1}$$

$$u_t - \Delta u = \sigma(u) |\nabla \varphi|^2 \quad \text{in } \Omega_T, \tag{1.2}$$

$$\varphi = \varphi_0 \text{ on } \Gamma_D^\varphi \times (0,t), \quad \frac{\partial \varphi}{\partial n} = 0 \quad \text{on } \Gamma_N^\varphi \times (0,t), \tag{1.3}$$

$$u = 0, \text{ on } \Gamma_D^u \times (0,t), \quad \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma_N^u \times (0,t), \tag{1.4}$$

$$u(x,0) = u_0(x), \quad \text{in } \Omega, \tag{1.5}$$

where n is the outward normal to $\partial\Omega$, and $\Gamma_D^\varphi, \Gamma_N^\varphi, \Gamma_D^u, \Gamma_N^u$ are relatively open subsets with $C^{2+\alpha}(N-2)$-dimensional boundaries such that

$$\Gamma_N^\varphi \cap \Gamma_D^\varphi = \phi, \quad \Gamma_N^u \cap \Gamma_D^u = \phi, \quad \bar{\Gamma}_N^\varphi \cup \bar{\Gamma}_D^\varphi = \partial\Omega, \quad \bar{\Gamma}_N^u \cup \bar{\Gamma}_D^\varphi = \partial\Omega$$

and meas $\Gamma_D^\varphi > 0$, meas $\Gamma_D^u > 0$.

The system (1.1) and (1.2), with first boundary value, has been investigated by several authors (see [1], [3], [4], [8], and [9]). Here we obtain the existence and uniqueness of the solution for problems (1.1)–(1.5) by using the single-layer potential analysis and fixed-point argument. After presenting the definition of the weak solution for the problem (P) and some auxiliary lemmas in §2, we demonstrate our main result in §3.

*Received by the editors October 1, 1992; accepted for publication May 10, 1993.

†Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing, 100088, China

‡Department of Mathematics, Suzhou University, Suzhou, 215006, China

2. Formulation and statement of results. We shall assume

$$\varphi_0 = \varphi_0(x,t) \in C^{1+\alpha,0}\left(\bar{\Omega}_T\right) \quad (0<\alpha<1), \tag{2.1}$$

$$u_0(x) \in C^{\alpha}\left(\bar{\Omega}\right) \cap H^1(\Omega), \qquad u_0 = 0 \quad \text{on } \Gamma_D^u \times \{0\}, \tag{2.2}$$

$$\sigma(s) \in C^1(\mathcal{R}^1), \quad 0 < \sigma_* \le \sigma(s) \le \sigma^* < +\infty \quad \forall s \in \mathcal{R}^1. \tag{2.3}$$

Introduce

$$V = \{v \in H^1(\Omega); \quad v = 0 \text{ on } \Gamma_D^{\varphi}\},$$
$$U = \{v \in H^1(\Omega); \quad v = 0 \text{ on } \Gamma_D^{u}\}.$$

DEFINITION 2.1. We say that a pair $\{u, \varphi\}$ is a weak solution of problem (P) if

$$u \in L^2(0,T;U), \qquad u \in C^{\alpha,\alpha/2}\left(\bar{\Omega}_T\right), \tag{2.4}$$

$$\varphi - \varphi_0 \in C\left(\bar{\Omega}_T\right) \cap C(0,T;V), \tag{2.5}$$

and

$$\int_{\Omega} \sigma(u) \nabla\varphi \cdot \nabla\psi = 0 \quad \forall \psi \in V \quad \forall t \in [0,T], \tag{2.6}$$

(2.7)

$$\int_{\Omega_T} \left\{-u\frac{\partial\eta}{\partial t} + \nabla u \cdot \nabla\eta\right\} dx\,dt = \int_{\Omega_T} \sigma(u)\left|\nabla\varphi\right|^2 \eta dx\,dt + \int_{\Omega} u_0(x)\,\eta(x,0)\,dx,$$
$$\forall \eta \in W_2^{1,1}(\Omega_T) \cap L^{\infty}(\Omega_T), \quad \text{with } \eta = 0 \quad \text{on } \Gamma_D^u \times (0,T), \quad \text{on } \Omega \times \{T\}.$$

The main result of this paper is the following.

THEOREM 2.2. *Assume that* (2.1)–(2.3) *hold. Then problem* (P) *admits a weak solution. Moreover, in addition to* (2.1)–(2.3), *assume the open portion* Γ_D^{φ} *of* $\partial\Omega$ *be closed as well, then the weak solution is unique.*

The following two lemmas will be useful in the proof of Theorem 2.2.

LEMMA 2.3. *Let* $u \in C(\bar{\Omega}_T)$ *and* $\varphi - \varphi_0 \in L^2(0,T;V)$ *satisfying*

$$\int_{\Omega} \sigma(u) \nabla\varphi \cdot \nabla\psi\,dx = 0 \quad \forall \psi \in V, \quad \forall t \in [0,T].$$

Then

$$\|\varphi(\cdot,t)\|_{C^{\alpha}(\bar{\Omega})} \le C \quad \forall t \in [0,T], \tag{2.8}$$

$$\operatorname*{ess\,sup}_{0\le t\le T} \int_{\Omega\cap B(x_0,R)} \sigma(u)\left|\nabla\varphi\right|^2 dx \le CR^{N-2+2\alpha} \quad \forall x_0 \in \bar{\Omega}, \quad \forall R > 0, \tag{2.9}$$

where $B(x_0,R)=\{x\in\mathcal{R}^N;|x-x_0|<R\}$, $\alpha\in(0,\frac{1}{2})$ *and the constant* C *depends only on* $\sigma_*,\sigma^*,\|\varphi_0\|_{L^\infty(0,T;C^1(\bar\Omega))}$ *and the smoothness of* $\partial\Omega$. *Moreover,*

$$\varphi-\varphi_0\in C\left(\bar\Omega_T\right)\cap C\left(0,T;V\right). \tag{2.10}$$

The estimate (2.8) is the classical Hölder estimate (see [7]). The inequality (2.9) is just [2, §5, Lem. 3]. And the proof of (2.10) can be shown as in §3.

LEMMA 2.4. *Set*

$$\Gamma(x-\xi,t-\tau)=\begin{cases}\frac{1}{[4\pi(t-\tau)]^{N/2}}\exp\left[-\frac{|x-\xi|^2}{4(t-\tau)}\right] & \text{for } t>\tau,\\ 0 & \text{for } t\le\tau,\end{cases}$$

$$w_f(x,t)=\int_0^t\int_{\mathcal{R}^N}\Gamma(x-\xi,t-\tau)f(\xi,\tau)\,d\xi\,d\tau,$$

where $f\in L^\infty(0,T;L^1(\Omega))$, $f=0$ *outside* Ω_T *and*

$$\operatorname*{ess\,sup}_{0\le t\le T}\int_{\Omega\cap B(x_0,R)}|f(x,t)|\,dx\le C_1R^{N-2+2\alpha}\quad\forall x_0\in\bar\Omega,\quad\forall R>0. \tag{2.11}$$

Then there exist constants $\beta\in(0,1)$ *and* C *depending only on the constant* C_1 *in* (2.11) *and* $\partial\Omega$ *and* T *such that*

$$\|w_f\|_{C^{\beta,\beta/2}(\bar\Omega_T)}\le C.$$

This Hölder estimate is just [9, Lems. 3.1 and 3.2].

3. The proof of the main result. To prove Theorem 2.2, we need the following lemma.

LEMMA 3.1. *Let* $f\in C_0^\infty(\Omega_T)$ *and* (2.11) *hold. Assume* v *is the solution of the following problem:*

$$v_t-\Delta v=f\quad\text{in }\Omega_T, \tag{3.1}$$

$$\frac{\partial v}{\partial n}=0\quad\text{on}\quad\partial\Omega\times(0,T),\qquad v(x,0)=0\quad\text{in }\Omega. \tag{3.2}$$

Then there exist constants $\beta\in(0,1)$ *and* C *depending only on the constant* C_1 *in* (2.11) *and* T *and* $\partial\Omega$, *such that*

$$\|v_f\|_{C^{\beta,\beta/2}(\bar\Omega_T)}\le C.$$

Proof. Let $F(x,t)=\int_0^t\int_\Omega(\partial\Gamma(x-\xi,t-\tau)/\partial n)f(\xi,\tau)d\xi d\tau$ for all $(x,t)\in\partial\Omega\times(0,T]$, where n is the outward normal to $\partial\Omega\times\{t\}$ at the point (x,t). We claim that

$$\|F(x,t)\|_{L^\infty(\Omega_T)}\le C. \tag{3.3}$$

In fact, by using the following inequality

$$\left|\frac{\partial\Gamma(x-\xi,t-\tau)}{\partial n}\right|\le\frac{C}{(t-\tau)^\mu|x-\xi|^{N+1-2\mu-\beta}} \tag{3.4}$$

for $t > \tau, \xi \in \Omega, x \in \partial\Omega$, for all $\beta \in (0,1)$, for all $\mu \in (1-\beta/2, 1)$ (see [5], Chap. 5, (2.12)), we have

$$\begin{aligned}
|F(x,t)| &\le \sum_{i=0}^{\infty} \int_0^t \int_{B(x,2^{i+1})\setminus B(x,2^i)} \frac{C\,|f(\xi,\tau)|}{(t-\tau)^{\mu_1}\,|x-\xi|^{N+1-2\mu_1-\beta_1}}\,d\xi\,d\tau \\
&\quad + \sum_{i=0}^{\infty} \int_0^t \int_{B(x,2^{-i})\setminus B(x,2^{-i-1})} \frac{C\,|f(\xi,\tau)|}{(t-\tau)^{\mu_2}\,|x-\xi|^{N+1-2\mu_2-\beta_2}}\,d\xi\,d\tau \\
&\le C\sum_{i=0}^{\infty} \left(\frac{1}{2^i}\right)^{N+1-2\mu_1-\beta_1} (2^{i+1})^{N-2+2\alpha} \int_0^t \frac{d\tau}{(t-\tau)^{\mu_1}} \\
&\quad + C\sum_{i=0}^{\infty} (2^{i+1})^{N+1-2\mu_2-\beta_2} \left(\frac{1}{2^i}\right)^{N-2+2\alpha} \int_0^t \frac{d\tau}{(t-\tau)^{\mu_2}} \\
&\le CT,
\end{aligned}$$

where μ_i and $\beta_i (i = 1, 2)$ satisfy $\mu_i \in (1-\beta_i/2, 1), \beta_i \in (0,1)$ and

$$0 < 1 - \mu_1 + \frac{1-\beta_1}{2} < \alpha < 1 - \mu_2 + \frac{1-\beta_2}{2}.$$

Denote

$$\begin{aligned}
M_1 &= M(x,t;\xi,\tau) = 2\frac{\partial\Gamma(x-\xi,t-\tau)}{\partial n}, \\
M_{i+1}(x,t;\xi,\tau) &= \int_0^t \int_{\partial\Omega} M(x,t;y,\sigma)\, M_i(y,\sigma;\xi,\tau)\, dS_y\, d\sigma, \\
\psi(x,t) &= 2F(x,t) + 2\sum_{i=1}^{\infty} \int_0^t \int_{\partial\Omega} M_i(x,t;\xi,\tau)\, F(\xi,\tau)\, dS_\xi\, d\tau.
\end{aligned}$$

Recall that if $0 \le a < N-1, 0 \le b < N-1$, then

$$\int_{\partial\Omega} \frac{dS_y}{|x-y|^a\,|y-\xi|^b} \le \begin{cases} C\,|x-\xi|^{N-1-a-b} & (a+b > N-1), \\ C & (a+b < N-1). \end{cases} \tag{3.5}$$

By using (3.3)–(3.5) it follows from direct calculation that $\psi(x,t)$ is continuous on $\partial\Omega \times [0,T]$ and

$$\|\psi\|_{L^\infty(\partial\Omega\times[0,T])} \le C, \tag{3.6}$$

where C depends only on $\|F\|_{L_\infty(\Omega_T)}$, T and $\partial\Omega$ (see, e.g., [5, Chaps. 1 and 5]).

Let

$$\Psi(x,t) = \int_0^t \int_{\partial\Omega} \Gamma(x-\xi, t-\xi)\, \psi(\xi,\tau)\, dS_\xi\, d\tau.$$

If we have proved that

$$\|\Psi\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} \le C, \tag{3.7}$$

where C depends only on $\|\psi\|_{L^\infty(\partial\Omega\times[0,T])}$, T and $\partial\Omega$, then the assertion of Lemma 3.1 follows from the fact that $v(x,t) = \Psi(x,t) + w_f(x,t)$ in Ω_T, and also from Lemma 2.4.

It remains to show (3.7) holds. For all $x, y \in \mathcal{R}^N, x \neq y$, denote $d = |x-y|$.

$$
\begin{aligned}
&|\Psi(x,t) - \Psi(y,t)| \\
&\leq \int_0^t \int_{B(x,2d)\cap\partial\Omega} \Gamma(x-\xi, t-\xi)\,|\psi(\xi,\tau)|\,d\xi\,d\tau \\
&\quad + \int_0^t \int_{B(y,2d)\cap\partial\Omega} \Gamma(y-\xi, t-\xi)\,|\psi(\xi,\tau)|\,d\xi\,d\tau \\
&\quad + \int_0^t \int_{\mathcal{R}^N\setminus B(x+y/2,d)} |\Gamma(x-\xi,t-\xi) - \Gamma(y-\xi,t-\xi)|\,|\psi(\xi,\tau)|\,d\xi\,d\tau \\
&= I_1 + I_2 + I_3.
\end{aligned}
$$

Observe that

$$
|\Gamma(x-\xi, t-\tau)| \leq C\,|t-\tau|^{-\beta_1}\,|x-\xi|^{-2\beta_1+N} \qquad (t>\tau, \quad 0<\beta_1<1).
$$

Thus

$$
\begin{aligned}
I_1 &= \sum_{i=0}^{\infty} \int_0^t \int_{\partial\Omega\cap(B(x,d/2^{i-1})\setminus B(x,d/2^i))} \Gamma(x-\xi,t-\tau)\,|\psi(\xi,\tau)|\,d\xi\,d\tau \\
&\leq C\sum_{i=0}^{\infty} \left(\frac{2^i}{d}\right)^{N-2\beta_1} \left(\frac{d}{2^{i-1}}\right)^{N-1} \int_0^t \frac{d\tau}{(t-\tau)^{\beta_1}} \\
&\leq Cd^{2\beta_1-1},
\end{aligned}
$$

where $\beta_1 \in (\frac{1}{2}, 1)$. Similarly, $I_2 \leq Cd^{2\beta_1-1}$ holds.

To estimate I_3, recall that if $|\xi - z| \geq |x-y|, 0<\beta_2<1, t>\tau$, then

$$
|\Gamma(x-\xi,t-\tau) - \Gamma(y-\xi,t-\tau)| \leq \frac{C\,|x-y|}{(t-\tau)^{\beta_2}\,|\xi-z|^{N+1-2\beta_2}},
$$

where $z = (x+y)/2$. Therefore,

$$
\begin{aligned}
I_3 &\leq C\int_0^t \int_{(\mathcal{R}^N\setminus B(z,d))\cap\partial\Omega} \frac{|x-y|\,|\psi(\xi,\tau)|}{(t-\tau)^{\beta_2}\,|\xi-z|^{N+1-2\beta_2}}\,d\xi\,d\tau \\
&= C\sum_{i=0}^{\infty} \int_0^t \int_{\partial\Omega\cap(B(0,2^{i+1}d)\setminus B(0,2^i d))} \frac{d\,|\psi(z-\xi,\tau)|}{(t-\tau)^{\beta_2}\,|\xi|^{N+1-2\beta_2}}\,d\xi\,d\tau \\
&\leq Cd\sum_{i=0}^{\infty} \left(\frac{1}{2^i d}\right)^{N+1-2\beta_2} (2^{i+1}d)^{N-1} \\
&\leq Cd^{2\beta_2-1},
\end{aligned}
$$

where $\beta_2 \in (\frac{1}{2}, 1)$. So we obtain that

$$
|\Psi(x,t) - \Psi(y,t)| \leq C\,|x-y|^{\beta}. \tag{3.8}
$$

Now let $0 \leq t_2 < t_1 \leq T, 0<\gamma<1$.

$$
\begin{aligned}
|\Psi(x,t_1) - \Psi(x,t_2)| &\leq \int_0^{t_2} \int_{\partial\Omega} |\Gamma(x-\xi,t_1-\tau) - \Gamma(x-\xi,t_2-\tau)|\,|\psi(\xi,\tau)|\,d\xi\,d\tau \\
&\quad + \int_{t_2}^{t_1} \int_{\partial\Omega} \Gamma(x-\xi,t_1-\tau)\,|\psi(\xi,\tau)|\,d\xi\,d\tau \\
&\equiv I_1 + I_2.
\end{aligned}
$$

One can estimate T_2 by

$$\begin{aligned}
I_2 &= C\sum_{i=0}^{\infty}\int_{t_2}^{t_1}\int_{\partial\Omega\cap(B(x,2^{i+1})\setminus B(x,2^i))}\Gamma(x-\xi,t_1-\tau)\,|\psi(\xi,\tau)|\,d\xi\,d\tau\\
&\quad + C\sum_{i=0}^{\infty}\int_{t_2}^{t_1}\int_{\partial\Omega\cap(B(x,2^{-i})\setminus B(x,2^{-i-1}))}\Gamma(x-\xi,t_1-\tau)\,|\psi(\xi,\tau)|\,d\xi\,d\tau\\
&\le C\sum_{i=0}^{\infty}\left(\frac{1}{2^i}\right)^{N-2\beta_1}2^{(i+1)(N-1)}\int_{t_2}^{t_1}\frac{d\tau}{(t_1-\tau)^{\beta_1}}\\
&\quad + C\sum_{i=0}^{\infty}2^{(i+1)(N-2\beta_2)}\left(\frac{1}{2^i}\right)^{N-1}\int_{t_2}^{t_1}\frac{d\tau}{(t_1-\tau)^{\beta_2}}\\
&\le C\,|t_1-t_2|^{1-\beta_1}\sum_{i=0}^{\infty}2^{(2\beta_1-1)i} + C\,|t_1-t_2|^{1-\beta_2}\sum_{i=0}^{\infty}2^{(1-2\beta_2)i}\\
&\le C\,|t_1-t_2|^{1-\beta_2},
\end{aligned}$$

where $\beta_1\in(0,\frac{1}{2}),\beta_2\in(\frac{1}{2},1)$. Next notice that

$$\left|\frac{\partial\Gamma(x-\xi,s-\tau)}{\partial s}\right|\le\frac{C}{(s-\tau)^{1+\gamma}\,|x-\xi|^{N-2\gamma}}\qquad(s>\tau,\quad 0<\gamma<1).$$

Hence

$$\begin{aligned}
I_1 &\le \int_0^{t_2}d\tau\int_{t_2}^{t_1}ds\int_{\partial\Omega}\left|\frac{\partial\Gamma(x-\xi,s-\tau)}{\partial s}\right||\psi(\xi,\tau)|\,d\xi\,d\tau\\
&\le C\sum_{i=0}^{\infty}\int_0^{t_2}d\tau\int_{t_2}^{t_1}ds\int_{\partial\Omega\cap(B(x,2^{i+1})\setminus B(x,2^i))}\frac{|\psi(\xi,\tau)|}{(s-\tau)^{1+\beta_1}\,|x-\xi|^{N-2\beta_1}}\,d\xi\,d\tau\\
&\quad + C\sum_{i=0}^{\infty}\int_0^{t_2}d\tau\int_{t_2}^{t_1}ds\int_{\partial\Omega\cap(B(x,2^{-i})\setminus B(x,2^{-i-1}))}\frac{|\psi(\xi,\tau)|}{(s-\tau)^{1+\beta_2}\,|x-\xi|^{N-2\beta_2}}\,d\xi\,d\tau\\
&\le C\sum_{i=0}^{\infty}\left(\frac{1}{2^i}\right)^{N-2\beta_1}2^{(i+1)(N-1)}\int_0^{t_2}d\tau\int_{t_2}^{t_1}\frac{ds}{(s-\tau)^{1+\beta_1}}\\
&\quad + C\sum_{i=0}^{\infty}2^{(i+1)(N-2\beta_2)}\left(\frac{1}{2^i}\right)^{N-1}\int_0^{t_2}d\tau\int_{t_2}^{t_1}\frac{ds}{(s-\tau)^{1+\beta_2}}\\
&\le C\,|t_1-t_2|^{1-\beta_2},
\end{aligned}$$

where β_1 and β_2 satisfy $0<\beta_1<\frac{1}{2}<\beta_2<1$. So we have

$$|\Psi(x,t_1)-\Psi(x,t_2)|\le C\,|t_1-t_2|^{\beta/2}. \tag{3.9}$$

Therefore, (3.7) follows from (3.8) and (3.9). □

COROLLARY 3.2. *Let f be the same as in Lemma* 3.1, $u\in L^2(0,T;U)\cap L^\infty(0,T;L^2(\Omega))$. *If u satisfies*

$$\int_{\Omega_T}\left\{-u\frac{\partial\xi}{\partial t}+\nabla u\cdot\nabla\xi\right\}dx\,dt=\int_{\Omega_T}f\xi\,dx\,dt+\int_\Omega u_0(x)\,\xi(x,0)\,dx,$$

$$\forall\xi\in W_2^{1,1}(\Omega_T),\ \textit{with } \xi=0 \textit{ on } \Gamma_D^u\times[0,T] \textit{ and } \xi=0 \textit{ on } \Omega\times\{T\},$$

then

$$\|u\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} \leq C. \tag{3.10}$$

Here constants $\beta_1 \in (0,1)$ and C depends only on the constant C_1 in (2.11), $\|u_0\|_{C^\alpha(\bar{\Omega})}, N, T$ and $\partial\Omega$.

Proof. Decompose u into the sum of v and $\tilde{v}$, where v is the same as in Lemma 3.1, and $\tilde{v}$ is the solution of the following problem:

$$\tilde{v}_t - \Delta\tilde{v} = 0 \quad \text{in } \Omega_T, \tag{3.11}$$

$$\tilde{v} = -v, \quad \text{on } \Gamma_D^u \times [0,T], \qquad \frac{\partial\tilde{v}}{\partial n} = 0 \quad \text{on } \Gamma_N^u \times [0,T], \tag{3.12}$$

$$\tilde{v} = u_0(x) \quad \text{on } \Omega \times \{0\}. \tag{3.13}$$

Then Corollary 3.2 follows from both Lemma 3.1 and the Hölder estimate for the mixed boundary value problem (3.11)–(3.13) (see [6], §4, Thm. 4). □

Proof of Theorem 2.2. Introduce the Banach space $B = C^{\alpha,\alpha/2}(\bar{\Omega}_T)$ and the closed convex subset $K = \{u \in B; \|u\|_B \leq C\}$, where $\alpha = \beta/2$, and β and C are the same constants as in (3.10). Let $u \in K$ and $t \in [0,T]$. Denote by $\varphi_u = \varphi_u(\cdot,t)$ the unique solution to the following problem:

$$(\varphi_u - \varphi_0)(\cdot,t) \in V, \qquad \int_\Omega \sigma(u)\nabla\varphi_u \cdot \nabla\eta\, dx = 0 \quad \forall\eta \in V.$$

By Lemma 2.3, we obtain that

$$\|\varphi_u(\cdot,t)\|_{C^\alpha(\bar{\Omega})} \leq C, \tag{3.14}$$

$$\operatorname{ess}\sup_{0\leq t\leq T} \int_{\Omega\cap B(x_0,R)} \sigma(u)|\nabla\varphi_u|^2\, dx \leq CR^{N-2+2\alpha} \quad \forall x_0 \in \bar{\Omega}, \quad \forall R > 0. \tag{3.15}$$

Here C is independent of u and t, $C = C(\sigma_*, \sigma^*, \|\varphi_0\|_{L^\infty(0,T;C^\alpha(\bar{\Omega}))}, \partial\Omega)$.

Set

$$f_n = \begin{cases} \sigma(u)|\nabla\varphi_u|^2 & \text{in } \Omega_T^{(n)}, \\ 0 & \text{in } \mathcal{R}^{N+1} \setminus \Omega_T^{(n)}, \end{cases}$$

where $\Omega_T^{(n)} = \{(x,t) \in \Omega_T; \operatorname{dist}\{(x,t), \partial\Omega_T\} > 2/n\}$, and

$$f_{\varepsilon n}^{(u)} \equiv f_{\varepsilon n}^{(u)}(x,t) = \frac{1}{\varepsilon^{N+1}} \int_{\mathcal{R}^1}\int_{\mathcal{R}^N} \rho_N\left(\frac{x-y}{\varepsilon}\right)\rho_1\left(\frac{t-s}{\varepsilon}\right) f_n(y,s)\, dy\, ds,$$

where $\rho_N(\cdot)$ and $\rho_1(\cdot)$ are mollifiers in x and in t, respectively.

Let $v = v_{\varepsilon n} = v_{\varepsilon n}^{(u)}$ is the unique solution to the problem: Find $v \in L^2(0,T;U) \cap L^2(0,T;L^2(\Omega))$ such that

$$\int_{\Omega_T} \left\{-v\frac{\partial\xi}{\partial t} + \nabla v \cdot \nabla\xi\right\} dx\, dt = \int_{\Omega_T} f_{\varepsilon n}^{(u)}\xi dx\, dt + \int_\Omega u_0(x)\xi(x,0)\, dx, \tag{3.16}$$

$$\forall\xi \in W_2^{1,1}(\Omega_T), \text{ with } \xi = 0 \text{ on } \Gamma_D^u \times [0,T] \text{ and } \xi = 0 \text{ on } \Omega \times \{T\}.$$

By Corollary 3.2, there exists a constant C independent of ε and n, such that

$$\|v\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} \leq C. \tag{3.17}$$

So we can define a mapping $\Lambda : K \to K$ as follows: $v = \Lambda u$. Obviously the image ΛK is precompact. To show that the mapping Λ has a fixed point, we need only to prove Λ is continuous. Let $u_i \in K (i = 1, 2, \ldots)$ converge to u in B. Denote $v_i = \Lambda u_i$, and $v = \Lambda u$. Using (3.17) and choosing suitable test function in (3.16) we get

$$\operatorname{ess} \sup_{0 \leq t \leq T} \int_\Omega v_i^2 (x,t)\, dt + \iint_{\Omega_T} |\nabla v_i|^2 \, dx\, dt \leq C. \tag{3.18}$$

Here C is a constant independent of i, ε, and n. So a subsequence out of $\{v_i\}$ can be selected (and relabeled with i) such that

$$\begin{aligned} v_i &\to \tilde{v} \quad \text{in } C^{\alpha,\alpha/2} \left(\bar{\Omega}_T\right), \\ v_i &\rightharpoonup \tilde{v} \quad \text{in } L^2 (0,T;U) \cap L^2 (0,T; L^2 (\Omega)). \end{aligned}$$

If we can prove that there exists a subsequence of $\{\nabla \varphi_{u_i}\}$ such that

$$\nabla \varphi_{u_i} \to \nabla \varphi_u \quad \text{a.e. in } \Omega_T \, (\text{as } i \to \infty), \tag{3.19}$$

then

$$\begin{aligned} \int_{\Omega_T} \left\{ -\tilde{v} \frac{\partial \xi}{\partial t} + \nabla \tilde{v} \cdot \nabla \xi \right\} dx\, dt &= \int_{\Omega_T} f_{\varepsilon n}^{(u)} \xi \, dx\, dt + \int_\Omega u_0 (x)\, \xi (x,0)\, dx, \\ \forall \xi \in W_2^{1,1} (\Omega_T), \quad &\text{with } \xi = 0 \text{ on } \Gamma_D^u \times [0,T] \text{ and } \xi = 0 \text{ on } \Omega \times \{T\}. \end{aligned}$$

We must have $\tilde{v} \equiv v = \Lambda u$, and hence the sequence $\{v_i\}$ itself converges to v in B. Then the Schauder fixed-point theorem yields a solution $\{u_{\varepsilon n}, \varphi_{\varepsilon n}\}$ to the following problem:

$$\begin{aligned} (\varphi_{\varepsilon n} - \varphi_0)(\cdot, t) \in V, \qquad & \int_\Omega \sigma (u_{\varepsilon n}) \nabla \varphi_{\varepsilon n} \cdot \nabla \eta \, dx = 0 \quad \forall \eta \in V, \quad \forall t \in [0,T], \\ & u_{\varepsilon n} \in L^2 (0,T;U) \cap C^{\beta,\beta/2} \left(\bar{\Omega}_T\right), \end{aligned}$$

and

$$\begin{aligned} \int_{\Omega_T} \left\{ -u_{\varepsilon n} \frac{\partial \xi}{\partial t} + \nabla u_{\varepsilon n} \cdot \nabla \xi \right\} dx\, dt &= \int_{\Omega_T} f_{\varepsilon n}^{(u_{\varepsilon n})} \xi dx\, dt + \int_\Omega u_0 (x)\, \xi (x,0)\, dx, \\ \forall \xi \in W_2^{1,1} (\Omega_T), \text{ with } \xi = 0 \text{ on } \Gamma_D^u \times [0,T] &\text{ and } \xi = 0 \text{ on } \Omega \times \{T\}. \end{aligned}$$

By the estimates (3.14), (3.15), (3.17), and (3.18), there exists a subsequence out of $\{u_{\varepsilon n}, \varphi_{\varepsilon n}\}$ such that

$$\begin{aligned} u_{\varepsilon n} &\to u \quad \text{in } C^{\alpha,\alpha/2} \left(\bar{\Omega}_T\right), \\ \nabla u_{\varepsilon n} &\rightharpoonup \nabla u \quad \text{in } L^2 (\Omega_T). \end{aligned}$$

If

$$\nabla \varphi_{\varepsilon n} \to \nabla \varphi_u \quad \text{a.e. in } \Omega_T, \tag{3.20}$$

then it is easy to see that the pair $\{u, \varphi_u\}$ satisfy (2.4), (2.6), and (2.7). It remains to prove (3.19), (3.20), and (2.5), and it is enough to show the following propositions. □

PROPOSITION A. *Let $u \in C(\bar{\Omega}_T)$. For each $t_0 \in [0,T]$, we have*

$$\varphi_u(\cdot,t) \to \varphi_u(\cdot,t_0) \quad \text{in } C^0(\bar{\Omega}) \quad (\text{as } t \to t_0), \tag{3.21}$$

$$\nabla\varphi_u(\cdot,t) \to \nabla\varphi_u(\cdot,t_0) \quad \text{in } L^2(\Omega) \quad (\text{as } t \to t_0). \tag{3.22}$$

PROPOSITION B. *Let $\{u_i\}$ converges to u in $C^{\alpha,\alpha/2}(\bar{\Omega}_T)(i \to \infty)$, then*

$$\varphi_{u_i} \to \varphi_u \quad \text{in } L^2(\Omega_T) \quad (\text{as } i \to \infty), \tag{3.23}$$

$$\nabla\varphi_{u_i} \to \nabla\varphi_u \quad \text{in } L^2(\Omega_T) \quad (\text{as } i \to \infty). \tag{3.24}$$

Proof of Proposition A. Denote $\varphi = \varphi_u$ for simplicity. Let $\{t_n\} \subset [0,T], t_n \to t_0 (n \to \infty)$. From (3.15) and (3.16) it follows that there exists a subsequence $\{t_{n_k}\}$ and a function $\tilde{\varphi}(x) \in H^1(\Omega)$ such that

$$\varphi(x,t_{n_k}) \to \tilde{\varphi}(x) \quad \text{in } C^0(\bar{\Omega}), \quad \nabla\varphi(x,t_{n_k}) \to \nabla\tilde{\varphi}(x) \quad \text{weakly in } L^2(\Omega).$$

So $\tilde{\varphi}(x) = \varphi_0(x,t_0)$ on $\partial\Omega$, and for any $\eta \in V$,

$$\begin{aligned}
&\left|\int_\Omega \sigma(u(x,t_0))\nabla\tilde{\varphi}(x)\cdot\nabla\eta(x)\,dx\right| \\
&\quad\le \left|\int_\Omega (\sigma(u(x,t_0)) - \sigma(u(x,t_{n_k})))\nabla\varphi(x,t_{n_k})\cdot\nabla\eta(x)\,dx\right| \\
&\qquad + \left|\int_\Omega \sigma(u(x,t_0))\nabla(\tilde{\varphi}(x) - \varphi(x,t_{n_k}))\cdot\nabla\eta(x)\,dx\right| \to 0 \,(\text{as } k \to \infty).
\end{aligned}$$

We conclude that $\tilde{\varphi}(x) = \varphi(x,t_0)$ in $\bar{\Omega}$ and (3.21) follows.

For any $\eta \in V$ we have

$$\begin{aligned}
&\int_\Omega \sigma(u(x,t_n))\nabla(\varphi(x,t_n) - \varphi(x,t_0))\cdot\nabla\eta(x)\,dx \\
&\qquad + \int_\Omega [\sigma(u(x,t_n)) - \sigma(u(x,t_0))]\nabla\varphi(x,t_0)\cdot\nabla\eta(x)\,dx = 0.
\end{aligned}$$

Choose $\eta(x) = \varphi(x,t_n) - \varphi(x,t_0) - (\varphi_0(x,t_n) - \varphi_0(x,t_0)) \in V$ to obtain

$$\begin{aligned}
\int_\Omega |\nabla(\varphi(x,t_n) - \varphi(x,t_0))|^2 &\le C\int_\Omega |\sigma(u(x,t_n)) - \sigma(u(x,t_0))|^2 |\nabla\varphi(x,t_0)|^2 \\
&\quad + C\int_\Omega |\nabla\varphi(x,t_0)|^2 |\nabla(\varphi_0(x,t_n) - \varphi_0(x,t_0))|^2.
\end{aligned}$$

Here $C = C(\sigma_*, \sigma^*)$. Therefore, (3.22) is proved. □

Remark. By using Proposition A and (3.14) we deduce that $\varphi_u(x,t) \in C(\bar{\Omega}_T)$, and since $\varphi_u(\cdot,t) \in C^{1+\alpha}_{\text{loc}}(\Omega), \nabla_x\varphi_u(x,t)$ is a measurable function on Ω_t. Thus $\|\nabla\varphi_u\|_{L^2(\Omega_T)} \le C$.

Proof of Proposition B. For any $\eta \in V$ and any $t \in [0,T]$ we have

$$\int_\Omega \{\sigma(u_i)\nabla(\varphi_{u_i} - \varphi_u)\cdot\nabla\eta + (\sigma(u_i) - \sigma(u))\nabla\varphi_u\cdot\nabla\eta\}\,dx = 0.$$

Let $\eta(x) = \varphi_{u_i}(x,t) - \varphi_u(x,t) \in V$ to obtain

$$\int_\Omega |\nabla(\varphi_{u_i} - \varphi_u)(x,t)|^2 \, dx \le \frac{1}{\sigma_*^2} \int_\Omega |\sigma(u_i(x,t)) - \sigma(u(x,t))|^2 \, |\nabla\varphi_u(x,t)|^2 \, dx.$$

Therefore,

$$\|\nabla(\varphi_{u_i} - \varphi_u)\|_{L^2(\Omega_T)} \le \frac{1}{\sigma_*} \|(\sigma(u_i) - \sigma(u)) \nabla\varphi_u\|_{L^2(\Omega_T)} \to 0 \qquad (i \to \infty).$$

Proposition B follows.

Now the existence of the weak solution to problem (P) has been proved. For the uniqueness of the weak solution it is enough to notice that by the standard elliptic estimate

$$\varphi \in L^\infty\left(0,T; C^{1+\alpha}(\bar{\Omega})\right)$$

holds under the assumption of Theorem 2.2, and then we can proceed as in [3] to derive the uniqueness. The proof of Theorem 2.2 is completed. □

Acknowledgment. The authors are very grateful to Prof. Lishang Jiang for his guidance and encouragement.

REFERENCES

[1] S. N. Antontsev and M. Chipot, *The thermistor problem: existence, smoothness, uniqueness, and blow up,* SIAM J. Math. Anal., this issue, pp. 1128–1156.

[2] X. Chen, *Existence and regularity of solutions of a nonlinear degenerate elliptic system arising from a thermistor problem,* J. Partial Differential Equations, 7 (1994), pp. 19–34.

[3] M. Chipot and G. Cimatti, *A uniqueness result for the thermistor problem,* European J. Appl. Math., 2 (1991), pp. 97–103.

[4] G. Cimatti, *Existence of weak solutions for the nonstationary problem of the Joule heating of a conductor,* Ann. Mat. Pura. Appl. (4), 162(1992), pp. 33–42.

[5] A. Friedman, *Partial differential equations of parabolic type,* Prentice-Hall, Englewood Cliffs, NJ, 1964.

[6] G. M. Lieberman, *Mixed boundary value problems for elliptic and parabolic differential equations of second order,* J. Math. Anal. Appl., 113 (1986), pp. 422–440.

[7] G. Stampacchia, *Problemi al contorno ellittici comdatti discontinui dotati soluzioni holderiane,* Ann. Math. Pura. Appl., 51 (1960), pp. 1–32.

[8] G. Yuan, *Existence of a weak solution for the phase change problem with Joule's heating,* J. Partial Differential Equations, 7 (1994), pp. 35–48.

[9] ———, *Regularity of solutions of the nonstationary thermistor problem,* Applicable Analysis, to appear.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1167–1175, July 1994

008

THE QUALITATIVE ANALYSIS OF A DYNAMICAL SYSTEM MODELING THE FORMATION OF MULTILAYER SCALES ON PURE METALS*

H. C. AKUEZUE†, R. L. BAKER‡, AND M. W. HIRSCH§

Abstract. Gesmundo and Viani modeled the growth rates of three-oxide scales by the system

$$\frac{dq_1}{dt} = m\frac{K_1}{2q_1} - \frac{m-1}{m}\frac{K_2}{2q_2},$$
$$\frac{dq_2}{dt} = -m\frac{K_1}{2q_1} + \left(\frac{m-1}{m} + \frac{n}{m}\right)\frac{K_2}{2q_2} - \frac{n-1}{n}\frac{K_3}{2q_3},$$
$$\frac{dq_3}{dt} = -\frac{n}{m}\frac{K_2}{2q_2} + \frac{K_3}{2q_3}.$$

The authors consider the more general n-dimensional dynamical system:

$$\frac{dq_i}{dt} = -\sum_{j=1}^{n}\frac{a_{ij}}{q_j}, \qquad q_i(t) > 0, \quad i = 1, \ldots, n.$$

Under mild algebraic conditions on the constant matrix $A = (a_{ij})$, it is shown that every solution $\mathbf{q}(t)$ extends to a solution defined for all $t \geq 0$, and $\lim_{t\to+\infty} q_i(t) = +\infty$. The difference between any two solutions is bounded as a function of t. When A is an irreducible tridiagonal matrix, then every solution is eventually increasing.

It is shown that when $m, n > 1$ the Gesmundo–Viani system admits a unique parabolic solution $q_i(t) = c_i\sqrt{t}$. The authors conjecture that this parabolic solution attracts all other solutions.

Key words. differential equations, dynamical system, nonlinear dynamical system, cooperative system

AMS subject classifications. 34C35, 70K05

1. Introduction.

The parabolic growth of complex oxide scales containing two or three components of pure metals has been studied by Gesmundo and Viani [3]. They obtained the following nonlinear three-dimensional dynamical system as a model for the growth rates of three-oxide scales:

$$\frac{dq_1}{dt} = m\frac{K_1}{2q_1} - \frac{m-1}{m}\frac{K_2}{2q_2}, \tag{1.1.1}$$

$$\frac{dq_2}{dt} = -m\frac{K_1}{2q_1} + \left(\frac{m-1}{m} + \frac{n}{m}\right)\frac{K_2}{2q_2} - \frac{n-1}{n}\frac{K_3}{2q_3}, \tag{1.1.2}$$

$$\frac{dq_3}{dt} = -\frac{n}{m}\frac{K_2}{2q_2} + \frac{K_3}{2q_3}. \tag{1.1.3}$$

Here $K_i > 0$ $(i = 1, 2, 3)$ are rate constants, $m > 0, n > 0$ are parameters, and $q_i > 0$ is the weight of oxygen contained in oxide i per unit area.

In the present paper we study (1.1.1)–(1.1.3) as a case of the more general n-dimensional system:

$$\frac{dq_i}{dt} = -\sum_{j=1}^{n}\frac{a_{ij}}{q_j}, \qquad q_i(t) > 0, \quad i = 1, \ldots, n. \tag{1.2}$$

* Received by the editors May 22, 1991; accepted for publication (in revised form) May 26, 1993.
† IIT Research Institute, Chicago, Illinois 60616.
‡ Department of Mathematics, The University of Iowa, Iowa City, Iowa 52242.
§ Department of Mathematics, University of California, Berkeley, California 94720.

We show that under mild algebraic conditions on the $n \times n$ constant matrix $A = (a_{ij})$, in the long run the trajectories of (1.2) are well behaved in the sense that every solution $\mathbf{q}(t) = (q_1(t), \ldots, q_n(t)), t \in [0, a], 0 < a < +\infty$, can be extended to a solution on $[0, +\infty)$, and $\lim_{t \to \infty} q_i(t) = +\infty, i = 1, \ldots, n$. Moreover the difference between any two solutions is bounded as a function of t. Furthermore, if A is irreducible and tridiagonal then all solutions are eventually monotone increasing.

Finally, we give a partial qualitative analysis of (1.1) in the case where the parameters m, n have values in the interval $(1, +\infty)$: If $m, n > 1$ there is a unique parabolic solution $q_i(t) = c_i \sqrt{t}, c_i > 0, t > 0$. We conjecture that this parabolic solution attracts all other solutions of (1.1). See the conjecture of Gesmundo and Viani [3].

For the two-dimensional case of (1.1) analogous results were obtained by Baker and Akuezue in [1].

In obtaining our results we have made essential application of algebraic techniques: the Kamke–Müller comparison principle, and a theorem of Smillie concerning the long-run monotonicity of flows of cooperative tridiagonal dynamical systems.

The following notation will be used:

$$\mathbf{R}^n_+ = \{(q_1, \ldots, q_n) \in \mathbf{R}^n | q_i \geq 0, i = 1, \ldots, n\},$$
$$\mathbf{R}^n_{++} = \{(q_1, \ldots, q_n) \in \mathbf{R}^n | q_i > 0, i = 1, \ldots, n\}.$$
$\|\mathbf{x}\|$ is the Euclidean norm of the vector $\mathbf{x}$.

Here are the main results:

THEOREM I. *Assume that the $n \times n$ matrix $A = (a_{ij})$ in (1.2) satisfies the following four conditions:*

(a) $\det A \neq 0$ *and* $a_{ij} \geq 0$, *for* $i \neq j$;

(b) *A is irreducible*;

(c) *for all* $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{R}^n_+$, *if* $x_i \sum_{j=1}^n a_{ij} x_j = 0$ *for* $i = 1, \ldots, n$, *then* $\mathbf{x} = 0$;

(d) *every real eigenvalue of A is negative.*

Then every solution of (1.2) *of the form*

$$\mathbf{q} = (q_1, \ldots, q_n) : [0, a] \to \mathbf{R}^n_{++}, \qquad 0 < a < +\infty,$$

extends uniquely to a solution

$$\mathbf{q} : [0, +\infty) \to \mathbf{R}^n_{++},$$

and

$$\lim_{t \to +\infty} q_i(t) = +\infty, \qquad i = 1, \ldots, n.$$

Moreover, if $\mathbf{r}(t) = (r_1(t), \ldots, r_n(t)), t \in [0, +\infty)$, *is any other solution of* (1.2) *in* $\mathbf{R}^n_{++}$, *then*

$$\sup_{0 \leq t < +\infty} ||\mathbf{q}(t) - \mathbf{r}(t)|| < +\infty,$$

and hence

$$\lim_{t \to +\infty} \frac{q_i(t)}{r_i(t)} = 1, \qquad i = 1, \ldots, n.$$

Finally, if the matrix A is tridiagonal, then every solution $\mathbf{q}(t), t \in [0, +\infty)$, *of* (1.2) *in* $\mathbf{R}^n_{++}$ *is eventually monotone increasing on* $[0, +\infty)$.

We define a solution $\mathbf{q}(t)$ of (1.1) to be *parabolic* provided it is defined for all $t \geq 0$, takes values in $\mathbf{R}^3_+$, and has the following form:

$$\mathbf{q}(t) = (q_1(t), q_2(t), q_3(t)), \quad q_i(t) = c_i\sqrt{t}, \quad c_i > 0, \quad i = 1, 2, 3.$$

THEOREM II. *Assume that* $m, n > 1$ *in the dynamical system* (1.1). *Then every solution* $\mathbf{p} : [0, a] \to \mathbf{R}^3_{++}, 0 < a < +\infty$, *extends uniquely to a solution* $[0, +\infty) \to \mathbf{R}^3_{++}$ *such that* $\lim_{t\to+\infty} p_i(t) = +\infty, i = 1, 2, 3$, *and this solution is eventually increasing. There exists a unique parabolic solution*

$$\mathbf{q}(t) = (q_1(t), q_2(t), q_3(t)), \quad q_i(t) = c_i\sqrt{t}, \quad c_i > 0, \quad i = 1, 2, 3, \quad 0 \leq t < \infty.$$

If $\mathbf{p} : [0, +\infty) \to \mathbf{R}^3_{++}$ *is any other solution, then*

$$\sup_{0\leq t<+\infty} \|\mathbf{p}(t) - \mathbf{q}(t)\| < +\infty;$$

therefore,

$$\lim_{t\to+\infty} \frac{p_i(t)}{q_i(t)} = 1, \quad \lim_{t\to+\infty} \frac{p_i(t)}{p_j(t)} = \frac{c_i}{c_j}, \quad 1 \leq i, j \leq n.$$

Based on numerical exploration of the system (1.1), we present the following conjecture.

CONJECTURE. *If* $\mathbf{q} : [0, +\infty) \to \mathbf{R}^3_+$ *is the unique parabolic solution of system* (1.1), *and if* $\mathbf{p} : [0, +\infty) \to \mathbf{R}^3_{++}$ *is any other solution, then*

$$\lim_{t\to+\infty} \|\mathbf{p}(t) - \mathbf{q}(t)\| = 0.$$

2. Preliminaries. In this section we present background material that we will use in the proofs of Theorems I and II. We also introduce a change of variable that transforms system (1.2) into a more convenient form.

Let $F : \mathbf{W} \to \mathbf{R}^n$ be a continuously differentiable vector field on an open set $\mathbf{W} \subseteq \mathbf{R}^n_+$, and consider the system

$$\frac{dx_i}{dt} = F_i(x_1, \ldots, x_n), \qquad \mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{W}, \quad i = 1, \ldots, n. \tag{2.1}$$

DEFINITION 2.1. An $n \times n$ real matrix M is *irreducible* if for each distinct pair of indices i, j with $1 \leq i \neq j \leq n$, there exists a finite sequence $i = k_0, \ldots, k_m = j$ such that $M_{k_{r-1},k_r} \neq 0, r = 1, \ldots, m$.

DEFINITION 2.2. Let $E \subseteq \mathbf{W}$ be any subset, where $\mathbf{W}$ is given in (2.1). System (2.1) is called *cooperative* in E if $\partial F_i/\partial x_j(\mathbf{a}) \geq 0$ for $i \neq j, \mathbf{a} \in E$; *irreducible* in E if each matrix $(\partial F_i/\partial x_j(\mathbf{a})), \mathbf{a} \in E$, is irreducible; and *tridiagonal* in E if $\partial F_i/\partial x_j(\mathbf{a}) = 0$ for $|i - j| > 1, \mathbf{a} \in E$.

The next lemma is the Kamke–Müller comparison principle (the last statement proved in Hirsch [5]). The following notation is used: For vectors $\mathbf{x}, \mathbf{y}$ we write $\mathbf{x} < \mathbf{y}$ to mean $x_i < y_i$ for all i. We write $\mathbf{x} \leq \mathbf{y}$ if $x_i \leq y_i$ for all i. If $\mathbf{x} \leq \mathbf{y}$ but $\mathbf{x} \neq \mathbf{y}$ we write $\mathbf{x} \prec \mathbf{y}$.

LEMMA 2.3. (Müller [8], Kamke [7]). *Assume that the system* (2.1) *is cooperative in a convex subset* $E \subseteq \mathbf{W}$ *having nonempty interior. Let* $\mathbf{x}(t), \mathbf{y}(t)$ *be solutions in* E

of (2.1) *for* $a \le t \le b$ *where* $a < b$. *If* $\mathbf{x}(a) \le \mathbf{y}(a)$ *then* $\mathbf{x}(b) \le \mathbf{y}(b)$. *If* $\mathbf{x}(a) < \mathbf{y}(a)$ *then* $\mathbf{x}(b) < \mathbf{y}(b)$. *When the system is irreducible in* E, *if* $\mathbf{x}(a) \prec \mathbf{y}(a)$ *then* $\mathbf{x}(b) < \mathbf{y}(b)$.

The next lemma, due to Smillie [10] with improvements by Smith [11], demonstrates long-run monotonicity of solutions to irreducible cooperative tridiagonal systems:

LEMMA 2.4. *Suppose that system* (2.1) *is cooperative, irreducible and tridiagonal in a convex subset* $E \subseteq \mathbf{W}$ *having nonempty interior. Let* $\mathbf{x}(t)$ *be a solution in* E *of* (2.1) *on a maximal interval of the form* $[0, a), 0 < a \le \infty$. *Then each coordinate* $x_i(t)$ *is eventually monotone increasing or decreasing.*

We conclude this section by making the change of variable $p_i = 1/q_i$ in (1.2); this transforms (1.2) into a more convenient form:

$$\frac{dp_i}{dt} = p_i^2 \sum_{j=1}^{n} a_{ij} p_j = G_i(p_1, \ldots, p_n), \qquad p_i(t) > 0, \quad i = 1, \ldots, n. \tag{2.2}$$

Notice that the vector field $(G_1, \ldots, G_n)$ is defined in all of $\mathbf{R}^n$.

3. Proof of Theorem I. We begin with the following well-known consequence of the Perron–Frobenius theorem.

LEMMA 3.1. *Suppose that the* $n \times n$ *matrix* $A = (a_{ij})$ *satisfies* (a) *and* (b) *of Theorem* I. *Then* A *has an eigenvector* $\mathbf{v} > 0$ *and a simple real eigenvalue* $\alpha \ne 0$ *such that* $A\mathbf{v} = \alpha\mathbf{v}$ *and* $\alpha > \mathcal{R}e\beta$ *for all other eigenvalues* β.

The following standard property of cooperative systems is very useful in analyzing the geometry of phase portraits.

LEMMA 3.2. *Suppose that the system* (2.1) *is cooperative in* $\mathbf{R}^n_+$, *and that* $F(\mathbf{0}) = \mathbf{0}$. *Let* $\mathbf{x}(t), t \in [0, a], 0 < a < +\infty$, *be a solution of* (2.1) *in* $\mathbf{R}^n_+$ *such that* $F(\mathbf{x}(0)) < \mathbf{0}$. *Then this solution extends to a unique solution in* $\mathbf{R}^n_{++}$ *defined for* $t \in [0, +\infty)$. *Moreover,* $\mathbf{x}(t)$ *is strictly decreasing on* $[0, +\infty)$ *and converges to an equilibrium as* $t \to \infty$.

Outline of proof (compare Selgrade [9, Thm. 2.2] and Hirsch [5, Thm. 2.5]). It follows from the Kamke–Müller Comparison Theorem (Lemma 2.3) that $\mathbf{x}(t; \mathbf{x}_0)$ is strictly decreasing on $[0, a]$, and that $\mathbf{x}(0) \ge \mathbf{0}$. The usual compactness argument then implies that the solution extends over the whole positive half-line. Since each coordinate of the solution is decreasing and bounded below by 0, the solution converges, necessarily to an equilibrium. □

Up to a change of variables (i.e., $p_i = 1/q_i$), the next lemma amounts to a special case of Theorem I.

LEMMA 3.3. *Suppose the matrix* $A = (a_{ij})$ *satisfies* (a)–(d) *of Theorem* I. *Let* $\mathbf{p}(t)$ *be a solution of* (2.2) *defined for* $t \in [0, a], 0 < a < +\infty$, *taking initial value* $\mathbf{p}_0 \in \mathbf{R}^n_{++}$. *Then* $\mathbf{p}(t)$ *uniquely extends to a solution defined for all* $t \in [0, +\infty)$, *and* $\lim_{t \to +\infty} \mathbf{p}(t) = 0$. *Moreover, if* A *is tridiagonal then* $\mathbf{p}(t)$ *is eventually decreasing on* $[0, +\infty)$.

Proof. By standard theory $\mathbf{p}(t)$ extends to a solution defined on a maximal interval $[0, b), 0 < b \le \infty$. We need to prove $b = \infty$.

Let $\mathbf{v} > 0$ be the eigenvector of A given in Lemma 3.1 with simple real eigenvalue $\alpha \ne 0$, so that $A\mathbf{v} = \alpha\mathbf{v}$. Condition (d) of Theorem I implies that $\alpha < 0$. Define $c = 1 + \max_{1 \le i \le n} \{p_{0i}/v_i\}$, where $\mathbf{p}_0 = (p_{01}, \ldots, p_{0n})$ and $\mathbf{v} = (v_1, \ldots, v_n)$. Note that $c\mathbf{v} > \mathbf{p}_0$. With G_i given in (2.2), it is easy to see that $G_i(0) = 0$ and $G_i(c\mathbf{v}) = \alpha c^3 v_i^3 < 0$ for $i = 1, \ldots, n$.

By Lemma 3.2 there is a solution $\mathbf{r}(t), t \in [0, +\infty)$, of (2.2) in $\mathbf{R}^n_+$, having initial value $c\mathbf{p}_0$, which is strictly decreasing on $[0, +\infty)$. Hence $\lim_{t\to\infty} \mathbf{r}(t)$ exists as an equilibrium point of (2.2) in $\mathbf{R}^n_+$. Since condition (c) of Theorem I implies that $\mathbf{0}$ is the only equilibrium point in $\mathbf{R}^n_+$, therefore $\lim_{t\to+\infty} \mathbf{r}(t) = \mathbf{0}$.

Condition (a) of Theorem I implies that (2.2) is cooperative in $\mathbf{R}^n_+$. We have $\mathbf{r}(0) = c\mathbf{v} > \mathbf{p}_0 = \mathbf{p}(0)$, hence the Comparison Theorem (Lemma 2.3) shows that $\mathbf{r}(t) > \mathbf{p}(t)$ for all $t \in [0, b)$.

Because the identically zero solution of (2.2) exists we have $\mathbf{p}(t) > \mathbf{0}$, and therefore $\mathbf{0} < \mathbf{p}(t) < \mathbf{r}(t) < c\mathbf{v}$, for all $t \in [0, b)$. The usual compactness argument now shows that $b = \infty$. Lemma 2.3 implies $\mathbf{r}(t) > \mathbf{p}(t) > \mathbf{0}$; therefore, $\lim_{t\to+\infty} \mathbf{p}(t) = \mathbf{0}$.

The eventual monotonicity of solutions follows from Lemma 2.4. □

LEMMA 3.4. *Let the $n \times n$ matrix $A = (a_{ij})$ satisfy conditions* (a), (b), *and* (d) *of Theorem* I. *Let* $\mathbf{q}(t)$ *and* $\mathbf{r}(t), t \in [0, +\infty)$, *be solutions of* (1.2). *Assume* $\mathbf{q}(0) < \mathbf{r}(0)$. *Then*

$$\sup_{0\le t<+\infty} \|\mathbf{q}(t) - \mathbf{r}(t)\| < +\infty. \tag{3.1}$$

Proof. For $t \in [0, +\infty)$ and $i = 1, \dots, n$, define

$$s_i(t) = r_i(t) - q_i(t), \quad u_i(t) = \frac{1}{q_i(t)}, \quad w_i(t) = \frac{1}{r_i(t)}.$$

By Lemma 2.3, $\mathbf{q}(t) < \mathbf{r}(t)$, and thus $\mathbf{s}(t) > 0$, for $t \in [0, +\infty)$.

Next define $H_i(x_1, \dots, x_{3n}), i = 1, \dots, 3n$, by

$$H_i(x_1, \dots, x_{3n}) = \begin{cases} \sum_{j=1}^{n} a_{ij} x_j x_{j+n} x_{j+2n} & \text{if } 1 \le i \le n; \\ x_i^2 \sum_{j=1}^{n} a_{ij} x_{j+n} & \text{if } n+1 \le i \le 2n; \\ x_i^2 \sum_{j=1}^{n} a_{ij} x_{j+2n} & \text{if } 2n+1 \le i \le 3n. \end{cases}$$

If we identify $(\mathbf{s}, \mathbf{u}, \mathbf{w})$ with the vector $\mathbf{x} = (x_1, \dots, x_{3n})$, then the vector function $(\mathbf{s}(t), \mathbf{u}(t), \mathbf{w}(t)), t \in [0, +\infty)$ is a trajectory of the following cooperative dynamical system in $\mathbf{R}^{3n}_{++}$:

$$\frac{dx_i}{dt} = H_i(\mathbf{x}), \qquad \mathbf{x} \in \mathbf{R}^{3n}_{++}, \quad i = 1, \dots, 3n. \tag{3.2}$$

Let $\mathbf{v} = (v_1, \dots, v_n) > 0$ be the positive eigenvector of A given in Lemma 3.1, and let $\alpha \neq 0$ be the simple real eigenvalue of A given in that lemma such that $A\mathbf{v} = \alpha\mathbf{v}$. By condition (d) of Theorem I, $\alpha < 0$. Define $\mathbf{z} = (v_1^{-1}, \dots, v_n^{-1}, v_1, \dots, v_n, v_1, \dots, v_n)$. Let $c > 0$ be arbitrary.

It is easy to check that $H_i(c\mathbf{z}) < 0$ and $H_i(0) = 0, i = 1, \dots, 3n$. It follows from Lemma 3.2 that there exists a solution $\mathbf{x}(t)$ to equation (3.2) defined for all $t \in [0, +\infty)$, having initial value $c\mathbf{z}$ and taking values in $\mathbf{R}^{3n}_+$, such that $\mathbf{x}(t)$ is monotone decreasing on $[0, +\infty)$.

Now fix $c > 0$ so large that $c\mathbf{z} > (\mathbf{s}(0), \mathbf{u}(0), \mathbf{w}(0))$. Then the comparison principle implies $\mathbf{x}(t) > (\mathbf{s}(t), \mathbf{u}(t), \mathbf{w}(t))$ for all t. Therefore, because $\mathbf{x}(t)$ is monotone decreasing with initial value $c\mathbf{z}$, we see that $0 < \mathbf{r}(t) - \mathbf{q}(t) = \mathbf{s}(t) < c(v_1^{-1}, \dots, v_n^{-1})$ for all $t \in [0, +\infty)$. This proves that (3.1) holds. □

Proof of Theorem I. Suppose the matrix $A = (a_{ij})$ satisfies conditions (a)–(d) of Theorem I. Let $\mathbf{q}(t) = (q_1(t), \ldots, q_n(t)), t \in [0, a], 0 < a < +\infty$, be a solution of (2.1) in $\mathbf{R}^n_{++}$. Define $\mathbf{p}(t) = (p_1(t), \ldots, p_n(t)), t \in [0, a]$, by

$$p_i(t) = \frac{1}{q_i(t)}, \qquad i = 1, \ldots, n.$$

Then $\mathbf{p}(t), t \in [0, a]$, is a solution of (2.2) in $\mathbf{R}^n_{++}$, and hence Lemma 3.3 implies that this solution can be uniquely extended to a solution $\mathbf{p}(t), t \in [0, +\infty)$, of (2.2) in $\mathbf{R}^n_{++}$ such that $\lim_{t \to +\infty} \mathbf{p}(t) = 0$. Moreover, when A is tridiagonal then this extended solution is eventually monotone decreasing on $[0, +\infty)$.

We uniquely extend $\mathbf{q}(t), t \in [0, a]$, to all of $[0, +\infty)$ by the definition

$$q_i(t) = \frac{1}{p_i(t)}, \qquad i = 1, \ldots, n, \qquad t \in [0, +\infty).$$

Then this extension is a solution of (2.1) in $\mathbf{R}^n_{++}$ such that $\lim_{t \to +\infty} q_i(t) = +\infty, i = 1, \ldots, n$. If A is tridiagonal then the extended solution $\mathbf{q}(t)$ is easily seen to be eventually monotone increasing. This proves the first part of Theorem I.

To prove the second part of Theorem I let $\mathbf{r}(t), t \in [0, +\infty)$, be an arbitrary solution of (2.1) in $\mathbf{R}^n_{++}$. Let $\mathbf{v} > 0$ be a positive eigenvector of A as given in Lemma 3.1. Select $c > 0$ so small that $c^{-1}(v_1^{-1}, \ldots, v_n^{-1}) > \mathbf{q}(0), \mathbf{r}(0)$. Using the same argument that we used in the proof of Lemma 3.3, we see that (2.2) has a solution $\mathbf{p}(t)$ in $\mathbf{R}^n_{++}$ defined for $t \in [0, +\infty)$, with initial value $c\mathbf{v}$, which is strictly decreasing. Define $\mathbf{u}(t) = (u_1(t), \ldots, u_n(t)), t \in [0, +\infty)$, by

$$u_i(t) = \frac{1}{p_i(t)}, \qquad i = 1, \ldots, n.$$

Then $\mathbf{u}(t), t \in [0, +\infty)$, is a strictly increasing solution of (2.1) and $\mathbf{u}(0) > \mathbf{q}(0), \mathbf{r}(0)$. By Lemma 3.4 we have

$$\sup_{0 \le t < +\infty} \|\mathbf{q}(t) - \mathbf{r}(t)\| < +\infty, \qquad \sup_{0 \le t < \infty} \|\mathbf{r}(t) - \mathbf{u}(t)\| < +\infty.$$

This implies

$$\sup_{0 \le t < +\infty} \|\mathbf{q}(t) - \mathbf{r}(t)\| < +\infty.$$

For $i = 1, \ldots, n$, we have

$$\left| \frac{q_i(t)}{r_i(t)} - 1 \right| \le \frac{\|\mathbf{q}(t) - \mathbf{r}(t)\|}{r_i(t)}, \qquad t \in [0, +\infty).$$

By the first part of Theorem I, we have $\lim_{t \to +\infty} r_i(t) = +\infty$; it follows that

$$\lim_{t \to +\infty} \left| \frac{q_i(t)}{r_i(t)} - 1 \right| = 0, \qquad i = 1, \ldots, n.$$

This completes the proof of Theorem I. □

4. Proof Of Theorem II. In this section we prove Theorem II. We first prove that under the hypothesis of Theorem II, there exists a unique parabolic solution of system (1.1). We prove the remainder of Theorem II by an application of Theorem I.

LEMMA 4.1. *Let* $K_1, K_2, K_3 > 0$ *and* $m, n > 1$. *Define the* 3×3 *matrix* $A = (a_{ij})$ *as follows:*

$$A = \begin{pmatrix} -\frac{mK_1}{2} & \frac{m-1}{m}\frac{K_2}{2} & 0 \\ \frac{mK_1}{2} & -\left[\frac{m-1}{m} + \frac{n}{m}\right]\frac{K_2}{2} & \frac{n-1}{n}\frac{K_3}{2} \\ 0 & \frac{n}{m}\frac{K_2}{2} & -\frac{K_3}{2} \end{pmatrix}.$$

Then A *satisfies conditions* (a)–(d) *of Theorem* I.

Proof. Condition (a) and (b) of Theorem I are easily verified. Because $A = (a_{ij})$ is tridiagonal and $a_{ij} \neq 0$ whenever $|i - j| = 1$, condition (b) is immediate. Condition (c) follows from $\det A \neq 0$ and the fact that all the principal minors of A are nonzero. Hence, to prove the lemma, it suffices to show that (d) holds.

To verify (d) we apply Gershgorin's Circle Theorem (Golub and Van Loan [4]) to the transpose of A, concluding that the eigenvalues of A are contained in the union of the following three closed disks in the complex plane:

$$D_1 : \text{center } = -\frac{mK_1}{2}, \qquad \text{radius } = \frac{mK_1}{2},$$

$$D_2 : \text{center } = -\frac{m+n-1}{m}\frac{K_2}{2}, \qquad \text{radius } = \frac{|m-1|+n}{m}\frac{K_2}{2},$$

$$D_3 : \text{center } = -\frac{K_3}{2}, \qquad \text{radius } = \frac{n-1}{n}\frac{K_3}{2}.$$

Because $m > 1, n > 1$ all three disks are in the closed left half plane, so all eigenvalues have nonpositive real parts. Because A is invertible, all eigenvalues have negative real parts. Thus real eigenvalues are negative. □

For the remainder of this section we will assume that $K_i > 0, i = 1, 2, 3$ and that $m > 1, n > 1$.

The next lemma is a key to proving the existence of a parabolic solution of (1.1). The following notation is used. We denote by Δ^2 the *standard 2-simplex,* i.e., the set of all points $\mathbf{x} = (x_1, x_2, x_3) \in \mathbf{R}^3_+$ such that $\sum_{i=1}^3 x_i = 1$; we denote the boundary of Δ^2 by $\partial\Delta^2$. Define $e_1 = (1, 0, 0), e_2 = (0, 1, 0), e_3 = (0, 0, 1)$. For $1 \leq i, j \leq 3$, we let $[e_i, e_j]$ be the boundary simplex determined by the pair e_i, e_j, that is, $[e_i, e_j]$ is the convex hull of the pair e_i, e_j. Observe that $\partial\Delta^2$ is the union of all the boundaries $[e_i, e_j], i \neq j, 1 \leq i, j \leq 3$.

LEMMA 4.2. *Let* $f : \Delta^2 \to \Delta^2$ *be a continuous map which maps each vertex to itself and each edge into itself. Then* $f(\Delta^2) = \Delta^2$.

Proof. Standard theorems in algebraic topology show that any extension to the simplex, of a continuous map of the boundary of a simplex to itself having nonzero degree, must map onto the simplex. By looking at each edge it is easy to prove that the restriction of f to the boundary is a map of the boundary to itself which is

homotopic to the identity; it is well known that this implies degree 1. Therefore f is onto. □

Remark. In fact, one can prove that f is a homeomorphism of $\partial\Delta^2$.

To introduce the notation used in the next lemma, define B to be the following matrix:

$$B = \begin{pmatrix} -m & \dfrac{m-1}{m} & 0 \\ m & -\left[\dfrac{m-1}{m} + \dfrac{n}{m}\right] & \dfrac{n-1}{n} \\ 0 & \dfrac{n}{m} & -1 \end{pmatrix}.$$

Then B is invertible, hence we may define the 3×3 matrix $P = (P_{ij})$ to be $-\frac{1}{2}B^{-1}$. The matrices A and B are related by the equation $A = B$ diag (K_1, K_2, K_3). A calculation shows that $P_{ij} > 0$ for all i, j.

LEMMA 4.3. *Define* $f = (f_1, f_2, f_3) : \Delta^2 \to \Delta^2$ *by*

$$f_i(\mathbf{x}) = \left(x_i \sum_{j=1}^{3} P_{ij}x_j\right) \Bigg/ \left(\sum_{k=1}^{3} x_k \sum_{j=1}^{3} P_{kj}x_j\right), \quad \mathbf{x} = (x_1, x_2, x_3) \in \Delta^2, \quad i = 1, 2, 3.$$

Then f maps Δ^2 onto itself.

Proof. It is easy to check that f is continuous and maps each edge of the simplex into itself. Therefore Lemma 4.2 implies f is onto. □

LEMMA 4.4. *There exists a unique parabolic solution of* (1.1).

Proof. Uniqueness: Let $\mathbf{q}(t)$ and $\mathbf{r}(t)$ be two parabolic solutions of (1.1), with $q_i(t) = c_i\sqrt{t}, r_i(t) = d_i\sqrt{t}, c_i > 0, d_i > 0, (i = 1, 2, 3)$. By Lemma 4.1, the matrix A in (1.1) satisfies conditions (a)–(d) of Theorem I, hence by that theorem we have $1 = \lim_{t\to+\infty} q_i(t)/r_i(t) = c_i/d_i, (i = 1, 2, 3)$. Existence: Let K_i be as in (1.1) and define $\mathbf{y} \in \Delta^2$ by

$$y_i = K_i \Bigg/ \left(\sum_{j=1}^{3} K_j\right), \qquad i = 1, 2, 3.$$

Let $f : \Delta^2 \to \Delta^2$ be defined as in Lemma 4.3; then by that lemma there exists a point $\mathbf{u} \in \Delta^2$, such that $\mathbf{y} = f(\mathbf{u})$. Define ζ, η by

$$\zeta = \left(\sum_{j=1}^{3} K_j\right)^{1/2}, \qquad \eta = \left(\sum_{i=1}^{3} u_i \sum_{j=1}^{3} P_{ij}u_j\right)^{1/2}$$

Let $\mathbf{c} = (\zeta/\eta)\mathbf{x}$. Then $\mathbf{y} = f(\mathbf{u})$ implies $K_i = c_i \sum_{j=1}^{3} P_{ij}c_j$ for $i = 1, 2, 3$. This last set of equations is equivalent to

$$\frac{1}{2}\, c_i = -\sum_{j=1}^{3} a_{ij}\left(\frac{1}{c_j}\right), \qquad i = 1, 2, 3.$$

Define $\mathbf{q}(t), t \in (0, +\infty)$, by $q_i(t) = c_i\sqrt{t}, i = 1, 2, 3$. The preceding equations imply that $\mathbf{q}(t)$ is a parabolic solution of (1.1). This proves the lemma. □

Proof of Theorem II. To prove Theorem II, let $\mathbf{p}(t) = (p_1(t), p_2(t), p_3(t)), t \in [0, a], 0 < a < +\infty$, be a solution of (1.1) in $\mathbf{R}^3_{++}$. By Lemma 4.1, we may apply Theorem I to the system (1.1), hence there exists a unique extension of $\mathbf{p}(t), t \in [0, a]$, to a solution $\mathbf{p}(t) = (p_1(t), p_2(t), p_3(t)), t \in [0, +\infty)$, of (1.1) in $\mathbf{R}^3_{++}$ such that $\lim_{t\to+\infty} p_i(t) = +\infty, i = 1, 2, 3$. Because the matrix A of the system (1.1) is tridiagonal, Theorem I implies that this extended solution is eventually monotone increasing on $[0, +\infty)$. By Lemma 4.4, there exists a unique parabolic solution $\mathbf{q}(t) = \left(c_1\sqrt{t}, c_2\sqrt{t}, c_3\sqrt{t}\right), (c_1, c_2, c_3) > 0, t \in (0, +\infty)$, of (1.1) in $\mathbf{R}^3_{++}$; and by Theorem I, we have

$$\sup_{0\le t<+\infty} \|\mathbf{p}(t) - \mathbf{q}(t)\| < +\infty,$$

and hence

$$\lim_{t\to+\infty} \frac{p_i(i)}{q_i(t)} = 1, \qquad i = 1, 2, 3.$$

Therefore, if $1 \le i, j \le 3$, then

$$\lim_{t\to+\infty} \frac{p_i(t)}{p_j(t)} = \left(\frac{c_i}{c_j}\right) \lim_{t\to+\infty} \left(\frac{p_i(t)/q_i(t)}{p_j(t)/q_j(t)}\right) = \frac{c_i}{c_j}.$$

this completes the proof of Theorem II. □

REFERENCES

[1] R. BAKER, *The qualitative analysis of a system of differential equations arising from the study of two-layer scales on pure metals,* Pro. Amer. Math. Soc., to appear.

[2] A. DOLD, *Algebraic Topology,* Springer-Verlag, New York, 1980, p. 56.

[3] F. GESMUNDO AND F. VIANI, *The formation of multilayer scales in the parabolic oxidation of pure metals,* J. Corrosion Sci., 18 (1978), pp. 217–230.

[4] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations,* Johns Hopkins University Press, Baltimore, MD, 1983.

[5] M. W. HIRSCH, *Systems of differential equations that are competitive or cooperative* II: *Convergence almost everywhere,* SIAM J. Math. Anal., 16 (1985), pp. 423–439.

[6] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra,* Academic Press, New York, 1974.

[7] E. KAMKE, *Zur Theorie der System gewöhnlicher Differentialgleichungen* II, Acta Math., 58 (1932), pp. 57–85.

[8] M. MÜLLER, *Über das Fundamentaltheorem in der Theorie der gewöhnlichen Differentialgleichungen,* Math. Z., 26 (1926), pp. 619–645.

[9] J. SELGRADE, *Asymptotic behavior of solutions to single loop positive feedback systems,* J. Differential Equations, 38 (1980), pp. 80–103.

[10] J. SMILLIE, *Competitive and cooperative tridiagonal systems of differential equations,* SIAM J. Math. Anal., 5 (1984), pp. 530–534.

[11] H. L. SMITH, *Periodic tridiagonal competitive and cooperative systems of differential equations,* SIAM J. Math. Anal., 22 (1991), pp. 1102–1109.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1176–1180, July 1994

009

SPLAY-PHASE ORBITS FOR EQUIVARIANT FLOWS ON TORI*

RENATO E. MIROLLO†

Abstract. This paper studies dynamical systems on the n-fold torus equivariant under a cyclic permutation of coordinates. It is proved that under a mild condition, these systems have splay-phase solutions. These are periodic orbits in which the n coordinates are given by the same function of time, but equally separated in phase. Applications to systems of equations used to model Josephson junction arrays are discussed.

Key words. Josephson junctions, ponies on a merry-go-round, splay-phase orbits

AMS subject classification. 34C25

This work is part of an ongoing effort to apply the techniques of nonlinear systems to understand the behavior of Josephson junction arrays. Josephson junction arrays are superconducting electronic devices, capable of generating very high frequency voltage oscillations, up to 10^{11} Hz or more. We refer the reader to [4], [5], and [7]–[14] for more information about Josephson junctions. To mathematicians, the most important feature of the system of equations governing Josephson junction arrays is the high degree of symmetry present in the system. We shall prove that certain types of solutions to these equations, which we call splay-phase solutions, exist under a mild condition. Moreover, the existence of these solutions essentially follows from nothing more than the symmetry present in the Josephson junction equations. Hence our results apply to any system of differential equations possessing this symmetry. Accordingly, we will present our results in as general a setting as possible.

The simplified equations we studied in [12] have the form

$$\theta_i' = c_1 + c_2 \sin\theta_i + c_3 \sum_{j=1}^{n} \sin\theta_j, \tag{1}$$

where $i = 1, \ldots, n, c_1, c_2, c_3$ are constants and $(\theta_1, \ldots, \theta_n)$ is a point on the n-fold torus T^n. This system is equivariant under any permutation of the coordinates θ_i (we will explain precisely what this means below). We studied two types of solutions to (1) in [12]: in-phase and splay-phase solutions. In-phase solutions are, of course, solutions $(\theta_1(t), \ldots, \theta_n(t))$, where $\theta_i(t) = \theta_j(t)$ for all t. Splay-phase solutions are solutions of the form

$$\left(\phi(t), \phi\left(t + \frac{1}{n}T\right), \ldots, \phi\left(t + \frac{n-1}{n}T\right)\right), \tag{2}$$

where ϕ has period exactly T. In other words, the coordinates each have the same periodic behavior, but are equally staggered in phase. (For simplicity, we order the phase shifts to correspond to the ordering of the coordinates, but note that solutions to (1) are preserved by any permutation of the coordinates.) Other authors call these "wagon wheel" or "ponies on a merry-go-round" solutions [4], [5].

Aronson, Golubitsky, and Mallet-Paret prove splay-phase solutions exist for this system in [5]. Their proof uses functional analysis methods, and applies more generally to a system like (1) with second derivative terms. We shall give a proof that (1) has

* Received by the editors August 5, 1991; accepted for publication (in revised form) June 17, 1993. This research was supported in part by National Science Foundation grant DMS-8906423.
† Department of Mathematics, Boston College, Chestnut Hill, Massachusetts 02167.

splay-phase solutions using the Lefschetz trace formula. The only property of (1) that the proof relies on is that the system is equivariant under a cyclic permutation of coordinates.

Consider a system of differential equations on the torus T^n:

$$\theta_i' = f_i(\theta_1, \ldots, \theta_n). \tag{3}$$

Define σ: $T^n \to T^n$ by the rule $\sigma(\theta_1, \ldots, \theta_n) = (\theta_2, \ldots, \theta_n, \theta_1)$. Let F_t be the time t flow for (3). Then we say (3) is *σ-equivariant* if for all t, $F_t \circ \sigma = \sigma \circ F_t$. This means a curve $(\theta_1(t), \ldots, \theta_n(t))$ is an orbit for (2) if and only if $(\theta_2(t), \ldots, \theta_n(t), \theta_1(t))$ is. In terms of the components of the flow, σ-equivariance means

$$f_i(\theta_1, \ldots, \theta_n) = f_1(\theta_i, \ldots, \theta_n, \theta_1, \ldots, \theta_{i-1}) \quad \text{for } i = 2, \ldots, n. \tag{4}$$

The Josephson junction model (1) is an example of a σ-equivariant flow on T^n. Another example is a "ring of coupled oscillators" given by equations

$$\theta_i' = c_1 + c_2 \sin(\theta_{i+1} - \theta_i) + c_3 \sin(\theta_{i-1} - \theta_i), \tag{5}$$

where $i = 1, \ldots, n, c_1, c_2, c_3$ are constants and we interpret all indices mod n. (See [1]–[3] and [6] for a discussion of this example. Ermentrout found conditions for the stability of splay-phase solutions for this model in [6].) We now state and prove the theorem.

THEOREM. *Let $\theta_i' = f_i(\theta_1, \ldots, \theta_n)$ be a σ-equivariant flow on T^n. Suppose also that*

$$\sum_{i=1}^{n} f_i(\theta_1, \ldots, \theta_n) > 0 \quad \textit{for all } (\theta_1, \ldots, \theta_n) \in T^n. \tag{6}$$

Then this system has splay-phase orbits.

Remark 1. For the Josephson array (1), this condition is

$$nc_1 + (c_2 + nc_3) \sum_{j=1}^{n} \sin\theta_j > 0 \tag{7}$$

which is true for all $(\theta_1, \ldots, \theta_n) \in T^n$ exactly if $c_1 > |c_2/n + c_3|$. In the ring of oscillators (5), a necessary condition guaranteeing (6) is $c_1 > |c_2 - c_3|$. (This condition is sufficient when $n \equiv 0 \bmod 4$, or as $n \to \infty$.)

Remark 2. The Josephson junction model (1) can have splay-phase solutions even when condition (6) does not hold. See [12] for details.

Remark 3. Obviously a similar result holds if instead of (6) we assume

$$\sum_{i=1}^{n} f_i(\theta_1, \ldots, \theta_n) < 0 \quad \text{for all } (\theta_1, \ldots, \theta_n) \in T^n. \tag{8}$$

Remark 4. As mentioned above, the Josephson junction model (1) is of course equivariant under any permutation of the coordinates θ_i. This system is also invariant under time reversal, in the sense that $(\theta_1(t), \ldots, \theta_n(t))$ is an orbit if and only if $(\pi - \theta_1(-t), \ldots, \pi - \theta_n(-t))$ is. Our arguments do not rely on these symmetries (again, see [12] for more details).

Proof of the Theorem. Consider the $n-1$ torus $\Sigma \subset T^n$ given by $\sum_{i=1}^n \theta_i \equiv 0 \bmod 2\pi$. Since

$$\sum_{i=1}^{n} \theta_i' = \sum_{i=1}^{n} f_i(\theta_1, \ldots, \theta_n) > 0, \tag{9}$$

there is a well-defined Poincaré first return map on Σ which we denote by $\Phi : \Sigma \to \Sigma$. Actually, this is the only place we use condition (6). We make two claims, which suffice to prove the theorem. □

Claim 1. Suppose $p \in \Sigma$ satisfies $\Phi(p) = \sigma(p)$. Then the orbit of p is a splay-phase solution.

Claim 2. The equation $\Phi(p) = \sigma(p)$ has solutions $p \in \Sigma$.

Proof of Claim 1. Suppose $\Phi(p) = \sigma(p)$ for some $p \in \Sigma$. Let $T/n > 0$ be the time required for p's first return to Σ. Recall that we denote the time t flow map $F_t : T^n \to T^n$. Now

$$F_{T/n}(p) = \Phi(p) = \sigma(p), \tag{10}$$

and since $F_t \circ \sigma = \sigma \circ F_t$ for all t,

$$F_T(p) = \left(F_{T/n}\right)^n(p) = \sigma^n(p) = p. \tag{11}$$

Hence the orbit of p is periodic, with period T/m for some integer $m > 0$. Let

$$(\theta_1(t), \ldots, \theta_n(t)) = F_t(p),$$

$$\phi(t) = \theta_n(t). \tag{12}$$

Then since

$$F_{t+T/n}(p) = F_t(\sigma(p)) = \sigma(F_t(p)), \tag{13}$$

we see that

$$\theta_i(t) = \phi\left(t + \frac{T}{n} i\right), \qquad i = 1, \ldots, n. \tag{14}$$

It remains to prove that $m = 1$. Now

$$\sum_{i=1}^{n} \theta_i'(t) = \sum_{i=1}^{n} \phi'\left(t + \frac{T}{n} i\right), \tag{15}$$

so

$$\int_0^{T/n} \sum_{i=1}^{n} \theta_i'(t)\, dt = \int_0^T \phi'(t)\, dt. \tag{16}$$

Now

$$\int_0^{T/n} \sum_{i=1}^{n} \theta_i'(t)\, dt = 2\pi, \tag{17}$$

since T/n is the first return time for p back to Σ. Hence

$$\int_0^T \phi'(t)\, dt = 2\pi. \tag{18}$$

But if ϕ has period T/m, then $\int_0^T \phi'(t)dt$ is a multiple of $2\pi m$. Hence $m = 1$. □

Proof of Claim 2. We need to prove that the map $\sigma^{-1} \circ \Phi$ has a fixed point on Σ. We shall do this by applying the Lefschetz fixed point theorem to the map $\sigma^{-1} \circ \Phi$. First we show that the map $\Phi : \Sigma \to \Sigma$ is homotopic to the identity map. We exhibit the homotopy as follows. For any $p \in \Sigma$, let $t(p)$ be the time required for p to return to Σ. For $0 \leq s \leq 1$, let

$$(\theta_1(p,s), \ldots, \theta_n(p,s)) = F_{s \cdot t(p)}(p). \tag{19}$$

Then set

$$G_s(p) = (\theta_1(p,s), \ldots, \theta_{n-1}(p,s), -(\theta_1(p,s) + \cdots + \theta_{n-1}(p,s))). \tag{20}$$

For $s = 0, G_0(p) = p$, and for $s = 1, G_1(p) = F_{t(p)}(p) = \Phi(p)$, so G_s is the required homotopy.

Finally, we recall the Lefschetz theorem. If $f : X \to X$ is a continuous map on, for example, a compact n-dimensional manifold X, then the Lefschetz number of f is given by

$$\lambda(f) = \sum_{p=0}^{n} (-1)^p \operatorname{Tr}(f_p, H_p(X, \mathbf{Q})), \tag{21}$$

where f_p is the induced map on the homology group $H_p(X, \mathbf{Q})$, and Tr denotes trace. Of course, $\lambda(f)$ depends only on the homotopy class of f. Lefschetz's theorem says that if $\lambda(f) \neq 0$, then f has at least one fixed point. Under additional assumptions, one may also calculate the Lefschetz number of a map as the sum of local contributions. Suppose that f is a smooth map, with finitely many fixed points $p_1, \ldots, p_m$, all nondegenerate. Then

$$\lambda(f) = \sum_{k=1}^{m} i_{p_k}(f), \tag{22}$$

where the local index $i_{p_k}(f)$ is given by

$$i_{p_k}(f) = \operatorname{sgn} \det (I - df_{p_k}), \tag{23}$$

where df_{p_k} is the derivative of f on the tangent space $T_{p_k}X$.

In our situation, we wish to prove that the map $\sigma^{-1} \circ \Phi$ on Σ has a fixed point. Since Φ is homotopic to the identity map, $\sigma^{-1} \circ \Phi$ is homotopic to σ^{-1}, so it suffices to show that $\lambda(\sigma^{-1}) \neq 0$. And since $\lambda(\sigma) = \lambda(\sigma^{-1})$, we might as well work on σ. So σ has n fixed points on Σ, given by

$$\left(\frac{2\pi k}{n}, \ldots, \frac{2\pi k}{n}\right) \quad \text{for } k = 0, 1, \ldots, n-1. \tag{24}$$

They all have the same local structure, so it suffices to calculate the local index at $(0, \ldots, 0)$. In local coordinates $(\alpha_1, \ldots, \alpha_{n-1}) \to (\alpha_1, \ldots, \alpha_{n-1}, -(\alpha_1 + \cdots + \alpha_{n-1}))$ near $(0, \ldots, 0)$ on Σ, σ is given by the linear map

$$A(\alpha_1, \ldots, \alpha_{n-1}) = (\alpha_2, \ldots, \alpha_{n-1}, -(\alpha_1 + \cdots + \alpha_{n-1})). \tag{25}$$

The characteristic equation of this linear map is

$$\det(\mu I - A) = \frac{\mu^n - 1}{\mu - 1} = 1 + \mu + \cdots + \mu^{n-1}. \tag{26}$$

Hence $\det(I-A) = n$, so $\operatorname{sgn}\det(I-A) = 1$. Therefore $\lambda(\sigma) = n$, so we are done. In fact, it suffices to observe that $\det(I-A) \neq 0$ since all the fixed points of σ on Σ have the same local type. And $\det(I - A) \neq 0$ because the equation $A\mathbf{x} = \mathbf{x}, \mathbf{x} \in \mathbf{R^{n-1}}$ has no nonzero solutions. □

Concluding remarks. The next question to ask about splay-phase solutions is whether they are stable. Based on computer simulation, we believe that the splay-phase orbits in the Josephson junction model (1) are neutrally stable. We can only prove this analytically when $n = 2$ (see [12]). In general, this question remains open. However, Strogatz and the author have made some progress towards understanding the stability of splay states. We have analyzed an infinite-dimensional analog of the Josephson junction system, and calculated the Floquet multipliers around the splay states. Our calculations predict neutral stability for the system (1) considered in this paper. We conjecture that the splay states in the finite-dimensional Josephson junction system (1) have the same stability behavior as in our infinite-dimensional model. See [11].

Acknowledgments. We thank Masato Kuwata, Paul Schweitzer, Steve Strogatz, and Kwok Tsang for many helpful discussions. We are also grateful for the referees' helpful comments.

REFERENCES

[1] J. C. ALEXANDER, SIAM J. Appl. Math., 46 (1986), pp. 199–221.
[2] J. C. ALEXANDER AND G. AUCHMUTY, Arch. Rational Mech. Anal., 93 (1986), pp. 253–270.
[3] J. C. ALEXANDER AND B. FIEDLER, *Lecture Notes In Pure and Applied Mathematics (Differential Equations)*, No. 118, Marcel-Dekker, New York, 1987, pp. 7–26.
[4] D. G. ARONSON, M. GOLUBITSKY, AND M. KRUPA, Nonlinearity, 4 (1991), pp. 861–902.
[5] D. G. ARONSON, M. GOLUBITSKY, AND J. MALLET-PARET, Nonlinearity, 4 (1991), pp. 903–910.
[6] B. ERMENTROUT, J. Math Biology, 23 (1985), pp. 55–74.
[7] P. HADLEY AND M. R. BEASLEY, Appl. Phys. Lett., 50 (1987), pp. 621–623.
[8] P. HADLEY, M. R. BEASLEY, AND K. WIESENFELD, Appl. Phys. Lett., 52 (1988), pp. 1619–1621.
[9] ———, Phys. Rev. B, 38 (1988), pp. 8712–8719.
[10] S. NICHOLS AND K. WIESENFELD, Phys. Rev. A, 45 (1992), pp. 8430–8435.
[11] S. STROGATZ AND R. MIROLLO, Phys. Rev. E, 47 (1993), pp. 220–227.
[12] K. TSANG, R. MIROLLO, S. STROGATZ, AND K. WIESENFELD, Phys. D, 48 (1991), pp. 102–112.
[13] K. TSANG, S. H. STROGATZ, AND K. WIESENFELD, Phys. Rev. Lett., 66 (1991), pp. 1094–1097.
[14] K. TSANG AND I. B. SCHWARTZ, Phys. Rev. Lett., 68 (1992), pp. 2265–2268.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1181-1194, July 1994

010

A VELOCITY FUNCTIONAL FOR AN ANALYSIS OF STABILITY IN DELAY-DIFFERENTIAL EQUATIONS *

JAMES LOUISELL†

Abstract. The author considers the linear delay-differential system $(*)\dot{x}(t) = A_0x(t) + \sum_1^p A_ix(t-h_i)$. It is shown that there is a velocity functional which along with its Lie derivative is analogous in the theory of delay-differential operators to the velocity Lyapunov function $V(x) = \langle\langle Ax, Ax\rangle\rangle$, which along with its Lie derivative $x^T(A+A^T)x$ is used in the analysis of the ordinary differential equation $\dot{x}(t) = Ax(t)$. The Lie derivative can be written as $2\langle(A_0 + \sum_1^p \sigma_iA_i)\phi, \phi\rangle$ in a suitable inner product $\langle,\rangle$ for C^1 vector functions ϕ given over $[-\eta, 0]$, where the σ_i signify delay operators having length h_i and $\eta = \max(h_1, \ldots, h_p)$. Next considering the *nonlinear* delay-differential equation $(\dagger)\dot{x}(t) = f(x(t), x(t-h_1), \ldots, x(t-h_p))$, the author gives a velocity functional having Lie derivative which locally resembles that for the linear system (*). It is proven that if the linear operator associated with this Lie derivative everywhere satisfies a certain stability property, then the nonlinear system (†) will be globally contractive to a unique equilibrium.

Key words. delay-differential equation, velocity functional, nonlinear system, stability

AMS subject classifications. 34K, 58F

1. Introduction.

In this paper the author presents a functional that will be used to analyze the stability of delay-differential systems. Initially the analysis will be applied to systems that are linear. The space of initial data will in this case be the space $C[-\eta, 0]$ of continuous functions mapping the interval $[-\eta, 0]$ into R^n, where $\eta = \max(h_1, \ldots, h_p)$ and $h_1, \ldots, h_p$ are positive real numbers representing the delays. We will use the notation $(R^p)_+$ to denote the set of members of R^p having all components positive. After the nature of the functional we use is made clear, we will investigate systems that are nonlinear.

To begin, we consider the familiar linear delay-differential equation $(*)\dot{x}(t) = A_0x(t) + \sum_1^p A_ix(t-h_i)$, where $A_0, \ldots, A_p \in R^{n\times n}$, and $h = (h_1, \ldots, h_p)$ is any member of $(R^p)_+$. We can define the linear delay transformation L: $C[-\eta, 0] \to R^n$ by $L\phi = A_0\phi(0) + \sum_1^p A_i\phi(-h_i)$. For any trajectory $x(\cdot)$ of the system (*) and $t \geq 0$, we let x_t denote the member of $C[-\eta, 0]$ given by $x_t(u) = x(t+u)$ for $-\eta \leq u \leq 0$. Then $Lx_t = A_0x(t) + \sum_1^p A_ix(t-h_i)$, and the delay-differential equation $(*)$ is written as $(*)\dot{x}(t) = Lx_t$. We will frequently consider the extended vector $x_e(t) = (x(t), x(t-h_1), \ldots, x(t-h_p)) \in (R^n)^{p+1}$, and the member of $(R^n)^{p+1}$ having each of its $p+1$ n-tuples equal to one fixed vector x will be written as $x_c = (x, x, \ldots, x)$. We will write $B = [A_1 \cdots A_p]$, and for members $k = (k_1, \ldots, k_p)$ of $(R^p)_+$, we will often have occasion to consider the negative definite matrix $D(k) = \text{diag}\,(-k_1I_n, \ldots, -k_pI_n)$. Before presenting our first theorem on stability, we give a lemma on a matrix having a special form which will occur in our analysis.

LEMMA 1.1. *Let $A_0, \ldots, A_p \in R^{n\times n}$ and let $k \in (R^p)_+$. Consider the $n(p+1) \times$*

*Received by the editors June 23, 1992; accepted for publication (in revised form) June 4, 1993.

†Department of Mathematics, University of Southern Colorado, Pueblo, Colorado 81001. This research was completed while the author was a visitor in the Department of Mathematics of the University of Minnesota. The author is appreciative of the support provided by this institution during this period.

$n(p+1)$ *matrix*

$$S_k = \begin{bmatrix} A_0 + A_0^T + \sum_1^p k_i I & B \\ B^T & D(k) \end{bmatrix}.$$

If S_k is sign definite, then the matrix $A_0 + \sum_1^p A_i$ is nonsingular.

Proof. If $A_0 + \sum_1^p A_i$ is singular, then one has $x \in R^n - \{0\}$ with $(A_0 + \sum_1^p A_i)x = 0$. In this case, a calculation shows that $x_c^T(S_k)x_c = 0$. We thus see if S_k is sign definite that the matrix $A_0 + \sum_1^p A_i$ is nonsingular. □

We can now introduce the basic functional used in this paper, in the process giving a theorem on the stability of the linear system (*).

THEOREM 1.1. *Consider the linear delay-differential equation* (*)$\dot{x}(t) = A_0x(t) + \sum_1^p A_ix(t-h_i)$. *For each* $k \in (R^p)_+$, *consider the previously defined matrix* S_k *of Lemma* 1.1. *If there exists* $k \in (R^p)_+$ *such that* S_k *is negative definite, then the system* (*) *is exponentially asymptotically stable.*

Proof. Take any $k = (k_1, \ldots, k_p) \in (R^p)_+$, and let $x(\cdot)$ be any solution of the linear delay-differential system (*) having range in $\mathbb{C}^n$. Now define the real-valued scalar function $V(x_t)$ of t by

$$V(x_t) = (Lx_t)^*(Lx_t) + \sum_1^p \left(k_i \cdot \int_{t-h_i}^t \dot{x}^*(u)\dot{x}(u)du \right) \quad \text{for all } t \geq \eta,$$

where $(\cdot)^*$ denotes the conjugate transpose. Calculating the time derivative of this function, particularly noting that $Lx_t = \dot{x}(t)$, we find that

$$\dot{V}(x_t) = \left(\dot{x}^*(t)A_0^T + \sum_1^p \dot{x}^*(t-h_i)A_i^T \right) \dot{x}(t) + \dot{x}^*(t) \left(A_0\dot{x}(t) + \sum_1^p A_i\dot{x}(t-h_i) \right) + \sum_1^p k_i(\dot{x}^*(t)\dot{x}(t) - \dot{x}^*(t-h_i)\dot{x}(t-h_i)),$$

i.e., $\dot{V}(x_t) = \dot{x}_e^*(t)S_k\dot{x}_e(t)$ for $t \geq \eta$. If there exists $k \in (R^p)_+$ such that $\lambda_{\max}(S_k) < 0$, then set $\lambda_{\max}(S_k) = 2\gamma$ and obtain

$$\dot{V}(x_t) \leq 2\gamma \cdot |\dot{x}_e(t)|^2 \leq 2\gamma \cdot |\dot{x}(t)|^2 \quad \text{for } t \geq \eta.$$

Examining the expression $\dot{V}(x_t) \leq 2\gamma \cdot |\dot{x}(t)|^2$, we see that if there were $\varepsilon > 0$ such that $|\dot{x}(t)|^2 \geq \varepsilon$ over $[0, \infty)$, then one would have $\dot{V}(x_t) \leq 2\gamma \cdot |\dot{x}(t)|^2 \leq 2\gamma \cdot \varepsilon$ for $t \geq \eta$, and integrating $\dot{V}(x_\tau)$ for $\eta \leq \tau \leq t$, we would find that $V(x_t) - V(x_\eta) \leq 2\gamma\varepsilon \cdot (t - \eta)$ for $t \geq \eta$. Thus there would be some $t > \eta$ making $V(x_t) < 0$. Since $V(\cdot) \geq 0$, we see that it is not possible that there exists $\varepsilon > 0$ with $|\dot{x}(t)|^2 \geq \varepsilon$ over $[0, \infty)$.

If one now considers the characteristic function $g(s) = |sI - A_0 - \sum_1^p A_ie^{-sh_i}|$, one can see that if there were some zero $\lambda = \lambda_1 + i\lambda_2$ of $g(s)$ having $\lambda_1 \geq 0$ and $\lambda \neq 0$, then one solution of (*) would be the function $c(t) = e^{\lambda t}w$, where w is a nonzero member of the null space of the matrix $\lambda I - A_0 - \sum_1^p A_ie^{-\lambda h_i}$. Noting that $\dot{c}(t) = \lambda e^{\lambda t}w$, one sees that $|\dot{c}(t)|^2 = |\lambda w|^2 e^{2\lambda_1 t}$, and setting $\varepsilon = |\lambda w|^2$, one would have $|\dot{c}(t)|^2 \geq \varepsilon$ for all $t \geq 0$, which is impossible. Thus $\lambda_1 < 0$ if $\lambda \neq 0$ and $\lambda = \lambda_1 + i\lambda_2$ is a zero of $g(s)$. Furthermore, since S_k is negative definite, we know that the matrix $A_0 + \sum_1^p A_i$ is nonsingular, so that $\lambda = 0$ is not a root of $g(s)$. Thus for any zero

$\lambda = \lambda_1 + i\lambda_2$ of $g(s)$, we have $\lambda_1 < 0$. We have now proven that the delay-differential system $(*)\dot{x}(t) = A_0x(t) + \sum_1^p A_ix(t-h_i)$ is exponentially asymptotically stable. □

It is interesting to examine the rate of decay for the solutions of the system (*). In fact, if $\lambda = \lambda_1 + i\lambda_2$ is any zero of $g(s)$, then one has the solution $c(t) = e^{\lambda t}w$ as in the theorem, where w is an eigenvector of $\lambda I - A_0 - \sum_1^p A_ie^{-\lambda h_i}$. We then have $|\dot{c}(t)|^2 = |\lambda w|^2 e^{2\lambda_1 t}$, so that

$$V(c_t) = |\dot{c}(t)|^2 + \sum_1^p \left(k_i \cdot \int_{t-h_i}^{t} |\dot{c}(u)|^2 \, du \right) = |\lambda w|^2 e^{2\lambda_1 t} + |\lambda w|^2 \sum_1^p \left(k_i \cdot \int_{t-h_i}^{t} e^{2\lambda_1 u} du \right),$$

and after integrating, one obtains

$$V(c_t) = e^{2\lambda_1 t} |\lambda w|^2 \left(1 + \sum_1^p k_i \cdot \frac{1 - e^{-2\lambda_1 h_i}}{2\lambda_1} \right).$$

This yields $\dot{V}(c_t) = 2\lambda_1 e^{2\lambda_1 t} |\lambda w|^2 (1 + \sum_1^p k_i \cdot (1 - e^{-2\lambda_1 h_i})/2\lambda_1)$, i.e., $\dot{V}(c_t) = 2\lambda_1(1 + \sum_1^p k_i \cdot (1 - e^{-2\lambda_1 h_i})/2\lambda_1) |\dot{c}(t)|^2$ for $t \geq 0$.

If, as in the theorem, one has $\lambda_{\max}(S_k) = 2\gamma < 0$, we then have $\dot{V}(c_t) \leq 2\gamma \cdot |\dot{c}(t)|^2$, so that $2\lambda_1(1 + \sum_1^p k_i \cdot (1 - e^{-2\lambda_1 h_i})/2\lambda_1)|\dot{c}(t)|^2 \leq 2\gamma \cdot |\dot{c}(t)|^2$, and noting that $|\dot{c}(t)|^2 > 0$, we arrive at the inequality $\lambda_1(1 + \sum_1^p k_i \cdot (1 - e^{-2\lambda_1 h_i})/2\lambda_1) \leq \gamma$. This inequality is satisfied by the real part λ_1 of any zero of $g(s)$, and thus setting $\alpha = \sup \{\beta : g(s)$ has a zero with Re $(s) = \beta\}$, we find that $\alpha(1 + \sum_1^p k_i \cdot (1 - e^{-2\alpha h_i})/2\alpha) \leq \gamma$, i.e. $\alpha + \sum_1^p .5k_i(1 - e^{-2\alpha h_i}) \leq \gamma$.

It is also noteworthy that the condition that there exist $k \in (R^p)_+$ making the matrix $S_k < 0$ does not in any way depend on the vector $h = (h_1, \ldots, h_p)$, i.e., the type of stability given in the theorem does not depend on the delay durations. This phenomenon is known as stability independent of delay, and has been examined in other contexts by several authors [1], [7]. Considering the inequality $\alpha(1 + \sum_1^p k_i \cdot (1 - e^{-2\alpha h_i})/2\alpha) \leq \gamma$ satisfied by the real part of any eigenvalue of the system (*), it is interesting that this expression does involve the delay duration. In fact, if $h = 0$, we learn that $\alpha \leq \gamma$, giving a very simple link between the stability exponent and the maximum eigenvalue of S_k. On the other hand, for fixed $i \in \{1, \ldots, p\}$ having $h_i \to \infty$, the information immediately evident from the inequality more closely resembles the mere fact that $\alpha < 0$, i.e., that the system is asymptotically stable.

2. In this section we establish an analogy between the theorem in the previous section and a problem in linear ordinary differential equations. We begin by considering the ordinary differential equation $\dot{x}(t) = Ax(t)$, where A is any member of $R^{n \times n}$. For vectors $x, y \in R^n$, we write $\langle\langle x, y \rangle\rangle$ for the usual inner product of x and y. We can now introduce the quadratic form $V(x) = \langle\langle Ax, Ax \rangle\rangle$, and for each solution $x(\cdot)$ of $\dot{x}(t) = Ax(t)$, we consider the real function $V(x(t))$. Noting that $\dot{x} = Ax$ and examining the time derivative of this function, we obtain

$$\dot{V}(x) = (A\dot{x})^T\dot{x} + (\dot{x})^TA\dot{x} = (\dot{x})^TA^T\dot{x} + (\dot{x})^TA\dot{x} = (\dot{x})^T(A^T + A)\dot{x} = 2\langle\langle A\dot{x}, \dot{x} \rangle\rangle.$$

Via an analysis similar to that seen in the proof of the theorem from the previous section, one finds that if the matrix $A + A^T$ is negative definite, then all solutions of $\dot{x}(t) = Ax(t)$ converge to zero at an exponential rate as $t \to \infty$.

This fact from linear ordinary differential equations has been used as a starting point for an analysis of the global stability of certain nonlinear ordinary differential

equations. To give an example, Markus and Yamabe [11] investigate systems of the form $\dot{x}(t) = f(x(t))$, where f: $R^n \to R^n$ is continuously differentiable. They use the Lyapunov function $V(x) = \langle\langle f(x), f(x)\rangle\rangle$, just as the Lyapunov function $V(x) = \langle\langle Ax, Ax\rangle\rangle$ is used in the linear system above, and they consider the matrix $M(x) = f'(x) + f'(x)^T$, defined for each $x \in R^n$, just as the matrix $A + A^T$ is considered above. Making fairly mild assumptions on det $(M(x))$ and trace $(M(x))$, they prove that if $M(x)$ is negative definite for all $x \in R^n$, then the system has a unique equilibrium in R^n, and all trajectories of the system approach this equilibrium at an exponential rate as $t \to \infty$. For further applications of this method in the area of nonlinear ordinary differential equations, including applications to the problem of estimating stability basins, one can refer to the papers by Markus and Yamabe [11], Hill and Mareels [8], and to the book by Krasovskii [10].

Returning now to the linear delay-differential equation $(*)\dot{x}(t) = A_0x(t) + \sum_1^p A_ix(t - h_i)$, we recall that the space of initial data for this system is $C[-\eta, 0]$, where $\eta = \max(h_1, \ldots, h_p)$. For each $k \in (R^p)_+$, we can define a bilinear functional $\langle , \rangle_k$ on $C[-\eta, 0]$ by $\langle \phi, \psi \rangle_k = \phi^T(0)\psi(0) + \sum_1^p (k_i \cdot \int_{-h_i}^0 \phi^T(u)\psi(u)du)$ for $\phi, \psi \in C[-\eta, 0]$. This bilinear functional $\langle , \rangle_k$ is actually an inner product on the space $C[-\eta, 0]$, and $C[-\eta, 0]$ becomes a normed space with the norm $\|\phi\|_k = (\langle \phi, \phi \rangle_k)^{1/2}$. We now consider the space consisting of all continuously differentiable vector functions mapping $[-\eta, 0]$ into R^n, and denote this space by $C^1[-\eta, 0]$. We define a mapping A on that subspace of members ϕ of $C^1[-\eta, 0]$ having $A_0\phi(0) + \sum_1^p A_i\phi(-h_i) = \lim_{u \to 0-} \dot{\phi}(u)$ by the formula $\zeta = A\phi$, where $\zeta(0) = A_0\phi(0) + \sum_1^p A_i\phi(-h_i)$, and $\zeta(u) = \dot{\phi}(u)$ if $-\eta \leq u < 0$. Motivated by the above formula $\dot{V}(x) = 2\langle\langle A\dot{x}, \dot{x}\rangle\rangle$ occurring in linear ordinary differential equations, we examine the functional $\langle A\phi, \phi \rangle_k$:

$$\begin{aligned}
\langle A\phi, \phi \rangle_k &= \left(A_0\phi(0) + \sum_1^p A_i\phi(-h_i) \right)^T \phi(0) + \sum_1^p \left(k_i \cdot \int_{-h_i}^0 \dot{\phi}^T(u)\phi(u)du \right) \\
&= \phi^T(0)A_0^T\phi(0) + \sum_1^p \phi^T(-h_i)A_i^T\phi(0) \\
&\quad + \sum_1^p \left(\frac{k_i}{2}\phi^T(0)\phi(0) - \frac{k_i}{2}\phi^T(-h_i)\phi(-h_i) \right) \\
&= \frac{1}{2}\left[\phi^T(0)\left(A_0 + A_0^T + \sum_1^p k_iI \right)\phi(0) \right. \\
&\quad \left. + \sum_1^p (\phi^T(-h_i)A_i^T\phi(0) + \phi^T(0)A_i\phi(-h_i)) - \sum_1^p \phi^T(-h_i)(k_iI)\phi(-h_i) \right].
\end{aligned}$$

Recalling that $B = [A_1 \cdots A_p]$, and writing $\phi_e^T(0) = [\phi^T(0)\ \phi^T(-h_1) \cdots \phi^T(-h_p)]$, this acquires the matrix form

$$\langle A\phi, \phi \rangle_k = \frac{1}{2}\phi_e^T(0)\begin{bmatrix} A_0 + A_0^T + \sum_1^p k_iI & B \\ B^T & D(k) \end{bmatrix}\phi_e(0),$$

i.e., $2\langle A\phi, \phi \rangle_k = \phi_e^T(0)S_k\phi_e(0)$, where S_k is the matrix from Lemma 1.1 and Theorem 1.1.

Now inspecting the proof of Theorem 1.1 in light of this inner product $\langle \cdot, \cdot \rangle_k$, we see that we had written $V(x_t) = \langle Ax_t, Ax_t \rangle_k$ for solutions $x(\cdot)$ of (*) $\dot{x}(t) = A_0x(t) + \sum_1^p A_ix(t-h_i)$, and we had obtained $\dot{V}(x_t) = 2\langle A\dot{x}_t, \dot{x}_t \rangle_k$ for $t \geq \eta$, where $\dot{x}_t$ is the member of $C[-\eta, 0]$ defined by $\dot{x}_t(u) = \dot{x}(t+u)$ for $-\eta \leq u \leq 0$. Thus the quadratic functional $V(\phi) = \langle A\phi, A\phi \rangle_k$ is the analogue, in the area of delay-differential equations, of the symmetric quadratic form $V(x)$ given above for ordinary differential equations.

In the next section we give stability theorems for nonlinear delay-differential systems. These theorems are deduced from the global behavior of the *derivative* of the function defining system velocities, rather than from the behavior of values of this function itself. With this perspective, we first explore the asymptotic convergence of system trajectories, and after it is established that all trajectories with continuously differentiable initial data do converge, in fact exponentially, we investigate the properties of possible equilibria, proving a type of continuous dependence of the equilibria on the initial data. With this we will be able to arrive at the conclusion that there is just one equilibrium for any nonlinear system of the type under consideration.

The vector function defining system velocities will be of the form $f\colon (R^n)^{p+1} \to R^n$. This function will display a kind of decreasing behavior made formal by our assumption on its derivative, an assumption expressed in terms of global negativity of a quadratic functional related to the nonlinear system in the same manner as the quadratic functional $V(\phi)$ above is related to the linear system. Although, in this paper, we will derive the existence of a unique equilibrium and of exponentially converging trajectories, it may be worthwhile to note that there is considerable literature on the wider topic of nonlinear delay-differential equations for which each solution has some constant asymptotic limit.

An early example of this is given by Kaplan, Sorg, and Yorke [9]. These authors proved that if the function f is an order relation which is locally Lipschitz in the first coordinate, then all bounded trajectories of the scalar delay-differential system $\dot{x}(t) = f(x(t), x(t-h))$ do have asymptotic finite limits. In a later paper, Cooke, Kaplan, and Sorg [2] gave similar theorems which applied specifically to the stability of motion for a radiating charged particle. More recently, Haddock, Nkashama, and Wu [4] have examined linear neutral scalar Volterra systems having unbounded delay, giving a class of such systems for which, again, each solution has a constant asymptotic limit. In a paper making extensive use of invariance principles applied to systems with infinite delay, Haddock and Terjeki [5] arrive at theorems giving asymptotic constancy of all solutions of certain types of nonlinear functional differential equations for which each constant function is itself a solution. For an interesting narrative on this topic, one can refer to the paper by Haddock [3].

3. In this section we investigate the stability of certain nonlinear delay-differential systems in terms of their linearizations. We will consider nonlinear delay-differential equations of the form $(\dagger)\dot{x}(t) = f(x_e(t))$, where $x_e(t) = (x(t), x(t-h_1), \ldots, x(t-h_p))$, as usual, and $f\colon (R^n)^{p+1} \to R^n$ is continuously differentiable throughout $R^{n(p+1)}$. We can define the nonlinear delay transformation $F\colon C[-\eta, 0] \to R^n$ by $F(\phi) = f(\phi_e(0))$, and with this the system $(\dagger)$ is written as $(\dagger)\dot{x}(t) = F(x_t)$. Given any such system, it is known that for each $\phi \in C[-\eta, 0]$, there exists $\beta > 0$ and a unique $x(\cdot) = x(\phi, \cdot)$ with $x(u) = \phi(u)$ for $-\eta \leq u \leq 0$, which satisfies $(\dagger)$ over $[0, \beta)$. It is known [6] that if there exists $\tilde{\beta} \geq \beta$ such that $x(\cdot) = x(\phi, \cdot)$ is a noncontinuable solution over the interval $[0, \tilde{\beta})$, then the solution $x(\cdot)$ is unbounded in R^n over $[0, \tilde{\beta})$. For $v_0, \ldots, v_p \in R^n$, we will write the derivative of f at $v = (v_0, \ldots, v_p)$ as $f'(v) = [A_0(v) \cdots A_p(v)]$, or merely

as $f'(v) = [A_0 \cdots A_p]$, where $A_i \in R^{n\times n}$ for $i = 0, \ldots, p$. Before proceeding, we prove the following facts from linear algebra which will be valuable in the stability analysis of the nonlinear system (†).

LEMMA 3.1. *Let*

$$J = \begin{bmatrix} A & C \\ C^T & E \end{bmatrix},$$

where $A = A^T \in R^{n\times n}, C \in R^{n\times np}$, *and* $E \in R^{np\times np}$. *Let* $E = \text{blockdiag}\,(E_1, \ldots, E_p)$, *with each* $E_i \in R^{n\times n}$. *If there exists* $i \in \{1, \ldots, p\}$ *such that* $E_i = 0$, *then* J *is not strictly sign definite.*

Proof. Suppose we have $i \in \{1, \ldots, p\}$ with $E_i = 0$. Consider any member $u = (u_1, \ldots, u_p)$ of $(R^n)^p$ having $u_j = 0$ for $j \neq i$, and having u_i nonzero. Then

$$[0 \quad u^T]J\begin{bmatrix} 0 \\ u \end{bmatrix} = [0 \quad u^T]\begin{bmatrix} A & C \\ C^T & E \end{bmatrix}\begin{bmatrix} 0 \\ u \end{bmatrix} = u^T E u = 0.$$

Since

$$[0 \quad u^T]J\begin{bmatrix} 0 \\ u \end{bmatrix} = 0$$

and $u \neq 0$, we see that J is not strictly sign definite. □

LEMMA 3.2. *Let* $k = (k_1, \ldots, k_p) \in (R^p)_+$. *Let*

$$J_k = \begin{bmatrix} A & C \\ C^T & D(k) \end{bmatrix},$$

where $A = A^T \in R^{n\times n}, C \in R^{n\times np}$, *and* $D(k)$ *is as previously defined. Then for each* $i = 1, \ldots, p$, *we have* $\lambda_{\max}(J_k) + k_i \geq 0$.

Proof. For each $i = 1, \ldots, p$, set $L_i = J_k + k_i I$, where I denotes $I_{n(p+1)}$. Now note from Lemma 3.1 that L_i is not sign definite, so that $\lambda_{\max}(L_i) \geq 0$. Since $\lambda_{\max}(L_i) = \lambda_{\max}(J_k) + k_i$, we see that $\lambda_{\max}(J_k) + k_i \geq 0$, and the lemma is proven. □

In the next three lemmas we provide a basis for comparing the solutions of a type of differential inequality with delay to the solutions of a corresponding delay-differential equation. This will be valuable in establishing an exponential decay rate for the derivatives of solutions of the nonlinear systems we will eventually analyze. Before giving the first of these lemmas, we introduce the notation $C_-[-\eta, 0]$ to denote the set of all functions ψ: $[-\eta, 0] \to R$ for which both (i) ψ is continuous over the half-open interval $[-\eta, 0)$, and (ii) $\lim_{u\to 0-}\psi(u)$ exists and is finite. Here it is not required that $\psi(0) = \lim_{u\to 0-}\psi(u)$. It will be useful to consider the norm $\|\cdot\|_-$ defined on this space by $\|\psi\|_- = (|\psi(0)|^2 + \int_{-\eta}^0 |\psi(u)|^2 du)^{1/2}$, i.e., $(\|\psi\|_-)^2 = |\psi(0)|^2 + (\|\psi\|_2)^2$, where $\|\psi\|_2$ is the norm of ψ in $L^2(-\eta, 0)$.

LEMMA 3.3. *Let* $\eta = \max\,(h_1, \ldots, h_p)$, *where each* $h_i > 0$, *let* $d > 0$, *and let* m: $[-\eta, d) \to R$ *have* $m_0 \in C_-[-\eta, 0]$, *i.e., suppose that* m *is continuous over* $[-\eta, 0)$, *and that* $\lim_{u\to 0-} m(u)$ *exists and is finite. Let* $a < 0$, *let* $b_1, \ldots, b_p$ *be constants with each* $b_i \geq 0$, *and suppose that* $m(\cdot)$ *is continuous and right differentiable over* $[0+, d)$, *with right derivative* m'_+ *satisfying the delay-differential inequality* $m'_+(t) \leq am(t) + \sum_1^p b_i m(t - h_i)$. *Now let* $n(\cdot)$ *be the solution over* $[0, \infty)$ *to the delay-differential* equation $n'(t) = an(t) + \sum_1^p b_i n(t - h_i)$ *having initial data* $n_0 = m_0 \in C_-[-\eta, 0]$. *Then* $m(t) \leq n(t)$ *for* $0 \leq t < d$.

Proof. Define f: $[-\eta, d) \to R$ by $f = m - n$. Then $f = 0$ over $[-\eta, 0)$, and $f(0) = 0$. Noting over $[0, d)$ that $m'_+(t) \leq am(t) + \sum_1^p b_i m(t - h_i), n'(t) = an(t) + \sum_1^p b_i n(t - h_i)$, we have $f'_+(t) \leq af(t) + \sum_1^p b_i f(t - h_i)$ for $0 \leq t < d$.

Now suppose there was a point in $[0,d)$ at which f attained a value greater than zero. Then let $\alpha = \inf\{t : f(t) > 0\}$, and note $\alpha \geq 0$. By definition of α, there would exist $\delta, \tilde{t}$ having $0 < \delta < \min(h_1, \ldots, h_p)$ and $\tilde{t} \in (\alpha, \alpha+\delta)$, with $f(\tilde{t}) > 0$. Now let $\tilde{\alpha} = \inf\{\tau : f(t) > 0 \text{ for all } t \in [\tau, \tilde{t}]\}$, and note that $f(\tilde{\alpha}) = 0$. Let t' be any member of $(\tilde{\alpha}, \tilde{t})$ having $f(t') < f(\tilde{t})$, and note that there would be some $t_0 \in (t', \tilde{t})$ with $f'_+(t_0) > 0$. Since $f(t) \leq 0$ for $t \leq \alpha$, and since $\alpha + \delta < \alpha + h_i$ for each i, we know for $i = 1, \ldots, p$ that $f(t_0 - h_i) \leq 0$. Furthermore, since $t_0 \in (\tilde{\alpha}, \tilde{t})$, we know that $f(t_0) > 0$. This would then yield $0 < f'_+(t_0) \leq af(t_0) + \sum_1^p b_i f(t_0 - h_i)$, a contradiction in light of the signs of $a, f(t_0)$, the b_i and the $f(t_0 - h_i)$. We now conclude that there is no point in $[0,d)$ at which f attains a positive value. Thus $f(t) \leq 0$ over $[0,d)$, i.e., $m(t) \leq n(t)$ for $0 \leq t < d$. □

LEMMA 3.4. *Let $a < 0, |a| > \sum_1^p |b_i|$, and consider the delay-differential equation $(\S) n'(t) = an(t) + \sum_1^p b_i n(t - h_i)$. Then there exist constants $\alpha < 0, C > 0$ such that the following holds: For each $\psi \in C_-[-\eta, 0]$, the solution $n(\psi, \cdot)$ of $(\S)$ satisfies $|n(\psi, t)| \leq (C\|\psi\|_-)e^{\alpha t}$ for all $t \geq 0$.*

Proof. Recalling the characteristic function $g(s) = s - a - \sum_1^p b_i e^{-sh_i}$, we know for $\operatorname{Re}(s) \geq 0$ that $|s - a| \geq |a| > \sum_1^p |b_i| \geq \sum_1^p |b_i e^{-sh_i}|$, and thus $|g(s)| \geq ||s - a| - |\sum_1^p b_i e^{-sh_i}|| \geq |a| - \sum_1^p |b_i| > 0$, i.e., $|g(s)| > 0$ for $\operatorname{Re}(s) \geq 0$. Noting the well-known fact that $g(s)$ has at most a finite number of zeros to the right of any vertical line in $\mathbb{C}$, we see, since $g(s)$ has no zeros in $\{\operatorname{Re}(s) \geq 0\}$, that there exists $\alpha < 0$ such that $g(s)$ has no zeros in $\{\operatorname{Re}(s) \geq \alpha\}$. The lemma now follows from well-established facts [6] in the theory of autonomous linear functional differential equations. □

LEMMA 3.5. *Let $\eta = \max(h_1, \ldots, h_p)$, where each $h_i > 0$. Let $a < 0$, again let $b_1, \ldots, b_p$ be nonnegative constants, and suppose $|a| > \sum_1^p b_i$. Then there exist constants $C > 0, \alpha < 0$ such that the following holds: If $d > 0$ and $m: [-\eta, d) \to [0, \infty)$ is any function having $m_0 \in C_-[-\eta, 0]$ and satisfying the delay-differential inequality $m'_+(t) \leq am(t) + \sum_1^p b_i m(t - h_i)$ for $0 \leq t < d$, then $0 \leq m(t) \leq (C\|m_0\|_-)e^{\alpha t}$ for $0 \leq t < d$.*

Proof. From Lemma 3.3, we know that $0 \leq m(t) \leq n(t)$ for $0 \leq t < d$, where $n(\cdot)$ is the solution over $[0, \infty)$ to the delay-differential equation $n'(t) = an(t) + \sum_1^p b_i n(t - h_i)$ having initial data $n_0 = m_0$. Since, by Lemma 3.4, we have $|n(t)| \leq (C\|m_0\|_-)e^{\alpha t}$ for $t \geq 0$, we see that $0 \leq m(t) \leq n(t) \leq (C\|m_0\|_-)e^{\alpha t}$ for $0 \leq t < d$, and the lemma is proven. □

We are now prepared to show how the functional from the preceding sections can be used in the analysis of nonlinear delay-differential systems. It is convenient to recall here that $C^1[-\eta, 0]$ denotes the space of all functions $\phi: [-\eta, 0] \to R^n$ which have continuous derivative $\dot{\phi}$ over $[-\eta, 0]$. For reasons involving the continuation of solutions, the remaining theorems will be stated for delay-differential equations having initial data in this space $C^1[-\eta, 0]$.

THEOREM 3.1. *Consider the delay-differential equation $(\dagger)\dot{x}(t) = f(x_e(t))$, where f is continuously differentiable throughout $R^{n(p+1)}$. For $v_0, \ldots, v_p \in R^n$ and $v = (v_0, \ldots, v_p)$, set $f'(v) = [A_0 \cdots A_p]$. Write $B = [A_1 \cdots A_p]$, and suppose that there exist $\gamma < 0, k \in (R^p)_+$ such that for each $v \in R^{n(p+1)}$, the matrix*

$$S_k = \begin{bmatrix} A_0 + A_0^T + \sum_1^p k_i I & B \\ B^T & D(k) \end{bmatrix}$$

satisfies $\lambda_{\max}(S_k) \leq 2\gamma$. Then for each $\phi \in C^1[-\eta, 0]$, the solution $x(\phi, \cdot)$ is defined over $[0, \infty)$, and in fact there exists $\bar{x} \in R^n$ such that $x(\phi, t)$ converges at an

exponential rate to $\bar{x}$ as $t \to \infty$. Furthermore, for $\bar{x}_c = (\bar{x}, \ldots, \bar{x})$, we have $f(\bar{x}_c) = 0$.

Proof. Take any $\phi \in C^1[-\eta, 0]$, and let β be any positive real number with the solution $x(\cdot) = x(\phi, \cdot)$ defined over $[0, \beta)$. Now form a real function of t by setting $V(x_t) = F(x_t)^T F(x_t) + \sum_1^p (k_i \cdot \int_{t-h_i}^t \dot{x}^T(u)\dot{x}(u)du)$. Noting for $0+ \leq t < \beta$ that $F(x_t) = \dot{x}(t)$, and calculating the right derivative of this function $V(x_t)$, we find for $0+ \leq t < \beta$ that

$$\begin{aligned}
\dot{V}(x_t) &= 2\,\langle\langle f'(x_e(t)) \cdot \dot{x}_e(t), \dot{x}(t)\rangle\rangle + \sum_1^p k_i(\dot{x}^T(t)\dot{x}(t) - \dot{x}^T(t-h_i)\dot{x}(t-h_i)) \\
&= 2\,\langle\langle [A_0 \cdots A_p]\,\dot{x}_e(t), \dot{x}(t)\rangle\rangle + \sum_1^p k_i(\dot{x}^T(t)\dot{x}(t) - \dot{x}^T(t-h_i)\dot{x}(t-h_i)) \\
&= \left(\dot{x}^T(t)A_0^T + \sum_1^p \dot{x}^T(t-h_i)A_i^T\right)\dot{x}(t) + \dot{x}^T(t)\left(A_0\dot{x}(t) + \sum_1^p A_i\dot{x}(t-h_i)\right) \\
&\quad + \dot{x}^T(t)\left(\sum_1^p k_i I\right)\dot{x}(t) - \sum_1^p \dot{x}^T(t-h_i)(k_i I)\dot{x}(t-h_i),
\end{aligned}$$

where the derivatives $\dot{x}(t)$ and $\dot{x}(t-h_i)$ are taken on the right. Writing $S_k = S_k(v)$ with $v = x_e(t)$, we see that $\dot{V}(x_t) = \dot{x}_e^T(t) S_k \dot{x}_e(t)$.

Noting that $\lambda_{\max}(S_k) \leq 2\gamma$, we have $\dot{V}(x_t) \leq 2\gamma|\dot{x}_e(t)|^2$, and recalling the definition of V, we obtain $d\dot{x}^T(t)\dot{x}(t)/dt + \sum_1^p k_i(|\dot{x}(t)|^2 - |\dot{x}(t-h_i)|^2) \leq 2\gamma|\dot{x}(t)|^2 + 2\gamma \cdot \sum_1^p |\dot{x}(t-h_i)|^2$. Thus we have $d|\dot{x}(t)|^2/dt \leq (2\gamma - \sum_1^p k_i)|\dot{x}(t)|^2 + \sum_1^p (2\gamma + k_i)|\dot{x}(t-h_i)|^2$ for $0+ \leq t < \beta$. Now set $m(t) = |\dot{x}(t)|^2$ for $-\eta \leq t < \beta$, with $\dot{x}(\cdot)$ taken as usual on the right. Set $a = 2\gamma - \sum_1^p k_i$, set $b_i = 2\gamma + k_i$ for $i = 1, \ldots, p$, and note that $m'_+(t) \leq am(t) + \sum_1^p b_i m(t-h_i)$ for $0 \leq t < \beta$. Since $\gamma < 0$, we know that $a < 0$, and also that $|a| = -2\gamma + \sum_1^p k_i > \sum_1^p k_i > \sum_1^p b_i$, i.e. $|a| > \sum_1^p b_i$. Furthermore, since $b_i = 2\gamma + k_i \geq \lambda_{\max}(S_k) + k_i$, one can note Lemma 3.2, and see that $b_i \geq 0$ for each $i = 1, \ldots, p$. Now for $0 \leq t < \beta$, $m(\cdot)$ satisfies the above delay-differential inequality with initial data $m_0 \in C_-[-\eta, 0]$ given by $m_0(u) = |\dot{\phi}(u)|^2$ for $-\eta \leq u < 0$, and $m_0(0) = |F(\phi)|^2 = |\dot{x}(0+)|^2$. Noting Lemma 3.5, we immediately see that there exist constants $C_0 > 0, \alpha_0 < 0$ with $m(t) \leq (C_0\|m_0\|_-)e^{\alpha_0 t}$ for $0 \leq t < \beta$. Setting $C = (C_0\|m_0\|_-)^{1/2}, \alpha = \alpha_0/2$, and noting $m(t) = |\dot{x}(t)|^2$, we have $|\dot{x}(t)| \leq Ce^{\alpha t}$ for $0 \leq t < \beta$.

From this inequality for $|\dot{x}|$, we see that for $0 \leq t \leq \tau < \beta$, one has

$$|x(\tau) - x(t)| = \left|\int_t^\tau \dot{x}\right| \leq \int_t^\tau |\dot{x}| \leq C\,\frac{e^{\alpha\tau} - e^{\alpha t}}{\alpha},$$

$$\text{i.e., } |x(\tau) - x(t)| \leq \left(\frac{C}{|\alpha|}\right)e^{\alpha t} \quad \text{for } 0 \leq t \leq \tau < \beta.$$

Thus $|x(t) - x(0)| \leq C/|\alpha|$ over $[0, \beta)$. Since β was an arbitrary positive real number for which the solution $x(\cdot)$ is defined over $[0, \beta)$, we see that there is a fixed bounded subset of R^n which $x(\cdot)$ does not escape over its interval of existence, and conclude that the solution $x(t)$ is defined over $[0, \infty)$.

Using the Cauchy criterion with the bound $|x(\tau) - x(t)| \leq C(|\alpha|)^{-1}e^{\alpha t}$, valid for $\tau \geq t \geq 0$, one can easily now see that $\lim_{t\to\infty} x(t)$ exists, and we denote this limit by $\bar{x}$. Writing $\bar{x} = \lim_{\tau\to\infty} x(\tau)$ and noting this bound, we have $|\bar{x} - x(t)| \leq C(|\alpha|)^{-1}e^{\alpha t}$

for $t \geq 0$, and thus the rate of convergence of $x(t)$ to $\bar{x}$ is exponential, with exponent less than or equal to αt. Since $x(t) \to \bar{x}$ as $t \to \infty$, we know for $i = 1, \ldots, p$ that $x(t - h_i) \to \bar{x}$ as $t \to \infty$, and we have $f(x_e(t)) \to f(\bar{x}_c)$ as $t \to \infty$. Recalling the inequality $|\dot{x}(t)| \leq Ce^{\alpha t}$, we see that $\dot{x}(t) \to 0$ as $t \to \infty$, and since $\dot{x}(t) = f(x_e(t))$, we know also that $\dot{x}(t) \to f(\bar{x}_c)$ as $t \to \infty$, and thus $f(\bar{x}_c) = 0$. □

COROLLARY 3.1. *Consider the delay-differential equation* $(\dagger)\dot{x}(t) = f(x_e(t))$, *with the hypotheses of Theorem* 3.1. *For each* $\phi \in C^1[-\eta, 0]$, *define* $m_\phi \in C_-[-\eta, 0]$ *by* $m_\phi(u) = |\dot{\phi}(u)|^2$ *for* $-\eta \leq u < 0$, *and* $m_\phi(0) = |F(\phi)|^2$. *Then for each* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that if* $\|m_\phi\|_- < \delta$, *then* $|x(\phi, t) - \phi(0)| < \varepsilon$ *for all* $t \geq 0$.

Proof. In the bound given in the theorem, we have $|x(\phi, t) - \phi(0)| \leq C(|\alpha|)^{-1}$ for $0 \leq t < \infty$, i.e., $|x(\phi, t) - \phi(0)| \leq (C_0\|m_\phi\|_-)^{1/2}(|\alpha|)^{-1}$ for $t \geq 0$. For any δ with $0 < \delta < (\alpha^2/C_0)\varepsilon^2$ and $t \geq 0$, we thus have $\|m_\phi\|_- < \delta \Longrightarrow |x(\phi, t) - \phi(0)| \leq (C_0)^{1/2}(|\alpha|)^{-1}(\|m_\phi\|_-)^{1/2} \Longrightarrow |x(\phi, t) - \phi(0)| < \varepsilon$, i.e., if $\|m_\phi\|_- < \delta$, then $|x(\phi, t) - \phi(0)| < \varepsilon$ for all $t \geq 0$. □

THEOREM 3.2. *Let* $\gamma < 0$ *and let* $k \in (R^p)_+$. *Set* $a = 2\gamma - \sum_1^p k_i$, *and set* $b_i = 2\gamma + k_i$ *for* $i = 1, \ldots, p$. *Consider the delay-differential equation* $(\dagger)\dot{x}(t) = f(x_e(t))$, *where* f *is continuously differentiable throughout* $R^{n(p+1)}$. *Let* $Z = V_0 \times \cdots \times V_p$, *where the* V_i *are open subsets of* R^n, *and* $V_i \supset V_0$ *for* $i = 1, \ldots, p$. *Suppose that for each* $v = (v_0, \ldots, v_p) \in Z$, *one has* $\lambda_{\max}(S_k) \leq 2\gamma$ *for the matrix* S_k *formed from* $f'(v) = [A_0 \cdots A_p]$. *Let* C_0, α_0 *be the constants obtained from* a *and the* b_i, h_i *as in Lemma* 3.5, *and let* $\alpha = \alpha_0/2$. *For each* $\phi \in C^1[-\eta, 0]$, *define* m_ϕ *by* $m_\phi(u) = |\dot{\phi}(u)|^2$ *for* $-\eta \leq u < 0$, *and* $m_\phi(0) = |F(\phi)|^2$, *and set* $C = (C_0\|m_\phi\|_-)^{1/2}$. *Then the following holds*: *If* $V_0 \supset \{x \in R^n : |x - \phi(0)| \leq C|\alpha^{-1}|\}$ *and* $V_i \supset$ range (ϕ) *for* $i = 1, \ldots, p$, *then the solution* $x(\phi, \cdot)$ *is defined over* $[0, \infty)$, *there exists* $\bar{x} \in \{|x - \phi(0)| \leq C|\alpha^{-1}|\}$ *such that* $x(\phi, t)$ *converges at an exponential rate to* $\bar{x}$ *as* $t \to \infty$, *and* $f(\bar{x}_c) = 0$.

Proof. Consider the set $\mathcal{B}$ of all nonnegative real numbers β for which both (1) the solution $x(\cdot) = x(\phi, \cdot)$ is defined over the interval $[0, \beta)$, and (2) for $0 \leq t < \beta$, one has $V_0 \ni x(t)$ and each $V_i \supset$ range (x_t) for $i = 1, \ldots, p$. Noting that $x(0) = \phi(0)$, each $V_i \supset$ range (ϕ), and $x(\cdot)$ is continuous, we see that $\mathcal{B}$ contains a nonempty half-open interval including 0. If it were the case that $\mathcal{B}$ was bounded, then one could set $\tilde{\beta} = \sup \mathcal{B}$. For $0 \leq t < \tilde{\beta}$, we could write the inequality $\dot{V}(x_t) \leq 2\gamma|\dot{x}_e(t)|^2$, as usual, with $\dot{x}(\cdot)$ taken on the right, so that for $m(t) = |\dot{x}(t)|^2$, we again have $m'_+(t) \leq am(t) + \sum_1^p b_i m(t - h_i)$ for $0 \leq t < \tilde{\beta}$. Noting the resulting inequality for $|x(\tau) - x(t)|$ as in the theorem above, we find that $|x(t) - x(0)| \leq C\alpha^{-1}(e^{\alpha t} - 1)$. Thus, $|x(t) - x(0)| \leq C|\alpha^{-1}|$ for $0 \leq t < \tilde{\beta}$. Since the solution $x(\cdot)$ is contained in $\{|x - \phi(0)| \leq C|\alpha^{-1}|\}$ over $[0, \tilde{\beta})$, we know that $x(\cdot)$ is continuable, i.e., there exists $\varepsilon_0 > 0$ such that the solution $x(\cdot)$ is defined over $[0, \tilde{\beta} + \varepsilon_0)$. Since we would now have $|x(t) - \phi(0)| \leq C|\alpha^{-1}|$ over $[0, \tilde{\beta}]$, there would then exist ε with $0 < \varepsilon < \varepsilon_0$ having $V_0 \ni x(t)$ over $[0, \tilde{\beta} + \varepsilon)$. Thus for each $i = 1, \ldots, p$ we would have $V_i \supset$ range (x_t) for $\tilde{\beta} \leq t < \tilde{\beta} + \varepsilon$, since if $t + u \leq 0$, then $x_t(u) = x(t + u) \in$ range (ϕ), and if $0 < t + u < \tilde{\beta} + \varepsilon$, then $x(t + u)$ lies in V_0, hence also $x_t(u) \in V_i$ for $i = 1, \ldots, p$. We would thus see for $0 \leq t < \tilde{\beta} + \varepsilon$ that $V_0 \ni x(t)$ and each $V_i \supset$ range (x_t), and immediately conclude that $\tilde{\beta} + \varepsilon \leq \sup \mathcal{B}$, which contradicts $\tilde{\beta} = \sup \mathcal{B}$. We now conclude that $\mathcal{B}$ is unbounded, i.e., $\sup \mathcal{B} = \infty$. We can now write $\dot{V}(x_t) \leq 2\gamma|\dot{x}_e(t)|^2$ for $0 \leq t < \infty$, and note the again resulting inequalities. Particularly noting that $x(t)$ lies in the compact set $\{|x - \phi(0)| \leq C|\alpha^{-1}|\}$ for $0 \leq t < \infty$, the theorem now follows from the inequalities for $|x(\tau) - x(t)|$ as in Theorem 3.1. □

The following lemma will be valuable in analyzing the issue of uniqueness of the equilibria of the nonlinear delay-differential equation $(\dagger)\dot{x}(t) = f(x_e(t))$.

LEMMA 3.6. *Let* $f(v)$ *be continuously differentiable on an open subset of* $R^{n(p+1)}$,

and let $k \in (R^p)_+$. *Again consider the matrix*

$$S_k = \begin{bmatrix} A_0 + A_0^T + \sum_1^p k_i I & B \\ B^T & D(k) \end{bmatrix},$$

where $f'(v) = [A_0 \cdots A_p]$ *and* $B = [A_1 \cdots A_p]$. *For* $x_c = (x, \ldots, x)$, *set* $d(x) = f(x_c)$. *If* $d(\bar{x}) = 0$ *and* S_k *is sign definite at* $v = \bar{x}_c$, *then* $\bar{x}$ *is an isolated zero of* $d(x)$.

Proof. Note that $d'(x) = f'(x_c) \cdot [I \cdots I]^T = [A_0 \cdots A_p] \cdot [I \cdots I]^T = \sum_0^p A_i(x_c)$. If S_k is sign definite at $v = \bar{x}_c$, we know from Lemma 1.1 that $\sum_0^p A_i(x, \ldots, x)$ is nonsingular at $x = \bar{x}$, and the lemma now follows from the inverse function theorem. □

Although the theorems in this section are as previously mentioned stated for delay-differential equations having initial data in the space $C^1[-\eta, 0]$, it is convenient here to introduce the supremum norm on the space $C[-\eta, 0]$, i.e. for each $\phi \in C[-\eta, 0]$, we define $\|\phi\|_s$ by $\|\phi\|_s = \sup\{|\phi(u)| : -\eta \leq u \leq 0\}$. We now examine the uniqueness issue for the equilibria of nonlinear delay-differential systems of the form $(\dagger)\dot{x}(t) = f(x_e(t))$. The most challenging aspect of this will come in demonstrating the continuity of the map taking members ϕ of $C^1[-\eta, 0]$ upon the limit as $t \to \infty$ of the trajectory $x(\phi, t)$. The idea will be that for time greater than some fixed value, the trajectory $x(\phi, t)$ will be close to a zero of the above function $d(x) = f(x_c)$, and in this vicinity $f(v)$ has small magnitude. Since over finite time trajectories can be made close to $x(\phi, t)$ by making initial data close to ϕ, trajectories with nearby initial data will at least temporarily be close to the same zero of $f(x_c)$ as the zero approached by $x(\phi, t)$. The previous lemma will tell us that this zero is isolated, and since trajectories are in a vicinity of low momentum, they will be unable to escape the vicinity of the zero, so that trajectories will converge to this zero.

THEOREM 3.3. *Again consider the delay-differential equation* $(\dagger)\dot{x}(t) = f(x_e(t))$, *where* f *is continuously differentiable throughout* $R^{n(p+1)}$. *Again suppose that there exist* $k \in (R^p)_+, \gamma < 0$ *such that for each* $v = (v_0, \ldots, v_p) \in (R^n)^{p+1}$, *the matrix*

$$S_k = \begin{bmatrix} A_0 + A_0^T + \sum_1^p k_i I & B \\ B^T & D(k) \end{bmatrix}$$

formed from $f'(v) = [A_0 \cdots A_p]$ *satisfies* $\lambda_{\max}(S_k) \leq 2\gamma$. *Then for each* $\phi \in C^1[-\eta, 0]$, *there exists* $r > 0$ *such that if* $\tilde{\phi} \in C[-\eta, 0]$ *and* $\|\tilde{\phi} - \phi\|_s < r$, *one has* $\lim_{t\to\infty} x(\tilde{\phi}, t) = \lim_{t\to\infty} x(\phi, t)$.

Proof. The fact that the second of the above limits exists is an immediate consequence of Theorem 3.1. The existence of the first limit will be established in the proof. Begin by setting $\lim_{t\to\infty} x(\phi, t) = \bar{x}$, and note from Theorem 3.1 that for $d(x) = f(x_c)$, we have $d(\bar{x}) = 0$. Examining the constant matrix $f'(\bar{x}_c) = [A_0 \cdots A_p]$, and recalling Theorem 1.1, we note that $S_k(\bar{x}_c)$ is negative definite, and conclude that the zeros of the function $g(s) = |sI - A_0 - \sum_1^p A_i e^{-sh_i}|$ have negative real part. Considering the member ϕ_0 of $C[-\eta, 0]$ defined by $\phi_0(u) = \bar{x}$ for $-\eta \leq u \leq 0$, it now follows from well-known facts in the theory of functional differential equations [6] that there exists $\delta > 0$ such that for each $\tilde{\phi} \in C[-\eta, 0]$ having $\|\tilde{\phi} - \phi_0\|_s < \delta$, the solution $x(\tilde{\phi}, t)$ exists for $0 \leq t < \infty$, and $\lim_{t\to\infty} x(\tilde{\phi}, t) = \bar{x}$.

Now take $t_0 > 0$ with $|x(\phi, t) - \bar{x}| < \frac{1}{2}\delta$ for all $t \geq t_0$. By continuous dependence of $x(\tilde{\phi}, t)$ upon $\tilde{\phi}$ over intervals of finite length, we know that there exists $r > 0$ such that

for each $\tilde{\phi} \in C[-\eta, 0]$ having $\|\tilde{\phi}-\phi\|_s < r$, the solution $x(\tilde{\phi}, t)$ exists over $[0, t_0+\eta]$, and $|x(\tilde{\phi}, t)-x(\phi, t)| < \frac{1}{2}\delta$ uniformly over $[0, t_0+\eta]$. For such $\tilde{\phi}$, we know from the triangle inequality that $|x(\tilde{\phi}, t) - \bar{x}| < \delta$ for $t_0 \leq t \leq t_0 + \eta$. Setting $\psi(u) = x(\tilde{\phi}, t_0 + \eta + u)$ for $-\eta \leq u \leq 0$, we see that $\|\psi - \phi_0\|_s < \delta$, so that the solution $x(\psi, t)$ exists over $[0, \infty)$, and $\lim_{t\to\infty} x(\psi, t) = \bar{x}$. Thus $\lim_{t\to\infty} x(\tilde{\phi}, t) = \bar{x}$. We have shown that $\lim_{t\to\infty} x(\tilde{\phi}, t) = \bar{x} = \lim_{t\to\infty} x(\phi, t)$ for any $\tilde{\phi} \in C[-\eta, 0]$ having $\|\tilde{\phi} - \phi\|_s < r$, and the theorem is now proven. □

THEOREM 3.4. *Again consider the delay-differential equation* (†)$\dot{x}(t) = f(x_e(t))$, *with the same hypotheses as in Theorem* 3.1 *and Theorem* 3.3. *Then there is a unique point* $\bar{x} \in R^n$ *having* $f(\bar{x}_c) = 0$. *For each* $\phi \in C^1[-\eta, 0]$, *one has* $x(\phi, t) \to \bar{x}$ *at an exponential rate as* $t \to \infty$.

Proof. Noting Theorem 3.1, one need only show the uniqueness of the zero of the function $d(x) = f(x_c)$. Let $C^1[-\eta, 0]$ have the supremum norm $\|\phi\|_s = \sup\{|\phi(u)| : -\eta \leq u \leq 0\}$, and let R^n have the standard norm. Define the map $T : C^1[-\eta, 0] \to R^n$ by $T\phi = \lim_{t\to\infty} x(\phi, t)$. Noting that $C^1[-\eta, 0]$ is connected, and noting from the above theorem that T is continuous, we see that the image of T is connected. From Lemma 3.6, we know that the zeros of the function $d(x)$ are isolated, and since the image of T is contained in the set of zeros of $d(x)$, we see that the image of T is a single point.

Now let $\{\bar{x}\}$ be the image of T. For any $x' \in R^n$ having $d(x') = 0$, let $\phi(u) = x'$ for $-\eta \leq u \leq 0$. Since $x(\phi, t) = x'$ for all $t \geq 0$, we have $x' = \lim_{t\to\infty} x(\phi, t) = T\phi = \bar{x}$, i.e. $x' = \bar{x}$. Thus we see that the zero of $d(x)$ in R^n is unique. □

4. In this section we give examples of the theorems in §3. Although, for the sake of simplicity, all examples presented are of nonlinear scalar delay-differential systems, one can use the techniques found below to give examples of similar phenomena occurring in nonlinear vector delay-differential systems. In the examples given we present nonlinear delayed systems which will be shown to satisfy the hypotheses of Theorem 3.1.

Example 4.1. Let ω be any fixed real number, let $h > 0$, and consider the nonlinear delay-differential equation (†)$\dot{x}(t) = 5\sin(\omega x(t) + \omega x(t-h)) - 3x(t)$.

In this case we write $v = (v_0, v_1) = (x, y)$, and we have $f(x, y) = 5\sin(\omega x + \omega y) - 3x$, so that $a_0(x, y) = D_1 f(x, y) = 5\omega\cos(\omega x + \omega y) - 3$, and $a_1(x, y) = D_2 f(x, y) = 5\omega\cos(\omega x + \omega y)$. Recalling the matrix

$$S_k = \begin{bmatrix} a_0 + a_0^T + kI & a_1 \\ a_1^T & -kI \end{bmatrix}, \text{ we have}$$

$$S_k = \begin{bmatrix} 10\omega\cos(\omega x + \omega y) - 6 + k & 5\omega\cos(\omega x + \omega y) \\ 5\omega\cos(\omega x + \omega y) & -k \end{bmatrix}.$$

One will find that the discriminant D for the characteristic polynomial of S_k is given by $\frac{1}{4}D = (k - 3 + 5\omega\cos(\omega x + \omega y))^2 + (5\omega\cos(\omega x + \omega y))^2$. If we set $k = 3$, then

$$\frac{1}{4}D = 50\omega^2\cos^2(\omega x + \omega y), \text{ and}$$

$$S_3 = \begin{bmatrix} -3 + 10\omega\cos(\omega x + \omega y) & 5\omega\cos(\omega x + \omega y) \\ 5\omega\cos(\omega x + \omega y) & -3 \end{bmatrix}.$$

After performing routine calculations, the eigenvalues of S_3 are found to be

$$\lambda_1 = -3 + 5\omega\cos(\omega x + \omega y) - 5\cdot\left(2^{1/2}\right)|\omega\cos(\omega x + \omega y)|,$$
$$\lambda_2 = -3 + 5\omega\cos(\omega x + \omega y) + 5\cdot\left(2^{1/2}\right)|\omega\cos(\omega x + \omega y)|.$$

For any ω having $|\omega| < .6/(1+2^{1/2})$, we have $5(1+2^{1/2})|\omega| < 3$, so that $-3+5(1+2^{1/2})|\omega| < 0$. Setting $-3+5(1+2^{1/2})|\omega| = -2\varepsilon$, we see that for each $(x,y) \in R^2$, one has $\lambda_{\max}(S_3(x,y)) \leq -2\varepsilon$.

Thus, for any ω with $|\omega| < .6/(1+2^{1/2})$, the hypotheses of Theorem 3.1 are satisfied, and for such ω we now know that there is exactly one solution to the nonlinear equation $f(x,x) = 0$. Writing $f(x,x) = 5\sin(2\omega x) - 3x$, and noting that $f(0,0) = 0$, we see that the origin in R is the sole solution to $f(x,x) = 0$. Thus for each $\phi \in C^1[-h,0]$, we have $x(\phi,t) \to 0$ as $t \to \infty$. In fact, given $\phi \in C^1[-h,0]$, there exists $r = r_\phi > 0$ such that if $\tilde{\phi} \in C[-h,0]$ and $\|\tilde{\phi} - \phi\|_s < r$, then $x(\tilde{\phi},t)$ exists for $0 \leq t < \infty$, and $\lim_{t\to\infty} x(\tilde{\phi},t) = 0$.

Example 4.2. Consider the nonlinear delay-differential equation $(\dagger)\dot{x}(t) = \cos(x(t)+x(t-h)) - 3x(t)$.

Again writing $v = (v_0, v_1) = (x,y)$, we have $f(x,y) = \cos(x+y) - 3x$, and proceeding with techniques similar to those found in the above example, one can show that the function $f(x,y)$ satisfies the hypotheses of Theorem 3.1, so that there is exactly one solution to the nonlinear equation $f(x,x) = 0$. Writing $f(x,x) = \cos(2x) - 3x$, we see that the solution $\bar{x}$ of $f(x,x) = 0$ is the x-coordinate of the unique point in R^2 where the line given by $y = 3x$ intersects the curve given by $y = \cos(2x)$. Thus for each $\phi \in C^1[-h,0]$, we have $x(\phi,t) \to \bar{x}$ as $t \to \infty$.

One can construct many interesting examples making use of the following principle, which we first prove.

Slowed perturbations principle. If $f : R^{n(p+1)} \to R^n$ is a function satisfying the hypotheses of Theorem 3.1, and $g : R^{n(p+1)} \to R^n$ is any C^1 function with bounded *derivative* over $R^{n(p+1)}$, then there exists $\omega_0 > 0$ such that for each ω having $|\omega| < \omega_0$, the function $p(v) = f(v) + g(\omega v)$ also satisfies the hypotheses of Theorem 3.1.

Proof. In fact, write $f'(v) = [F_0 \cdots F_p]$, $g'(v) = [G_0 \cdots G_p]$, and write

$$F = \begin{bmatrix} F_0 + F_0^T & B_f \\ B_f^T & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} G_0 + G_0^T & B_g \\ B_g^T & 0 \end{bmatrix},$$

where $B_f = [F_1 \cdots F_p]$ and $B_g = [G_1 \cdots G_p]$. Now suppose that there exist some fixed $k \in (R^p)_+, \gamma < 0$ such that for all $v \in R^{n(p+1)}$, one has $\lambda_{\max}(U_k) \leq 2\gamma$ for the matrix U_k given by $U_k = F + \operatorname{diag}(k_0 I_n, -k_1 I_n, \ldots, -k_p I_n)$, where $k_0 = \sum_1^p k_i$. Note that if g has bounded derivative, we can set $m = \sup\{\|G(v)\| : v \in R^{n(p+1)}\}$, with $m < \infty$. For $j(v) = g(\omega v)$, one knows that $j'(v) = \omega g'(\omega v)$, so that if one writes $j'(v) = [J_0 \cdots J_p]$ and

$$J = \begin{bmatrix} J_0 + J_0^T & B_j \\ B_j^T & 0 \end{bmatrix}$$

with $B_j = [J_1 \cdots J_p]$, one then has $J(v) = \omega G(\omega v)$. Thus we know that $|\omega m| = \sup\{\|J(v)\| : v \in R^{n(p+1)}\}$. Now set $S_k = U_k + J$, and note that S_k is the matrix for the function $p(v) = f(v) + g(\omega v)$. Examining the quadratic form $v^T S_k v = v^T(U_k + J)v$ for $v \in R^{n(p+1)}$, one will easily find that $\lambda_{\max}(S_k) \leq \lambda_{\max}(U_k) + \|J\|$. For any ω having $|\omega| < 2|\gamma|/m$, we have $|\omega m| < 2|\gamma|$, so that $\|J\| < |\lambda_{\max}(U_k)|$ for each $v \in R^{n(p+1)}$, and $\lambda_{\max}(S_k) \leq 2\gamma + |\omega m| < 0$. With $\omega_0 = 2|\gamma|/m$, we have shown that if $|\omega| < \omega_0$, then for each $v \in R^{n(p+1)}$, one has $\lambda_{\max}(S_k) \leq 2(\gamma + |\omega m/2|)$, with $\gamma + |\omega m/2| < 0$. Thus for $|\omega| < \omega_0$ the function $p(v)$ does indeed satisfy the hypotheses of Theorem 3.1. □

Example 4.3. We examine Example 4.1 in light of the above principle. Beginning with $\omega = 0$, the delay-differential equation is written $\dot{x}(t) = -3x(t)$, and again writing $v = (v_0, v_1) = (x, y)$, we have $f(x, y) = -3x$. Then $F = \text{diag}\,(-6, 0)$, and for any $k \in (0, 3)$, we have $2\gamma = \lambda_{\max}(U_k) = -k$ for U_k as above. Writing $g(x, y) = 5\sin(x + y)$, we have $j(x, y) = g(\omega x, \omega y)$. This gives $p(x, y) = f(x, y) + g(\omega x, \omega y)$, so that $p(x, y) = 5\sin(\omega x + \omega y) - 3x$, and the associated delay-differential equation $(\dagger)\dot{x}(t) = p(x(t), x(t-h))$ is $\dot{x}(t) = 5\sin(\omega x(t) + \omega x(t-h)) - 3x(t)$, as in Example 4.1.

Now $g'(x, y) = 5(\cos(x + y), \cos(x + y))$, so that

$$G(x, y) = \begin{bmatrix} 10\cos(x + y) & 5\cos(x + y) \\ 5\cos(x + y) & 0 \end{bmatrix}.$$

Calculating the eigenvalues of the symmetric matrix $G(x, y)$, one can find that $\|G(x, y)\| \leq 5(1 + 2^{1/2})$ for all $(x, y) \in R^2$. Letting $S_k(\cdot, \cdot)$ be the matrix for $p(\cdot, \cdot)$ as above, we now see that for any $k \in (0, 3)$, one has $\lambda_{\max}(S_k) \leq -k + |\omega m| \leq -k + 5(1 + 2^{1/2})|\omega|$ for all $(x, y) \in R^2$. For any ω having $|\omega| < 3/5(1 + 2^{1/2})$, we take k with $5(1 + 2^{1/2})|\omega| < k < 3$, and set $-k + 5(1 + 2^{1/2})|\omega| = -2\varepsilon$. We then see for all $(x, y) \in R^2$ that $\lambda_{\max}(S_k(x, y)) \leq -2\varepsilon$. For each ω with $|\omega| < .6/(1 + 2^{1/2})$, then, the function $p(x, y)$ satisfies the hypotheses of Theorem 3.1, and for such ω we again arrive at the conclusions of Example 4.1.

Finally, using MATLAB and noting the *slowed perturbations principle*, one can easily construct many interesting examples for the case $n \geq 2, p \geq 2$ of functions $f : R^{n(p+1)} \to R^n$ which satisfy the hypotheses of Theorem 3.1. It is particularly interesting here for the perturbation function $j(v) = g(\omega v)$ that g may be unbounded over $R^{n(p+1)}$, provided that the derivative g' is bounded over $R^{n(p+1)}$. The range of the *derivative* of the perturbation, not the range of the perturbation itself, is the object of interest following from the theorems of §3 which is relevant to the stability of the system $(\dagger)\dot{x}(t) = p(x_e(t))$.

Acknowledgments. An anonymous referee made suggestions leading to simplifications of several of the proofs in this paper. The author is grateful to this referee for these suggestions, as well as for remarks encouraging the author to broaden the scope of the basic theorems of this paper.

REFERENCES

[1] F. Brauer, *Absolute Stability in Delay Equations*, Journal of Differential Equations, Vol. 69, 1987, pp. 185–191.

[2] K. Cooke, J. Kaplan, and M. Sorg, *Stability of a functional differential equation for the motion of a radiating charged particle*, Nonlinear Anal., 5 (1981), pp. 1133–1139.

[3] J. Haddock, *Functional differential equations for which each constant function is a solution: A narrative*, Proceedings of the Eleventh International Conference on Nonlinear Oscillations, Janos Bolyai Mathematical Society, Budapest, Hungary, 1987, pp. 86–93.

[4] J. Haddock, M. Nkashama, and J. Wu, *Asymptotic constancy for linear neutral Volterra integrodifferential equations*, Tohoku Mathematical Journal, 41 (1989), pp. 689–710.

[5] J. Haddock and J. Terjeki, *On the location of positive limit sets for autonomous functional differential equations with infinite delay*, J. of Differential Equations, 86 (1990), pp. 1–32.

[6] J. K. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[7] J. K. Hale, E. F. Infante, and F. Tsen, *Stability in linear delay equations*, J. Math. Anal. Appl., 105 (1985), pp. 533–555.

[8] D. Hill and M. Mareels, *Stability theory for differential/algebraic systems with application to power systems*, IEEE Trans. Circuits and Systems, 37 (1990), pp. 1416–1423.

[9] J. KAPLAN, M. SORG, AND J. YORKE, *Solutions of $x'(t) = f(x(t), x(t-l))$ have limits when f is an order relation*, Nonlinear Anal., 3 (1979), pp. 53–58.

[10] N. KRASOVSKII, *Stability of Motion*, Moscow, 1959. Translation, Stanford University Press, Stanford, CA, 1963.

[11] L. MARKUS AND H. YAMABE, *Global stability criteria for differential systems*, Osaka Math. J., 12 (1960), pp. 305–317.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1195–1202, July 1994

011

STABLE INHOMOGENEOUS ITERATIONS OF NONLINEAR POSITIVE OPERATORS ON BANACH SPACES*

TAKAO FUJIMOTO† AND ULRICH KRAUSE‡

Abstract. For a sequence $(f_n)_n$ of nonlinear positive operators on a Banach space which converges to some operator f, conditions are specified under which the inhomogeneous iterates $f_n \circ f_{n-1} \circ \cdots \circ f_2 \circ f_1$, after normalization, converge to the unique positive and normalized eigenvector of f. This stability result extends, for discrete dynamical systems, the property of strong ergodicity from finite to infinite dimensions.

Key words. positive discrete dynamical systems, nonlinear positive operators, inhomogeneous iterations, strong ergodicity, Hilbert's projective metric

AMS subject classifications. 47H07, 47H09, 47A35

1. Introduction.

For a sequence $(f_n)_{n\geq 1}$ of nonlinear positive operators on a Banach space, consider the inhomogeneous iterates $f_n \circ f_{n-1} \circ \cdots \circ f_2 \circ f_1(x)$. It is well known that under certain conditions on the f_n there holds weak ergodicity, meaning that the path defined by the normalized inhomogeneous iterates becomes finally independent of the starting point x. (Cf. [2], [3], and [5]–[7].) Despite weak ergodicity, the path itself, however, does in general not converge, as can be seen already from simple examples of affine functions in two dimensions. The property of convergence for inhomogeneous iterates, also called strong ergodicity, has been dealt with for nonlinear positive operators in the finite-dimensional case in [1] and [2]. The results obtained there are extended in the present paper to Banach spaces, whereby the proof given cannot rely on compactness arguments as in finite dimensions, but uses some refinement of a technique employed in [4]. Strong ergodicity, as well as the notion of inhomogeneous products (of matrices), stems from the field of Markov chains and appears in applications to nonautonomous discrete dynamical systems, in particular those from biology and economics. (Cf. [1]–[3], and [6].)

2. Notation and definitions.

Let $(E, \|\cdot\|)$ denote a real Banach space, and let $K \subset E$ be a closed and normal convex cone. K induces an *ordering* $\leq$ by $x \leq y$ if and only if $y - x \in K$ for $x, y \in K$. Without restriction the norm $\|\cdot\|$ may be assumed to be *increasing* on K, i.e., $\|x\| \leq \|y\|$ for $0 \leq x \leq y$. An operator $f : K \to K$ is called *ray-preserving* [2] if for every $x \in K$ and every scalar $\lambda > 0$ there exists a scalar $\lambda(x) > 0$ such that $f(\lambda x) = \lambda(x) f(x)$. Denote by U the set $U = \{x \in K \mid \|x\| = 1\}$. An operator $f : K \to K$ is called *ascending* with respect to the norm $\|\cdot\|$, if the following two conditions are satisfied [4]:

(i) There exists a continuous mapping $\varphi : [0,1] \to [0,1]$ with $\varphi(\lambda) > \lambda$ for $0 < \lambda < 1$ such that for every $\lambda \in [0,1]$ and every $x, y \in U$,

$$\lambda x \leq y \quad \text{implies } \varphi(\lambda) f(x) \leq f(y);$$

(ii) for every $x, y \in U$ there exists a number

$$a = a(x,y) > 0 \quad \text{such that } af(x) \leq f(y).$$

* Received by the editors November 13, 1991; accepted for publication (in revised form) April 30, 1993.

† University of Okayama, Okayama 700, Japan.

‡ University of Bremen, 2800 Bremen 33, Germany.

An operator $f : K \to K$ is called *U-bounded*, if there exists scalars $0 < a \leq b$ and $u, v \in U$ such that

$$au \leq f(x) \leq bv \quad \text{for all } x \in U.$$

For an operator $f : K \to K$ with $f(x) \neq 0$ for all $x \neq 0$ the *rescaled operator* to f is defined by

$$Tx = \frac{f(x)}{\|f(x)\|} \quad \text{for } x \in U.$$

As a tool we shall use *Hilbert's projective metric* d on U which is defined as follows [2]–[6]: $d(x, y) = -\log[\lambda(x, y) \cdot \lambda(y, x)]$, where $\lambda(x, y) = \sup\{\lambda \geq 0 | \lambda x \leq y\}$ for $x, y \in U$.

3. Stable inhomogeneous iterations. The main result of the present paper is the following theorem.

THEOREM. *Let $(f_n)_{n\geq 1}$ be a sequence of operators $f_n : K \to K, f_n(x) \neq 0$ for $x \neq 0$, which converges uniformly on U to some operator $f : K \to K$, which is uniformly continuous on U, U-bounded and ascending (for the given norm). Then the equation $f(x) = \lambda x$ with $x \in U, \lambda > 0$ has a unique solution $x = x^*, \lambda = \lambda^*$ and for the rescaled operators T_n to f_n there holds the stability property*

$$\lim_{n\to\infty} T_n \circ T_{n-1} \circ \cdots \circ T_1(x) = x^* \quad \textit{uniformly on } U.$$

In case the f_n are ray-preserving the stability property becomes

$$\lim_{n\to\infty} \frac{f_n \circ f_{n-1} \circ \cdots \circ f_1(x)}{\|f_n \circ f_{n-1} \circ \cdots \circ f_1(x)\|} = x^* \quad \textit{uniformly on } K\backslash\{0\}.$$

In proving the theorem, we will need the following two lemmas.

LEMMA 1. (i) *If $f : K \to K$ is U-bounded and uniformly continuous on U, then the rescaled operator T to f is uniformly continuous on U.*

(ii) *If $(f_n)_{n\geq 1}$ is a sequence of operators $f_n : K \to K, f_n(x) \neq 0$ for $x \neq 0$, which converges uniformly to an U-bounded operator $f : K \to K$, then the sequence of the rescaled operators T_n of f_n converges uniformly on U to the rescaled operator T of f.*

Proof. (i) Since f is U-bounded, $au \leq f(x)$, and hence $a\|u\| \leq \|f(x)\|$, for some $a > 0, u \in U$ and all $x \in U$. In particular, $f(x) \neq 0$ for $x \in U$. For $x, y \in U$ it follows that

$$\begin{aligned}
\|T(x) - T(y)\| &= \left\| \frac{f(x)}{\|f(x)\|} - \frac{f(y)}{\|f(y)\|} \right\| \\
&= \frac{1}{\|f(x)\|} \left\| f(x) - f(y) + \left(1 - \frac{\|f(x)\|}{\|f(y)\|}\right) f(y) \right\| \\
&\leq \frac{1}{a}(\|f(x) - f(y)\| + |\|f(y)\| - \|f(x)\||) \leq \frac{2}{a}\|f(x) - f(y)\|.
\end{aligned}$$

Hence T is uniformly continuous on U.

(ii) Similarly, it holds for $x \in U$,

$$\begin{aligned}\|T(x) - T_n(x)\| &= \left\| \frac{f(x)}{\|f(x)\|} - \frac{f_n(x)}{\|f_n(x)\|} \right\| \\ &= \frac{1}{\|f(x)\|} \left\| f(x) - f_n(x) + \left(1 - \frac{\|f(x)\|}{\|f_n(x)\|}\right) f_n(x) \right\| \\ &\leq \frac{1}{a}(\|f(x) - f_n(x)\| + |\|f_n(x)\| - \|f(x)\||) \\ &\leq \frac{2}{a}\|f(x) - f_n(x)\|. \qquad \square\end{aligned}$$

LEMMA 2. *If $\varphi : [0,1] \to [0,1]$ is continuous with $\varphi(\lambda) > \lambda$ for $0 < \lambda < 1$, then the function $\psi : [0,1] \to [0,1]$ defined by $\psi(x) = \inf\{\sup\{\varphi(u) \cdot \varphi(v) | 0 \leq u \leq x_1, 0 \leq v \leq x_2\} | x_i \in [0,1], x_1 \cdot x_2 = x\}$ is continuous with $\psi(x) > x$ for $0 < x < 1$ and $\psi(x) \leq \psi(y)$ for $x \leq y$.*

Proof. (1) First, we show continuity of the function f defined by $f(x) = \sup\{\varphi(u) | 0 \leq u \leq x\}$. Obviously, $f(x) \leq f(y)$ for $x \leq y$. By continuity of φ, to $\varepsilon > 0$ there exists $\delta > 0$ such that $|\varphi(u) - \varphi(v)| \leq \varepsilon$ for $u, v \in [0,1], |u - v| \leq \delta$. Fix $x, y \in [0,1]$ with $|x - y| \leq \delta$. If $x \leq y$, then $f(x) \leq f(y) + \varepsilon$ holds trivially. Suppose $y \leq x$ and $u \leq x$. If $u \leq y$, then $\varphi(u) \leq f(y) + \varepsilon$ holds trivially. If $y \leq u \leq x$, then $|u - y| \leq \delta$ and hence $\varphi(u) \leq \varphi(y) + \varepsilon \leq f(y) + \varepsilon$. Hence $\varphi(u) \leq f(y) + \varepsilon$ for all $u \leq x$, and therefore $f(x) \leq f(y) + \varepsilon$. Exchanging x and y gives $f(y) \leq f(x) + \varepsilon$. Thus $|f(y) - f(x)| \leq \varepsilon$ for $|x - y| \leq \delta$.

(2) With f from step (1), ψ becomes $\psi(x) = \inf\{f(x_1) \cdot f(x_2) | x_i \in [0,1], x_1 \cdot x_2 = x\}$. Since f is uniformly continuous on $[0,1]$, to $\varepsilon > 0$ we may choose $\delta > 0$ such that $|f(u) - f(v)| \leq \varepsilon$ for $u, v \in [0,1], |u - v| \leq \delta$. Fix $0 < \varepsilon \leq 1$ and consider $x, y \in [0,1]$. There exist $y_1, y_2 \in [0,1]$ such that $y = y_1 y_2$ and $\psi(y) \geq f(y_1) f(y_2) - \varepsilon$. For $r, s \in [0,1]$ define

$$\rho(r,s) = \begin{cases} \min\{1-r, 1-s\} & \text{if } r < 1, s < 1, \\ 1 - r & \text{if } r < 1, s = 1, \\ 1 - s & \text{if } r = 1, s < 1, \\ 1 & \text{if } r = 1, s = 1. \end{cases}$$

Let $\rho = \min\{\delta, \rho(y_1, y_2)\}$. Obviously, $0 < \rho \leq 1$. Now assume $|x - y| \leq \rho^2$. Consider first the case when $y_1 < 1$ and $y_2 < 1$ and put $x_1 = y_1 + \rho$, $x_2 = x/x_1 \leq (y_1 y_2 + \rho^2)/(y_1 + \rho) \leq y_2 + \rho$. It follows that $x_1, x_2 \in [0,1]$ and $x = x_1 x_2$. Furthermore, $f(x_1) \leq f(y_1) + \varepsilon$ and $f(x_2) \leq f(y_2) + \varepsilon$, which implies

$$(*) \qquad f(x_1) f(x_2) \leq f(y_1) f(y_2) + 3\varepsilon \leq \psi(y) + 4\varepsilon.$$

For $y_1 = y_2 = 1$ these inequalities hold trivially for any $x_1, x_2 \in [0,1], x_1 x_2 = x$.

It remains to consider, without restriction, the case when $y_1 < 1, y_2 = 1$. $(*)$ holds by choosing $x_1 = x, x_2 = 1$. From $(*)$ it follows that $\psi(x) \leq \psi(y) + 4\varepsilon$ for $|x - y| \leq \rho^2$. Similarly, by exchanging the roles of x and y,

$$\psi(y) \leq \psi(x) + 4\varepsilon \quad \text{for } |x - y| \leq \overline{\rho}^2 \quad \text{with } \overline{\rho} = \min\{\delta, \rho(x_1, x_2)\}.$$

Hence

$$|\psi(x) - \psi(y)| \leq 4\varepsilon \quad \text{for } |x - y| \leq \gamma^2 \quad \text{with } \gamma = \min\{\delta, \rho(x_1, x_2), \rho(y_1, y_2)\},$$

which shows the continuity of ψ on $[0, 1]$.

(3) Let $0 < x < 1$. Since f is continuous on $[0, 1]$ by step (1), there exist $x_1, x_2 \in [0, 1]$ such that $x = x_1 x_2$ and $\psi(x) = f(x_1) f(x_2)$. From the definition of f and $\psi(\lambda) > \lambda$ for $0 < \lambda < 1$ it follows that $f(x_1) f(x_2) \geq \varphi(x_1)\varphi(x_2) > x_1 x_2 = x$ because of $x_1, x_2 > 0$, and $\min\{x_1, x_2\} < 1$. This shows $\psi(x) > x$. Finally, it will be shown that ψ is increasing on $[0, 1]$. Let $x, y \in [0, 1], x \leq y$ and suppose $y = y_1 y_2$. If $y = 0$, then $x = 0$, and $\psi(x) \leq \psi(y)$ holds trivially. Assume $y > 0$ and choose $x_1 = y_1, x_2 = x/y_1 \leq y_2$. Obviously, $x_1, x_2 \in [0, 1], x = x_1 x_2$ and, by the definition of ψ,

$$\begin{aligned} \psi(x) &\leq \{\sup \varphi(u)\varphi(v) | 0 \leq u \leq x_1, 0 \leq v \leq x_2\} \\ &\leq \{\sup \varphi(u)\varphi(v) | 0 \leq u \leq y_1, 0 \leq v \leq y_2\}. \end{aligned}$$

By taking the infimum over $y_1 \in [0, 1]$ it follows that $\psi(x) \leq \psi(y)$. □

Proof of the theorem. (1) By induction we show that for every $\varepsilon > 0$ and $m \in \mathbb{N}$ there exists an $N = N(\varepsilon, m)$ such that for the rescaled operators $\|T_{n+m} \circ T_{n+m-1} \circ \cdots \circ T_{n+1}(x) - T^m(x)\| \leq \varepsilon$ for all $x \in U$ and $n \geq N$.

By Lemma 1, the T_n converge uniformly on U to the uniformly continuous operator T. Hence there exists for $\varepsilon > 0$ $\delta(\varepsilon) > 0, N(\varepsilon)$ such that $\|x - y\| \leq \delta(\varepsilon)$ implies that $\|T(x) - T_n(y)\| \leq \|T(x) - T(y)\| + \|T(y) - T_n(y)\| \leq (\varepsilon/2) + (\varepsilon/2) = \varepsilon$ for all $n \geq N(\varepsilon)$. Settng $N(\varepsilon, 1) = N(\varepsilon)$, this shows the desired approximation for $m = 1$. Suppose the approximation is true for some $m \geq 1$.

Then $\|T_{n+m} \circ \cdots \circ T_{n+1}(x) - T^m(x)\| \leq \delta(\varepsilon)$ for all $u \in U$ and all $n \geq N(\delta(\varepsilon), m)$, and hence

$$\begin{aligned} &\|T_{n+m+1} \circ \cdots \circ T_{n+1}(x) - T^{m+1}(x)\| \\ &= \|T_{n+m+1}(T_{n+m} \circ \cdots \circ T_{n+1}(x)) - T(T^m(x))\| \leq \varepsilon \end{aligned}$$

for all $x \in U$ and all $n \geq N(\delta(\varepsilon), m), n \geq N(\varepsilon)$. Setting $N(\varepsilon, m + 1) = \max\{N(\varepsilon), N(\delta(\varepsilon), m)\}$ completes the induction.

(2) Next we show that for an appropriate $x^* \in U$, for every $\varepsilon > 0$ there exists some $m = m(\varepsilon)$ such that

$$\|T^m(x) - x^*\| \leq \varepsilon \quad \text{for all } x \in U.$$

This we shall derive from the corresponding statement for Hilbert's projective metric d, because of the inequality

$$\|x - y\| \leq 3(1 - e^{-d(x,y)}) \quad \text{for all } x, y \in U.$$

To see this inequality, let $x, y \in U$ and $\lambda x \leq y, \mu y \leq x$ with $\lambda, \mu \geq 0$. Obviously, $\lambda, \mu \leq 1$. It follows

$$0 \leq x - y + (1 - \mu)y \leq (1 - \lambda)x + (1 - \mu)y \leq (1 - \lambda\mu)(x + y).$$

Hence

$$\|x - y\| - (1 - \mu)\|y\| \leq \|x - y + (1 - \mu)y\| \leq (1 - \lambda\mu)\|x + y\|$$

and therefore

$$\|x-y\| \leq (1-\lambda\mu)(\|x+y\|+\|y\|) \leq 3(1-\lambda\mu).$$

This proves

$$\|x-y\| \leq 3(1-\lambda(x,y)\cdot\lambda(y,x)),$$

and hence the desired inequality.

(3) To prove the above-mentioned statement

$$d(T^m(x),x^*) \leq \varepsilon \quad \text{for all } x \in U,$$

we must show that for some $x^* \in U$ and for $\varepsilon > 0$ given there exists $m = m(\varepsilon)$ such that $\lambda(x^*,T^m(x))\cdot\lambda(T^m(x),x^*) \geq 1-\varepsilon$ for all $x \in U$. Since K is a closed and normal cone in the Banach space $(E,\|\cdot\|)$, K does not contain affine lines, order intervals (for $\leq$ induced by K) are bounded, and K is sequentially complete in the norm topology. Since f is ascending for $\|\cdot\|$ we may apply Theorem 3 in [4] to obtain for the equation $f(x) = \lambda x$ a unique solution $x^* \in U, \lambda^* > 0$. Consider $x, y \in U$ such that $\lambda x \leq y, \mu y \leq x$ with $\lambda, \mu \in [0,1]$. Because f is ascending, it follows that

$$\varphi(\lambda)f(x) \leq f(y), \qquad \varphi(\mu)f(y) \leq f(x)$$

for some continuous function $\varphi : [0,1] \to [0,1], \varphi(r) > r$ for $0 < r < 1$. By the definition of $\lambda(.,.)$ and using the definition of ψ in Lemma 2, it follows that

$$\begin{aligned}&\lambda(f(x),f(y))\cdot\lambda(f(y),f(x))\\ &\geq \sup\{\varphi(\lambda)\cdot\varphi(\mu)|0 \leq \lambda \leq \lambda(x,y), 0 \leq \mu \leq \lambda(y,x)\} \geq \psi(\lambda(x,y)\cdot\lambda(y,x)).\end{aligned}$$

Furthermore, by Lemma 2, $\psi(r) > r$ for $0 < r < 1$ and $\psi(r) \leq \psi(s)$ for $r \leq s$. Define for $x \in U, n \in \mathbb{N}$,

$$a_n(x) = \lambda(x^*,T^n(x))\cdot\lambda(T^n(x),x^*).$$

From $T(x^*) = f(x^*)/\|f(x^*)\| = \lambda^*x^*/\|\lambda^*x^*\| = x^*$ it follows that

$$a_{n+2}(x) = \lambda(T^{n+2}(x^*),T^{n+2}(x))\cdot\lambda(T^{n+2}(x),T^{n+2}(x^*)).$$

By induction we show that

$$a_{n+2}(x) \geq \psi^n(a_2(x)) \quad \text{for } n \in \mathbb{N} \quad (\psi^n \text{ the } n\text{th iterate of } \psi).$$

For $n = 0$ this inequality holds trivially. From $\lambda(f(x),f(y))\cdot\lambda(f(y),f(x)) \geq \psi(\lambda(y,x)\cdot\lambda(y,x))$ it follows that $\lambda(T(x),T(y))\cdot\lambda(T(y),T(x)) \geq \psi(\lambda(x,y)\cdot\lambda(y,x))$. Hence

$$a_{n+1+2}(x) \geq \psi(a_{n+2}(x)) \geq \psi(\psi^n(a_2(x))) = \psi^{n+1}(a_2(x))$$

by using the induction hypothesis and the monotonicity of ψ. This completes the induction. Next, we show that for some constant $s, 0 < s \leq 1, a_2(x) \geq s$ for all $x \in U$. From $au \leq f(x) \leq bv$, where $u, v \in U$ and $0 < a \leq b$ for all $x \in U$ it follows that $T(x) = f(x)/\|f(x)\| \leq (b/\|f(x)\|)v \leq (b/a)v$ and, since f is ascending,

$f(T(x)) \leq (b/a)f(v)$. By condition (ii) in the definition of an ascending operator, there exists some $c_1 \geq 1$ such that $f(v) \leq c_1 f(x^*) = c_1\lambda^* x^*$, and hence

$$T^2(x) = \frac{f(T(x))}{\|f(T(x))\|} \leq \frac{(b/a)f(v)}{a} \leq \frac{bc_1\lambda^*}{a^2}x^*.$$

That is, $\lambda(T^2(x), x^*) \geq a^2/bc_1\lambda^* > 0$ for all $x \in U$. Similarly, $T(x) = f(x)/\|f(x)\| \geq au/\|f(x)\| \geq (a/b)u$, and $f(T(x)) \geq (a/b)f(u)$.

Since $f(u) \geq c_2 f(x^*) = c_2\lambda^* c^*$ for some $0 < c_2 < 1$, it follows that

$$T^2(x) = \frac{f(T(x))}{\|f(T(x))\|} \geq \frac{(a/b)f(u)}{b} \geq \frac{ac_2\lambda^*}{b^2}x^*;$$

hence $\lambda(x^*, T^2x) \geq ac_2\lambda^*/b^2$. Thus, by setting $s = a^3/b^3 \cdot c_2/c_1, a_2(x) \geq s$ for all $x \in U$, and $0 < s \leq 1$. From $a_{n+2}(x) \geq \psi^n(a_2(x))$ and the monotonicity of ψ it follows that

$$a_{n+2}(x) \geq \psi^n(s) \quad \text{for all } x \in U \text{and } n \in \mathbb{N}.$$

Consider the sequence $(b_n)_{n\geq 1}$ where $b_n = \psi^n(s)$. $b_{n+1} = \psi(\psi^n(s)) \geq \psi^n(s) = b_n$, because of $\psi(r) \geq r$ on $[0,1]$, and hence $(b_n)_{n\geq 1}$ converges to some $b \in [0,1]$. By the continuity of $\psi, \psi(b) = b$, which is possible only for $b = 0$ or $b = 1$. If $b = 0$, then $\psi(s) = b_1 = 0$, which contradicts $s > 0$. Hence $b = 1$. Thus we arrive at the conclusion that for $\varepsilon > 0$ there exists $n(\varepsilon) \in \mathbb{N}$ such that

$$a_{n+2}(x) \geq 1 - \varepsilon \quad \text{for all } x \in U \text{and } n \geq n(\varepsilon).$$

(4) Now, from steps (2) and (3) we obtain $\|T^{n+2}(x) - x^*\| \leq 3(1 - a_{n+2}(x)) \leq 3\varepsilon$ for all $x \in U$ and all $n \geq n(\varepsilon)$. Setting $m = n(\varepsilon) + 2$, together with step (1) we obtain $\|T_{n+m} \circ T_{n+m-1} \circ \cdots \circ T_{n+1}(x) - x^*\| \leq 4\varepsilon$ for all $x \in U$ and all $n \geq N(\varepsilon, n(\varepsilon) + 2)$. Finally, choose a starting point $x_0 \in U$ and set $x = T_n \circ T_{n-1} \circ \cdots \circ T_1(x_0)$. Then $\|T_k \circ T_{k-1} \circ \cdots \circ T_1(x_0) - x^*\| \leq 4\varepsilon$ for all $k \geq N(\varepsilon, n(\varepsilon)+2)+n(\varepsilon)+2$ and all $x_0 \in U$. This shows the uniform convergence of $T_n \circ T_{n-1} \circ \cdots \circ T_1(x)$ on U to x^* for $n \to \infty$. In case the f_n are ray-preserving, one has for i, j arbitrary

$$T_i \circ T_j(x) = \frac{f_i\left(f_j(x)/\|f_j(x)\|\right)\|}{\|f_i\left(f_j(x)/\|f_j(x)\|\right)\|} = \frac{\lambda_{ij} f_i \circ f_j(x)}{\|\lambda_{ij} f_i \circ f_j(x)\|} = \frac{f_i \circ f_j(x)}{\|f_i \circ f_j(x)\|}$$

with certain $\lambda_{ij} > 0$. Also due to the ray-preserving property, we may admit arbitrary starting point in $K\backslash\{0\}$. This proves the last statement of the theorem. □

Remark 1. Step (3) in the proof of the theorem shows that it is sufficient to require some iterate of f to be ascending (for the given norm), provided f is ray-preserving.

Remark 2. As a special case one may admit in the theorem that $f_n = f$ for all n. The theorem then yields for a single operator $f : K \to K$, which is uniformly continuous on U, U-bounded and ascending, that

$$\lim_{n\to\infty} T^n(x) = x^* \quad \text{and} \quad \lim_{n\to\infty} \frac{f^n(x)}{\|f^n(x)\|} = x^*, \quad \text{uniformly in } x.$$

(Cf. also Theorem 3 in [4].) When specialized to the standard cone in finite dimensions, the theorem yields Theorem 6 in [2].

COROLLARY *Let $(f_n)_{n\geq 1}$ be a sequence of operators $f_n : K \to K, f_n(x) \neq 0$ for $x \neq 0$, which converges uniformly on U to some operator $f : K \to K$, which is*

uniformly continuous on U, concave, and satisfies $ae \le f(x) \le be$ for some $0 < a \le b$, some $0 \ne e \in K$, and all $x \in U$. Then the conclusions of the theorem hold.

Proof. By assumption, f is U-bounded. To apply the theorem, it remains to show that f is ascending (for the norm). For $x, y \in K$ with $x \le y, y = (1-(1/n))x + (1/n)(n(y-x)+x)$ for all $n \in \mathbb{N}$ and, by the concavity of f, it follows that $f(y) \ge (1-(1/n))f(x)$ for all $n \in \mathbb{N}$. Since K is closed, $f(y) \ge f(x)$, showing that f is increasing on K. Suppose, $\lambda x \le y$ with $x, y \in U$ and $0 \le \lambda < 1$. If $z = y - \lambda x \in K$, then $y = \lambda x + (1-\lambda)(z/(1-\lambda))$, and concavity implies $f(y) \ge \lambda f(x) + (1-\lambda)f(z/(1-\lambda))$. From $y = \lambda x + z$ it follows that $1 - \lambda \le \|z\|$, and, since f is increasing, $f(z/(1-\lambda)) \ge f(z/\|z\|)$. By assumption, $f(z/\|z\|) \ge (a/b)f(x)$, and hence $f(y) \ge \lambda f(x) + (1-\lambda)(a/b)f(x) = (\lambda + (1-\lambda)(a/b))f(x)$. Since f is increasing, this formula holds trivially for $\lambda = 1$. Thus f is ascending with $\varphi(\lambda) = \lambda + (1-\lambda)(a/b)$. □

4. Examples. The following two examples are concerned with $E = C(T)$, the space of all continuous real-valued functions on a compact (Hausdorff-)space T, equipped with the sup-norm $\|\cdot\|$. $K = C_+(T) = \{x \in C(T) \mid x(t) \ge 0 \text{ for all } t \in T\}$ is a closed cone in E and the sup-norm is monotone on K. Furthermore, $U = \{x \in C_+(T) \mid \|x\| = 1\}$.

4.1. Linear operators. Let $f_n : C(T) \to C(T), n \in \mathbb{N}$ be a sequence of linear operators mapping $C_+(T)$ into itself and converging uniformly on U to some continuous linear operator f on $C(T)$. Suppose for f that

$$c = \inf\{f(x)(t) \mid x \in U, t \in T\} > 0.$$

Obviously, f maps $C_+(T)$ into itself. Being linear, the continuous operator f is uniformly continuous on U. Let $e(t) = c$ for all $t \in T$ and let 1 denote the function being constant 1 on T. There exists some $b > 0$ such that $f(1) \le be$. Hence for $x \in U$,

$$e \le f(x) \le f(1) \le be.$$

Thus the corollary yields a unique eigenfunction $x^* \in C_+(T), \|x^*\| = 1$, $f(x^*) = \lambda^* x^*, \lambda^* > 0$, and

$$\lim_{n\to\infty} \frac{f_n \circ f_{n-1} \circ \cdots \circ f_1(x)}{\|f_n \circ f_{n-1} \circ \cdots \circ f_1(x)\|} = x^*$$

uniformly on $C_+(T)\backslash\{0\}$. For the case of a finite T, i.e., $C(T) = \mathbb{R}^N$, this result is well known as strong ergodicity for inhomogeneous Markov chains (or, more general, for nonnegative matrices; cf. [6]). In that special case, the mappings f_n, f can be represented by nonnegative matrices A_n and A, respectively. The assumptions are made to reduce to that of elementwise convergence of A_n to a strictly positive matrix. (It would be sufficient to require some power of A to be strictly positive; see Remark 1.)

4.2. Min-max combinations of linear operators. The stability for inhomogeneous iterations of linear operators as seen above can be extended to the nonlinear operators obtained by forming finite minima or maxima of linear operators. To derive this directly from the result about linear operators would be rather clumsy, but it can be derived easily as follows by using the corollary and the theorem, respectively. It is enough to consider the minimum or the maximum of two operators.

Let $f_n, g_n : C(T) \to C(T), n \in \mathbb{N}$ be two sequences of linear operators with $f_n \to f, g_n \to g$ uniformly on U for $n \to \infty$ and satisfying the assumptions made in

§4.1. Define $h_n, h : C(T) \to C(T)$ by $h_n(x)(t) = \min\{f_n(x)(t), g_n(x)(t)\}, h(x)(t) = \min\{f(x)(t), g(x)(t)\} (x \in C(T), t \in T)$. Still, $h_n \to h$ uniformly on U for $n \to \infty$ and h is uniformly continuous on U. Although no longer linear, h_n and h are ray-preserving and concave. Furthermore, as in §4.1, we have $e \leq f(x) \leq be$ and $e' \leq g(x) \leq b'e'$, and hence $\min\{e, e'\} \leq h(x) \leq \max\{b, b'\} \min\{e, e'\}$. The corollary therefore yields

$$\lim_{n\to\infty} \frac{h_n \circ h_{n-1} \circ \cdots \circ h_1(x)}{\|h_n \circ h_{n-1} \circ \cdots \circ h_1(x)\|} = x^*$$

uniformly on $C_+(T)\backslash\{0\}$. The corresponding case for the maximum, however, is not covered by the corollary, because the maximum of two linear operators is a convex operator. In this case we need to go back to the theorem. Let

$$h_n(x)(t) = \max\{f_n(x)(t), g_n(x)(t)\}, \qquad h(x)(t) = \max\{f(x)(t), g(x)(t)\}.$$

As in the case of the minimum, $h_n \to h$ uniformly on U for $n \to \infty$, h is uniformly continuous on U, and there holds an inequality $e \leq h(x) \leq be$ with $0 \neq e \in C_+(T), b > 0$. Hence h is U-bounded. To apply the theorem, it remains to show that h is ascending. From the assumptions on f and g it follows, as in the proof of the corollary, that $\lambda x \leq y$ for $x, y \in U, \lambda \in [0, 1]$ implies $\varphi(\lambda)f(x) \leq f(y)$ and $\psi(\lambda)g(x) \leq g(y)$ with $\varphi, \psi : [0, 1] \to [0, 1]$ continuous and $\varphi(\lambda) > \lambda, \psi(\lambda) > \lambda$ for $0 < \lambda < 1$. Therefore,

$$\min\{\varphi(\lambda), \psi(\lambda)\}\ h(x) \leq h(y),$$

showing that h is ascending.

This, too, can be specialized to the case of nonnegative matrices. In the finite-dimensional case the minimum of linear operators and the related stability problem for inhomogeneous iterations appears in applications to economics [1].

Acknowledgment. The authors would like to thank the referee for his useful remarks.

REFERENCES

[1] T. FUJIMOTO AND U. KRAUSE, *Ergodic price setting with technical progress*, in Competition, Instability, and Nonlinear Cycles, W. Semmler, ed., Springer-Verlag, Berlin, 1986.

[2] ———, *Asymptotic properties for inhomogeneous iterations of nonlinear operators*, SIAM J. Math. Anal., 19 (1988), pp. 841–853.

[3] H. INABA, *Weak ergodicity of population evolution processes*, Math. Biosci., 96 (1989), pp. 195–219.

[4] U. KRAUSE, *A nonlinear extension of the Birkhoff–Jentzsch theorem*, J. Math. Anal. Appl., 114 (1986), pp. 552–568.

[5] R. D. NUSSBAUM, *Some nonlinear weak ergodic theorems*, SIAM J. Math. Anal., 21 (1990), pp. 436–460.

[6] E. SENETA, *Non-Negative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, Berlin, 1980.

[7] H. R. THIEME, *Asymptotic proportionality (weak ergodicity) and conditional asymptotic equality of solutions to time-heterogeneous sublinear difference and differential equations*, J. Differential Equations, 73 (1988), pp. 237–268.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1203–1226, July 1994

012

CANONICAL FORMS OF DIFFERENTIAL EQUATIONS FREE FROM ACCESSORY PARAMETERS*

YOSHISHIGE HARAOKA†

Abstract. Systems of differential equations free from accessory parameters are defined and studied by Okubo [Seminar Reports of Tokyo Metropolitan University, 1987]. They are Fuchsian on the complex projective line, and there is an algorithm of determining monodromy representations for such systems. The Gauss hypergeometric equation, the generalized hypergeometric equation, the Pochhammer equation and a one-dimensional section of the Appell hypergeometric system F_3 are known to be reduced to such systems. Recently Yokoyama classified all the systems of differential equations which are irreducible and free from accessory parameters in terms of multiplicities of characteristic exponents. This paper presents canonical forms of all such systems and will define a new class of special functions.

Key words. systems of Okubo normal form, accessory parameters, characteristic exponents, hypergeometric functions

AMS subject classifications. 33C20, 33C65, 33E30

The Gauss hypergeometric function is one of the most important special functions, and there have been many efforts to extend it. Thus obtained are the generalized hypergeometric function and the Pochhammer function, which are functions of one variable, the Appell–Lauricella hypergeometric functions and the Aomoto–Gelfand hypergeometric functions, which are functions of several variables, and so on; they are also interesting special functions. Note that together with the Gauss hypergeometric function, they are characterized as solutions of linear differential equations. Then, if we find a *good* differential equation, it will define a new special function. From this point of view, we follow the Okubo theory and then determine a class of extensions of the Gauss hypergeometric differential equation.

In [5] Okubo developed a global theory of Fuchsian differential equations on the complex projective line $\mathbf{P}^1(\mathbf{C})$. The theory consists of the following three parts: (i) Reduction to a normal form: *Every Fuchsian differential equation on* $\mathbf{P}^1(\mathbf{C})$ *is reduced to a system of a normal form which we call the Okubo normal form.* (ii) Definition of systems free from accessory parameters. (iii) Algorithm of determining monodromy representations for systems free from accessory parameters. Once a system free from accessory parameters is given, we can apply this theory to find a monodromy representation for the system (e.g., [7] and [8]). The Gauss hypergeometric equation, the generalized hypergeometric equation ${}_nE_{n-1}$, and the Pochhammer equation are free from accessory parameters. Moreover it is known that a one-dimensional section of the Appell hypergeometric function F_3 satisfies a system of Okubo normal form of rank 4 that is free from accessory parameters [1], [7]. Then it is natural to ask what is the whole set of systems free from accessory parameters. Recently Yokoyama has classified the set of irreducible systems free from accessory parameters [9]. Using this result, we determine all systems which are irreducible and free from accessory parameters. Thus a new class of extensions of the Gauss hypergeometric function is defined.

In §1 we review part (ii) of the Okubo theory, introduce the result by Yokoyama,

*Received by the editors May 26, 1992; accepted for publication (in revised form) May 25, 1993.
†Graduate School of Science and Technology, Kumamoto University, Kumamoto, 860, Japan

and give the canonical forms of irreducible systems free from accessory parameters in Theorems I, I*, II, II*, III, III*, IV, and IV* according to Yokoyama's classification. These theorems are proved in §3. Lemmas and propositions employed in proving the theorems are collected in §2. The main tool in §2 is expansion in partial fractions of rational functions, which we owe to the work of Kimura and Okamoto [3].

Notation. I_k: the identity matrix of size k, for $k \in \mathbf{N}$. $M(k,l)$: the set of $k \times l$ matrices with entries in $\mathbf{C}$, for $k, l \in \mathbf{N}$.

1. Systems free from accessory parameters. Let n be a fixed positive integer. We consider a system of differential equations

$$(xI_n - T)\frac{dY}{dx} = AY \tag{1.1}$$

on $\mathbf{P}^1(\mathbf{C})$ of rank n, which we call a *system of Okubo normal form,* where

$$\begin{aligned} &T = t_1 I_{n_1} \oplus \cdots \oplus t_p I_{n_p}, \\ &\quad t_i \in \mathbf{C}\,(1 \le i \le p), \qquad t_i \ne t_j\,(i \ne j), \\ &\quad n_1 + \cdots + n_p = n, \\ &A \in \mathrm{End}\,(n, \mathbf{C})\,. \end{aligned}$$

We denote the partition $(n_1, \ldots, n_p)$ of n by Δ. Corresponding to the partition Δ, we decompose A into $(n_1, \ldots, n_p)$ blocks:

$$A = (A_{ij})_{1 \le i,j \le p}, \qquad A_{ij} \in M(n_i, n_j i\mathbf{C}).$$

We denote the set of eigenvalues of A_{ii} by Λ_i for $i = 1, \ldots, p$, and the set of eigenvalues of A by Λ_∞. By virtue of the invariance of the trace of the matrix A we have

$$\sum_{i=1}^{p} \sum_{\lambda \in \Lambda_i} \lambda = \sum_{\rho \in \Lambda_\infty} \rho. \tag{1.2}$$

The system (1.1) is Fuchsian over $\mathbf{P}^1(\mathbf{C})$ with regular singular points at $x = t_1, \ldots, t_p$, ∞. The set of the characteristic exponents at $x = t_i$ is Λ_i for $i = 1, \ldots, p$, and the set of the characteristic exponents at $x = \infty$ is Λ_∞. The relation (1.2) is the Riemann–Fuchs relation ([5, Chap. II, Thm. 1.1.]).

With the partition Δ we associate the subgroup G_Δ of $\mathrm{GL}(n, \mathbf{C})$ by

$$G_\Delta = \mathrm{GL}\,(n_1, \mathbf{C}) \oplus \cdots \oplus \mathrm{GL}\,(n_p, \mathbf{C})\,.$$

By a G_Δ-valued gauge transformation

$$Y = PZ, \qquad P \in G_\Delta,$$

(1.1) is transformed into

$$(xI_n - T)\frac{dZ}{dx} = P^{-1}APZ.$$

Note that $\Lambda_1, \ldots, \lambda_p, \Lambda_\infty$ are invariant under this transformation.

We assume the following:

(1.3) A is diagonalizable.

(1.4) A_{ii} is diagonalizable for every $i = 1, \dots, p$.

DEFINITION 1. We assume (1.3) and (1.4). If the system (1.1) is uniquely determined by $T, \Lambda_1, \dots, \Lambda_p, \Lambda_\infty$ up to G_Δ-valued gauge transformations, (1.1) is said to be *free from accessory parameters.*

Let $L = (m_1, \dots, m_q)$ denote the multiplicities of the eigenvalues of A:

$$\Lambda_\infty = (\underbrace{\rho_1, \dots, \rho_1}_{m_1}, \dots, \underbrace{\rho_q, \dots, \rho_q}_{m_q}), \qquad \rho_i \neq \rho_j \ (i \neq j), \quad m_1 + \cdots + m_q = n.$$

We assume further that

(1.5) For every $i = 1, \dots, p$, the entries of Λ_i are mutually distinct.

Then (1.5) implies (1.4). Define the integer $N(\Delta, L)$ by

$$N(\Delta, L) = n^2 - n + 2 - \sum_{i=1}^{p} n_i^2 - \sum_{j=1}^{q} m_j^2.$$

Then Okubo showed that, on the assumptions (1.3) and (1.5), (1.1) is free from accessory parameters if $N(\Delta, L) = 0$ [5, Chap. 2, Thm. 1.2].

DEFINITION 2. A system of differential equations

$$\frac{dY}{dx} = A(x)Y, \qquad A(x) \in \operatorname{End}(n, \mathbf{C}(x)),$$

is said to be *reducible* if there is a transformation $Y = P(x)Z, P(x) \in \mathrm{GL}(n, \mathbf{C}(x))$, such that the coefficient matrix $B(X)$ of the transformed system

$$\frac{dZ}{dx} = B(x)Z$$

is of block triangular form. Otherwise the system is said to be *irreducible.*

Now we introduce the result by Yokoyama.

THEOREM (Yokoyama [9, Thm. 2]). *If the system* (1.1) *with the assumptions* (1.3) *and* (1.5) *is irreducible, and if* $N(\Delta, L) = 0$, *then one of the following holds.*

(I) $\Delta = (n-1, 1), \quad L = (1, 1, \dots, 1),$

(I*) $\Delta = (1, 1, \dots, 1), \quad L = (n-1, 1),$

(II) $\Delta = (m, m), \quad L = (m, m-1, 1),$

where $n = 2m$ *is an even integer equal to or greater than* 4.

(II*) $\Delta = (m, m-1, 1), \quad L = (m, m),$

where $n = 2m$ *is an even integer equal to or greater than* 4.

(III) $\Delta = (m+1, m), \quad L = (m, m, 1),$

where $n = 2m+1$ *is an odd integer equal to or greater than* 5.

(III*) $$\Delta = (m, m, 1), \qquad L = (m+1, m),$$

where $n = 2m+1$ *is an odd integer equal to or greater than* 5.

(IV) $$\Delta = (4, 2), \qquad L = (2, 2, 2),$$

where $n = 6$.

(IV*) $$\Delta = (2, 2, 2), \qquad L = (4, 2),$$

where $n = 6$.

For each case in the above theorem, we give a canonical form of the system (1.1).

THEOREM I. *The case is* $\Delta = (n-1, 1), L = (1, 1, \dots, 1)$. *Set* $T = t_1 I_{n-1} \oplus t_2, \Lambda_1 = (\lambda_1, \dots, \lambda_{n-1}), \Lambda_2 = (\mu)$ *and* $\Lambda_\infty = (\rho_1, \dots, \rho_n)$, *where*

(F$_\mathrm{I}$) $$\sum_{i=1}^{n-1} \lambda_i + \mu = \sum_{j=1}^{n} \rho_j.$$

We assume (1.5) *and*

(1.6) $\lambda_i \neq \rho_j$ *for every* $i \in \{1, \dots, n-1\}$ *and* $j \in \{1, \dots, n\}$.

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T)\frac{dZ}{dx} = A_\mathrm{I} Z$$

with

$$A_\mathrm{I} = \begin{pmatrix} \lambda_1 & & & 1 \\ & \ddots & & \vdots \\ & & \lambda_{n-1} & 1 \\ \beta_1 & \cdots & \beta_{n-1} & \mu \end{pmatrix},$$

where

$$\beta_i = -\frac{\prod_{1 \le j \le n} (\lambda_i - \rho_j)}{\prod_{\substack{1 \le k \le n-1 \\ k \neq i}} (\lambda_i - \lambda_k)}, \qquad i = 1, \dots, n-1.$$

THEOREM I*. *The case is* $\Delta = (1, 1, \dots, 1), L = (n-1, 1)$. *Set* $T = t_1 \oplus t_2 \oplus \cdots \oplus t_n, \Lambda_i = (\lambda_i), i = 1, \dots, n$ *and* $\Lambda_\infty = (\underbrace{\rho_1, \dots, \rho_1}_{n-1}, \rho_2)$, *where*

(F$_{\mathrm{I}^*}$) $$\sum_{i=1}^{n} \lambda_i = (n-1)\rho_1 + \rho_2.$$

We assume (1.5) *and*

(1.7) $\lambda_i \neq \rho_j$ *for every* $i \in \{1, \dots, n\}$ *and* $j \in \{1, 2\}$.

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T)\frac{dZ}{dx} = A_{\mathrm{I}^*}Z$$

with

$$A_{\mathrm{I}^*} = \begin{pmatrix} \lambda_1 & \lambda_1 - \rho_1 & \cdots & \lambda_1 - \rho_1 \\ \lambda_2 - \rho_1 & \lambda_2 & \cdots & \lambda_2 - \rho_1 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_n - \rho_1 & \lambda_n - \rho_1 & \cdots & \lambda_n \end{pmatrix}.$$

THEOREM II. *The case is* $\Delta = (m, m), L = (m, m-1, 1)$, *where* $n = 2m$ *is an even integer equal to or greater than* 4. *Set* $T = t_1 I_m \oplus t_2 I_m, \Lambda_1 = (\lambda_1, \ldots, \lambda_m), \Lambda_2 = (\mu_1, \ldots, \mu_m)$ *and* $\Lambda_\infty = (\underbrace{\rho_1, \ldots, \rho_1}_{m}, \underbrace{\rho_2, \ldots, \rho_2}_{m-1}, \rho_3)$, *where*

$$(\mathrm{F_{II}}) \qquad \sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i = m\rho_1 + (m-1)\rho_2 + \rho_3.$$

We assume (1.5) *and*

$$(1.8) \qquad \begin{cases} \lambda_i \neq \rho_1, \mu_i \neq \rho_1 \quad \text{for every } i \in \{1, \ldots, m\}, \\ \lambda_i + \mu_k \neq \rho_1 + \rho_2 \quad \text{for every } i, k \in \{1, \ldots, m\}. \end{cases}$$

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T)\frac{dZ}{dx} = A_{\mathrm{II}}Z$$

with

$$A_{\mathrm{II}} = \left(\begin{array}{ccc|ccc} \lambda_1 & & & & & \\ & \ddots & & & (\alpha_{ij}) & \\ & & \lambda_m & & & \\ \hline & & & \mu_1 & & \\ & (\beta_{ij}) & & & \ddots & \\ & & & & & \mu_m \end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1) \prod_{\substack{1 \leq k \leq m \\ k \neq i}} \left(\frac{\lambda_k + \mu_j - \rho_1 - \rho_2}{\lambda_i - \lambda_k}\right),$$

$$\beta_{ij} = (\mu_i - \rho_1) \prod_{\substack{1 \leq k \leq m \\ k \neq i}} \left(\frac{\mu_k + \lambda_j - \rho_1 - \rho_2}{\mu_i - \mu_k}\right), \qquad i, j = 1, \ldots, m.$$

THEOREM II*. *The case is* $\Delta = (m, m-1, 1), L = (m, m)$, *where* $n = 2m$ *is an even integer equal to or greater than* 4. *Set* $T = t_1 I_m \oplus t_2 I_{m-1} \oplus t_3, \Lambda_1 = (\lambda_1, \ldots, \lambda_m), \Lambda_2 = (\mu_1, \ldots, \mu_{m-1}), \Lambda_3 = (\nu)$ *and* $\Lambda_\infty = (\underbrace{\rho_1, \ldots, \rho_1}_{m}, \underbrace{\rho_2, \ldots, \rho_2}_{m})$, *where*

$$(\mathrm{F_{II^*}}) \qquad \sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m-1} \mu_i + \nu = m\rho_1 + m\rho_2.$$

We assume (1.5) *and*

$$(1.9)\qquad \begin{cases} \lambda_i \neq \rho_j \quad \text{for every } i \in \{1,\dots,m\} \text{ and } j \in \{1,2\}, \\ \lambda_i + \mu_k \neq \rho_1 + \rho_2 \quad \text{for every } i \in \{1,\dots,m\} \text{ and } k \in \{1,\dots,m-1\}. \end{cases}$$

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T)\frac{dZ}{dx} = A_{\mathrm{II}^*} Z$$

with

$$A_{\mathrm{II}^*} = \left(\begin{array}{ccc|ccccc} \lambda_1 & & & & & & & \\ & \ddots & & & (\alpha_{ij}) & & & \\ & & \lambda_m & & & & & \\ \hline & & & \mu_1 & & & & \xi_1 \\ & (\beta_{ij}) & & & \ddots & & & \vdots \\ & & & & & & \mu_{m-1} & \xi_{m-1} \\ & & & \eta_1 & \cdots & & \eta_{m-1} & \nu \end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1) \prod_{\substack{1 \le \ell \le m \\ \ell \neq i}} \left(\frac{\lambda_\ell + \mu_j - \rho_1 - \rho_2}{\lambda_i - \lambda_\ell}\right), \qquad i = 1,\dots,m,\ j = 1,\dots,m-1,$$

$$\alpha_{im} = (\lambda_i - \rho_1) \prod_{\substack{1 \le \ell \le m \\ \ell \neq i}} \frac{1}{\lambda_i - \lambda_\ell}, \qquad i = 1,\dots,m,$$

$$\beta_{ij} = (\lambda_j - \rho_2) \prod_{\substack{1 \le k \le m-1 \\ k \neq i}} \left(\frac{\lambda_j + \mu_k - \rho_1 - \rho_2}{\mu_i - \mu_k}\right), \qquad i = 1,\dots,m-1,\ j = 1,\dots,m,$$

$$\beta_{mj} = -(\lambda_j - \rho_2) \prod_{k=1}^{m-1} (\lambda_j + \mu_k - \rho_1 - \rho_2), \qquad j = 1,\dots,m,$$

$$\xi_i = \prod_{\substack{1 \le k \le m-1 \\ k \neq i}} \frac{1}{\mu_i - \mu_k}, \qquad i = 1,\dots,m-1,$$

$$\eta_j = -\prod_{\ell=1}^{m} (\lambda_\ell + \mu_j - \rho_1 - \rho_2), \qquad j = 1,\dots,m-1.$$

THEOREM III. *The case is* $\Delta = (m+1, m)$, $L = (m, m, 1)$, *where* $n = 2m+1$ *is an odd integer equal to or greater than* 5. *Set* $T = t_1 I_{m+1} \oplus t_2 I_m$, $\Lambda_1 = (\lambda_1,\dots,\lambda_{m+1})$, $\Lambda_2 = (\mu_1,\dots,\mu_m)$ *and* $\Lambda_\infty = (\underbrace{\rho_1,\dots,\rho_1}_{m}, \underbrace{\rho_2,\dots,\rho_2}_{m}, \rho_3)$, *where*

$$(\mathrm{F_{III}})\qquad \sum_{i=1}^{m+1} \lambda_i + \sum_{i=1}^{m} \mu_i = m\rho_1 + m\rho_2 + \rho_3.$$

We assume (1.5) *and*

$$(1.10)\qquad \begin{cases} \lambda_i \neq \rho_j \quad \text{for every } i \in \{1,\dots,m+1\} \text{ and } j \in \{1,2\}, \\ \lambda_i + \mu_k \neq \rho_1 + \rho_2 \quad \text{for every } i \in \{1,\dots,m+1\} \text{ and } k \in \{1,\dots,m\}. \end{cases}$$

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T)\frac{dZ}{dx} = A_{\mathrm{III}} Z$$

with

$$A_{\mathrm{III}} = \left(\begin{array}{ccc|ccc} \lambda_1 & & & & & \\ & \ddots & & & (\alpha_{ij}) & \\ & & \lambda_{m+1} & & & \\ \hline & & & \mu_1 & & \\ & (\beta_{ij}) & & & \ddots & \\ & & & & & \mu_m \end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1) \prod_{\substack{1 \le \ell \le m+1 \\ \ell \neq i}} \left(\frac{\lambda_\ell + \mu_j - \rho_1 - \rho_2}{\lambda_i - \lambda_\ell}\right), \qquad i = 1,\dots,m+1,\ j = 1,\dots,m,$$

$$\beta_{ij} = (\lambda_j - \rho_1) \prod_{\substack{1 \le k \le m \\ k \neq i}} \left(\frac{\lambda_j + \mu_k - \rho_1 - \rho_2}{\mu_i - \mu_k}\right), \qquad i = 1,\dots,m,\ j = 1,\dots,m+1.$$

THEOREM III*. *The case is* $\Delta = (m,m,1), L = (m+1,m)$, *where* $n = 2m+1$ *is an odd integer equal to or greater than* 5. *Set* $T = t_1 I_m \oplus t_2 \oplus t_3 I_m, \Lambda_1 = (\lambda_1,\dots,\lambda_m), \Lambda_2 = (\nu), \Lambda_3 = (\mu_1,\dots,\mu_m)$ *and* $\Lambda_\infty = (\underbrace{\rho_1,\dots,\rho_1}_{m+1}, \underbrace{\rho_2,\dots,\rho_2}_{m})$ *where*

$$(\mathrm{F}_{\mathrm{III}^*})\qquad \sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i + \nu = (m+1)\rho_1 + m\rho_2.$$

We assume (1.5) *and*

$$(1.11)\qquad \begin{cases} \lambda_i \neq \rho_1, \mu_i \neq \rho_1 \quad \text{for every } i \in \{1,\dots,m\}, \\ \lambda_i + \mu_k \neq \rho_1 + \rho_2 \quad \text{for every } i \in \{1,\dots,m\} \text{ and } k \in \{1,\dots,m\}. \end{cases}$$

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T)\frac{dZ}{dx} = A_{\mathrm{III}^*} Z$$

with

$$A_{\mathrm{III}*} = \left(\begin{array}{ccccccc} \lambda_1 & & & \gamma_1 & & & \\ & \ddots & & \vdots & & (\alpha_{ij}) & \\ & & \lambda_m & \gamma_m & & & \\ \zeta_1 & \cdots & \zeta_m & \nu & \eta_1 & \cdots & \eta_m \\ & & & \xi_1 & \mu_1 & & \\ & (\beta_{ij}) & & \vdots & & \ddots & \\ & & & \xi_m & & & \mu_{m-1} \end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1) \prod_{\substack{1 \le k \le m \\ k \ne i}} \left(\frac{\rho_1 + \rho_2 - \lambda_k - \mu_j}{\lambda_i - \lambda_k} \right), \qquad i, j = 1, \dots, m,$$

$$\beta_{ij} = (\mu_i - \rho_1) \prod_{\substack{1 \le k \le m \\ k \ne i}} \left(\frac{\rho_1 + \rho_2 - \lambda_j - \mu_k}{\mu_i - \mu_k} \right), \qquad i, j = 1, \dots, m,$$

$$\gamma_i = \frac{\lambda_i - \rho_1}{\prod_{\substack{1 \le k \le m \\ k \ne i}} (\lambda_k - \lambda_i)}, \qquad i = 1, \dots, m,$$

$$\xi_i = \frac{\mu_i - \rho_1}{\prod_{\substack{1 \le k \le m \\ k \ne i}} (\mu_i - \mu_k)}, \qquad i = 1, \dots, m,$$

$$\zeta_j = \prod_{k=1}^{m} (\rho_1 + \rho_2 - \lambda_j - \mu_k), \qquad j = 1, \dots, m,$$

$$\eta_j = -\prod_{k=1}^{m} (\lambda_k + \mu_j - \rho_1 - \rho_2), \qquad j = 1, \dots, m.$$

THEOREM IV. *The case is* $n = 6, \Delta = (4,2), L = (2,2,2)$. *Set* $T = t_1 I_4 \oplus t_2 I_2$, $\Lambda_1 = (\lambda_1, \lambda_2, \lambda_3, \lambda_4), \Lambda_2 = (\mu_1, \mu_2)$, *and* $\Lambda_\infty = (\rho_1, \rho_1, \rho_2, \rho_2, \rho_3, \rho_3)$, *where*

$$\text{(F}_{\text{IV}}\text{)} \qquad \sum_{i=1}^{4} \lambda_i + \sum_{i=1}^{2} \mu_i = 2\rho_1 + 2\rho_2 + 2\rho_3.$$

We assume (1.5) *and*

(1.12)

$$\begin{cases} \lambda_i \ne \rho_j \quad \text{for every } i \in \{1,2,3,4\} \text{ and } j \in \{1,2,3\}, \\ \lambda_i + \lambda_j + \mu_k \ne \rho_1 + \rho_2 + \rho_3 \quad \text{for } i, j \in \{1,2,3,4\} \text{ with } i \ne j \text{ and } k \in \{1,2\}. \end{cases}$$

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T) \frac{dZ}{dx} = A_{\text{IV}} Z$$

with

$$A_{\text{IV}} = \begin{pmatrix} \lambda_1 & & & & \alpha_{11} & \alpha_{12} \\ & \lambda_2 & & & \alpha_{21} & \alpha_{22} \\ & & \lambda_3 & & \alpha_{31} & \alpha_{32} \\ & & & \lambda_4 & \alpha_{41} & \alpha_{42} \\ \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} & \mu_1 & \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} & & \mu_2 \end{pmatrix},$$

where

$$\alpha_{ij} = \frac{\prod_{\ell=1,2,3}(\lambda_i - \rho_\ell)}{(\mu_j - \mu_{j'})\prod_{\substack{1\le k\le 4\\ k\neq i}}(\lambda_i - \lambda_k)} \cdot a_{ij}, \qquad i = 1,\dots,4,\ j = 1,2, \{j,j'\} = \{1,2\},$$

$$a_{11} = \prod_{k=2}^{4}(\lambda_1 + \lambda_k + \mu_2 - \rho_1 - \rho_2 - \rho_3),$$

$$a_{12} = \prod_{k=2}^{4}(\lambda_1 + \lambda_k + \mu_1 - \rho_1 - \rho_2 - \rho_3),$$

$$a_{ij} = \lambda_1 + \lambda_i + \mu_{j'} - \rho_1 - \rho_2 - \rho_3, \qquad i = 2,3,4,\ j = 1,2, \{j,j'\} = \{1,2\},$$

$$\beta_{11} = \beta_{21} = 1,$$

$$\beta_{ji} = \prod_{\substack{k=2,3,4\\ k\neq i}}(\lambda_1 + \lambda_k + \mu_j - \rho_1 - \rho_2 - \rho_3), \qquad i = 2,3,4,\ j = 1,2.$$

THEOREM IV*. *The case* $n = 6, \Delta = (2,2,2), L = (4,2)$. *Set* $T = t_1 I_2 \oplus t_2 I_2 \oplus t_3 I_2, \Lambda_1 = (\lambda_1, \lambda_2), \Lambda_2 = (\mu_1, \mu_2), \Lambda_3 = (\nu_1, \nu_2)$ *and* $\Lambda_\infty = (\rho_1, \rho_1, \rho_1, \rho_1, \rho_2, \rho_2)$, *where*

$$(\mathrm{F}_{\mathrm{IV}^*}) \qquad \lambda_1 + \lambda_2 + \mu_1 + \mu_2 + \nu_1 + \nu_2 = 4\rho_1 + 2\rho_2.$$

We assume (1.5) *and*

$$(1.13) \qquad \begin{cases} \lambda_i \neq \rho_1, \mu_i \neq \rho_1, \nu_i \neq \rho_1 & \text{for every } i \in \{1,2\}, \\ \lambda_i + \mu_j + \nu_k \neq 2\rho_1 + \rho_2 & \text{for every } i,j,k \in \{1,2\}. \end{cases}$$

Then the system (1.1) *is transformed by a* G_Δ*-valued transformation into*

$$(xI_n - T)\frac{dZ}{dx} = A_{\mathrm{IV}^*} Z$$

with

$$A_{\mathrm{IV}^*} = \begin{pmatrix} \lambda_1 & & \alpha_{13} & \alpha_{14} & \alpha_{15} & \alpha_{16} \\ & \lambda_2 & \alpha_{23} & \alpha_{24} & \alpha_{25} & \alpha_{26} \\ \beta_{11} & \beta_{12} & \mu_1 & & \beta_{15} & \beta_{16} \\ \beta_{21} & \beta_{22} & & \mu_2 & \beta_{25} & \beta_{26} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \nu_1 & \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & & \nu_2 \end{pmatrix},$$

where

$$\alpha_{ij} = \frac{\lambda_i - \rho_1}{\lambda_i - \lambda_{i'}} \cdot a_{ij} \quad \text{for } i = 1,2, \text{with } \{i,i'\} = \{1,2\}, j = 3,4,5,6,$$

$$\beta_{ij} = \frac{\mu_i - \rho_1}{\mu_i - \mu_{i'}} \cdot b_{ij} \quad \text{for } i = 1,2, \text{with } \{i,i'\} = \{1,2\}, j = 1,2,5,6,$$

$$\gamma_{ij} = \frac{\nu_i - \rho_1}{\nu_i - \nu_{i'}} \cdot c_{ij} \quad \text{for } i = 1,2, \text{with } \{i,i'\} = \{1,2\}, j = 1,2,3,4,$$

$$a_{13} = [121], \quad a_{14} = [112], \quad a_{15} = [221], \quad a_{16} = [212],$$
$$a_{23} = [222], \quad a_{24} = [211], \quad a_{25} = [222], \quad a_{26} = [211],$$
$$b_{11} = [211], \quad b_{12} = [112], \quad b_{15} = [112], \quad b_{16} = [122],$$
$$b_{21} = [222], \quad b_{22} = [121], \quad b_{25} = [111], \quad b_{26} = [121],$$
$$c_{11} = [122], \quad c_{12} = [121], \quad c_{13} = [121], \quad c_{14} = [122],$$
$$c_{21} = [111], \quad c_{22} = [112], \quad c_{23} = [111], \quad c_{24} = [122].$$

Here we have set

$$[ijk] = \lambda_i + \mu_j + \nu_k - 2\rho_1 - \rho_2 \tag{1.14}$$

for $i, j, k = 1, 2$.

For each J = I, I*, ..., IV, we call the system of differential equations

$$(xI_n - T)\frac{dY}{dx} = A_J Y$$

the *system* (J). The system (I) is transformed into a single differential equation of order n which is known to be the generalized hypergeometric equation [4], [6]. The system (I*) is also transformed into a single differential equation of order n which is known to be the Pochhammer equation [2]. The system (II*) with $m = 2$ is transformed into a one-dimensional section of Appell's hypergeometric system F_3 [1], [7].

After preparing several lemmas and propositions in §2, we shall prove this theorem in §3.

2. Lemmas and propositions. Throughout this section we assume that $p_1, \ldots, p_m, q_1, \ldots, q_m$ are mutually distinct complex numbers.

LEMMA 1.

$$\sum_{j=1}^{m} \frac{\prod_{\substack{1\le k\le m \\ k\neq i}} (p_j - q_k)}{\prod_{\substack{1\le \ell\le m \\ \ell\neq j}} (p_j - p_\ell)} = 1$$

for $i = 1, \ldots, m$.

Proof. Put $x = q_i$ into the both sides of the partial fraction expansion

$$\frac{\prod_{k=1}^{m} (x - q_k)}{\prod_{j=1}^{m} (x - p_j)} = 1 + \sum_{j=1}^{m} \frac{1}{x - p_j} \cdot \frac{\prod_{k=1}^{m} (p_j - q_k)}{\prod_{\substack{1\le \ell\le m \\ \ell\neq j}} (p_j - p_\ell)}$$

to show the lemma.

LEMMA 2.

$$\sum_{j=1}^{m} \frac{\prod_{k=1}^{m} (p_j - q_k)}{\prod_{\substack{1\le \ell\le m \\ \ell\neq j}} (p_j - p_\ell)} + \sum_{k=1}^{m} q_k - \sum_{\ell=1}^{m} p_\ell = 0.$$

Proof. We use an auxiliary variable q_0. Put $x = q_0$ into the both sides of the partial fraction expansion

$$\frac{\prod_{k=0}^{m} (x - q_k)}{\prod_{j=1}^{m} (x - p_j)} = x + \sum_{\ell=1}^{m} p_\ell - \sum_{k=0}^{m} q_k + \sum_{j=1}^{m} \frac{1}{x - p_j} \cdot \frac{\prod_{k=0}^{m} (p_j - q_k)}{\prod_{\substack{1\le \ell\le m \\ \ell\neq j}} (p_j - p_\ell)}$$

to show the lemma.

PROPOSITION 1. *The inverse matrix of*

$$C = \left(\frac{1}{p_i - q_j}\right)_{1\le i,j\le m}$$

is $C^{-1} = (\gamma_{ij})$ *with*

$$\gamma_{ij} = \frac{\prod_{k=1}^{m} (p_j - q_k)}{\prod_{\substack{1\le \ell\le m \\ \ell\neq j}} (p_j - p_\ell)} \cdot \frac{\prod_{\substack{1\le \ell\le m \\ \ell\neq j}} (q_i - p_\ell)}{\prod_{\substack{1\le k\le m \\ k\neq i}} (q_i - q_k)}$$

for $i, j = 1, \dots, m$.

Proof. Put $x = p_{i'}$ into the both sides of the partial fraction expansion

$$\frac{\prod_{\substack{1 \le \ell \le m \\ \ell \ne i}} (x - p_\ell)}{\prod_{k=1}^{m} (x - q_k)} = \sum_{j=1}^{m} \frac{1}{x - q_j} \cdot \frac{\prod_{\substack{1 \le \ell \le m \\ \ell \ne i}} (q_j - p_\ell)}{\prod_{\substack{1 \le k \le m \\ k \ne j}} (q_j - q_k)}$$

to show that

$$\sum_{j=1}^{m} \frac{1}{p_{i'} - q_j} \cdot \gamma_{ji} = \delta_{ii'},$$

which proves the proposition.

The following proposition can be shown in a similar manner.

PROPOSITION 2. *The inverse matrix of*

$$C = \begin{pmatrix} \frac{1}{p_1 - q_1} & \cdots & \frac{1}{p_1 - q_m} \\ \vdots & \cdots & \vdots \\ \frac{1}{p_{m-1} - q_1} & \cdots & \frac{1}{p_{m-1} - q_m} \\ 1 & \cdots & 1 \end{pmatrix}$$

is $C^{-1} = (\gamma_{ij})$ *with*

$$\gamma_{ij} = \frac{\prod_{k=1}^{m} (p_j - q_k)}{\prod_{\substack{1 \le \ell \le m-1 \\ \ell \ne j}} (p_j - p_\ell)} \cdot \frac{\prod_{\substack{1 \le \ell \le m-1 \\ \ell \ne j}} (q_i - p_\ell)}{\prod_{\substack{1 \le k \le m \\ k \ne i}} (q_i - q_k)},$$

$$\gamma_{im} = \frac{\prod_{\ell=1}^{m-1} (q_i - p_\ell)}{\prod_{\substack{1 \le k \le m \\ k \ne i}} (q_i - q_k)},$$

for $i = 1, \dots, m, j = 1, \dots, m-1$.

PROPOSITION 3. *Let*

$$P = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_m \end{pmatrix}, \qquad Q = \begin{pmatrix} q_1 & & \\ & \ddots & \\ & & q_m \end{pmatrix}$$

be diagonal matrices. Suppose that there are

$$\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_m \end{pmatrix}, \quad \eta = (\eta_1 \cdots \eta_m), \quad U \in \mathrm{GL}(m, \mathbf{C})$$

satisfying

$$U^{-1} Q U + \xi \eta = P. \tag{2.1}$$

Then we have

$$\xi_j \eta_j = \frac{\prod_{k=1}^{m} (p_j - q_k)}{\prod_{\substack{1 \le \ell \le m \\ \ell \ne j}} (p_j - p_\ell)}, \qquad j = 1, \dots, m, \tag{2.2}$$

so that $\xi_j \neq 0$ *for every* j. *Moreover we have*

$$U = DC^{-1} \begin{pmatrix} \xi_1 & & \\ & \ddots & \\ & & \xi_m \end{pmatrix}^{-1},$$

where D *is any nonsingular diagonal matrix and*

$$C = \left(\frac{1}{p_i - q_j} \right)_{1 \leq i,j \leq m},$$

which appeared in Proposition 1.

Proof. Denote the proper polynomial of $P - \xi\eta$ by $f(x)$. We shall calculate $f(x)$ in two ways. Following the definition, we have

$$\begin{aligned} f(x) &= \det (x - (P - \xi\eta)) \\ &= \det \left[\begin{pmatrix} x - p_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \eta_1 \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{pmatrix}, \begin{pmatrix} 0 \\ x - p_2 \\ \vdots \\ 0 \end{pmatrix} + \eta_2 \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{pmatrix}, \right. \\ &\qquad \left. \cdots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ x - p_m \end{pmatrix} + \eta_m \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{pmatrix} \right] \\ &= \sum_{j=1}^{m} \prod_{\substack{1 \leq k \leq m \\ k \neq j}} (x - p_k)\, \eta_j \xi_j + \prod_{k=1}^{m} (x - p_k). \end{aligned}$$

On the other hand, since the eigenvalues of $P - \xi\eta$ are $q_1, \ldots, q_m$ by (2.1), we have

$$f(x) = \prod_{k=1}^{m} (x - q_k).$$

Comparing the above two expressions of $f(x)$, we obtain

$$\sum_{j=1}^{m} \frac{\eta_j \xi_j}{x - p_j} + 1 = \prod_{k=1}^{m} \frac{x - q_k}{x - p_k},$$

and the partial fraction expansion of the right-hand side proves (2.2). It is easy to see that

$$v_k = \begin{pmatrix} \frac{\xi_1}{p_1 - q_k} \\ \vdots \\ \frac{\xi_m}{p_m - q_k} \end{pmatrix}$$

is a q_k eigenvector of $P - \xi\eta$ for $k = 1, \ldots, m$. Therefore, we have

$$U^{-1} = \begin{pmatrix} \xi_1 & & \\ & \ddots & \\ & & \xi_m \end{pmatrix} \left(\frac{1}{p_i - q_j} \right)_{1 \leq i,j \leq m} D$$

with a nonsingular diagonal matrix D, which completes the proof.

3. Proofs of theorems. We prove Theorems I–IV* in §1.

Proof of Theorem I. By using G_Δ-valued transformation if necessary, we may assume that

$$A = \begin{pmatrix} \lambda_1 & & & a_1 \\ & \ddots & & \vdots \\ & & \lambda_{n-1} & a_{n-1} \\ b_1 & \cdots & b_{n-1} & \mu \end{pmatrix}.$$

The proper polynomial $f_A(x)$ of A is

$$\begin{aligned} f_A(x) &= \det [xI_n - A] \\ &= \varphi(x) \left[(x-\mu) - \sum_{j=1}^{n-1} \frac{a_j b_j}{x - \lambda_j} \right], \end{aligned}$$

where $\varphi(x) = \prod_{j=1}^{n-1} (x - \lambda_j)$. Since $\rho_1, \ldots, \rho_n$ are eigenvalues of A, and since we have assumed that $\rho_i \neq \lambda_j$ for any i, j, we have

$$f_A(\rho_i) = 0, \qquad \varphi(\rho_i) \neq 0$$

for $i = 1, \ldots, n$, and hence

$$\sum_{j=1}^{n-1} \frac{a_j b_j}{\rho_i - \lambda_j} = \rho_i - \mu, \qquad i = 1, \ldots, n.$$

Using Proposition 1, we can solve this system of linear equations in $a_1 b_1, \ldots, a_{n-1} b_{n-1}$ to obtain

$$\begin{aligned} a_i b_i &= \sum_{j=1}^{n-1} \frac{\prod_{k=1}^{n-1} (\rho_j - \lambda_k)}{\prod_{\substack{1 \le \ell \le n-1 \\ \ell \neq j}} (\rho_j - \rho_\ell)} \cdot \frac{\prod_{\substack{1 \le \ell \le n-1 \\ \ell \neq j}} (\lambda_i - \rho_\ell)}{\prod_{\substack{1 \le k \le n-1 \\ k \neq i}} (\lambda_i - \lambda_k)} \cdot (\rho_j - \mu) \\ &= -\frac{\prod_{\ell=1}^{n} (\lambda_i - \rho_\ell)}{\prod_{\substack{1 \le k \le n-1 \\ k \neq i}} (\lambda_i - \lambda_k)}, \qquad i = 1, \ldots, n-1, \end{aligned}$$

where we have used Lemmas 1 and 2 and ($\mathrm{F_I}$) to show the last equality. By assumption we see that $a_i \neq 0$ and $b_i \neq 0$ for every i. Then if we set

$$P = \begin{pmatrix} a_1 & & & \\ & \ddots & & \\ & & a_{n-1} & \\ & & & 1 \end{pmatrix},$$

$P \in G_\Delta$ and $P^{-1} A P = A_{\mathrm{I}}$, which proves the theorem.

For later use we note that, if $U \in \mathrm{GL}(n, \mathbf{C})$ satisfies

$$U^{-1} A_{\mathrm{I}} U = \begin{pmatrix} \rho_1 & & \\ & \ddots & \\ & & \rho_n \end{pmatrix},$$

then we have

$$U = \begin{pmatrix} \frac{1}{\rho_1-\lambda_1} & \cdots & \frac{1}{\rho_n-\lambda_1} \\ \vdots & & \vdots \\ \frac{1}{\rho_1-\lambda_{n-1}} & \cdots & \frac{1}{\rho_n-\lambda_{n-1}} \\ 1 & \cdots & 1 \end{pmatrix} \cdot D,$$

where D is a nonsingular diagonal matrix.

Proof of Theorem I*. Let k be an integer between 2 and n. Let $A^{(k)}$ be a $k \times k$ matrix such that

$$A^{(k)} = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ * & & \lambda_k \end{pmatrix} \sim \begin{pmatrix} \rho_1 I_{k-1} & \\ & \rho_2^{(k)} \end{pmatrix},$$

where $\rho_2^{(k)} = \sum_{i=1}^{k} \lambda_i - (k-1)\rho_1$. Let $G^{(k)}$ be the group of $k \times k$ nonsingular diagonal matrices. We prove that

$(3.1)_k$ there is a $P^{(k)} \in G^{(k)}$such that

$$P^{(k)^{-1}} A^{(k)} P^{(k)} = \begin{pmatrix} \lambda_1 & \lambda_1-\rho_1 & \cdots & \lambda_1-\rho_1 \\ \lambda_2-\rho_1 & \lambda_2 & \cdots & \lambda_2-\rho_1 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_k-\rho_1 & \lambda_k-\rho_1 & \cdots & \lambda_k \end{pmatrix}$$

by induction on k; then the theorem follows when $k = n$. $(3.1)_2$ is easily shown. We assume $(3.1)_{k-1}$, and shall show $(3.1)_k$. Set

$$A^{(k)} = \begin{pmatrix} A_1^{(k)} & \xi \\ \eta & \lambda_k \end{pmatrix}$$

with $A_1^{(k)} \in M(k-1, k-1), \xi \in M(k-1, 1)$, and $\eta \in M(1, k-1)$. Since rank $(A_{(k)} - \rho_1 I_k) = 1$, and since $\lambda_k - \rho_1 \neq 0$, we have

$$A_1^{(k)} - \rho_1 I_{k-1} = (\lambda_k - \rho_1)^{-1} \xi\eta. \tag{3.2}$$

The (1,1) entry of $A_1^{(k)} - \rho_1 I_{k-1}$ is $\lambda_1 - \rho_1$, which differs from 0 by assumption (1.7), so that $A_1^{(k)} - \rho_1 I_{k-1} \neq O$. Then it follows from (3.2) that

$$\operatorname{rank}\left(A_1^{(k)} - \rho_1 I_{k-1}\right) = 1.$$

Hence ρ_1 is $(k-2)$-ple eigenvalue of $A_1^{(k)}$, and the other eigenvalue is

$$\sum_{i=1}^{k-1} \lambda_i - (k-2)\rho_1 = \rho_2^{(k-1)}$$

by virtue of the invariance of the trace. If, for any $(k-1)$ elements $i_1, \ldots, i_{k-1}$ in $\{1, \ldots, n\}$,

$$\sum_{j=1}^{k-1} \lambda_{i_j} = (k-1)\rho_1 \tag{3.3}$$

would hold; we should have $\lambda_1 = \lambda_2 = \cdots = \lambda_n$, and hence

$$\lambda_i = \rho_1$$

should hold for every i by virtue of (3.3). This contradicts the assumption, and then

$$\sum_{j=1}^{k-1} \lambda_{i_j} \neq (k-1)\rho_1$$

for some $i_1, \ldots, i_{k-1}$. Thus, by an appropriate exchange of rows and columns, we may assume that $\sum_{i=1}^{k-1} \lambda_i \neq (k-1)\rho_1$, and hence

$$\rho_2^{(k-1)} \neq \rho_1.$$

Then we have

$$A_1^{(k)} \sim \rho_1 I_{k-2} \oplus \rho_2^{(k-1)}.$$

Now we can apply the induction assumption $(3.1)_{k-1}$, and by the help of (3.2) we see that $(3.1)_k$ holds. This completes the induction and hence the proof of the theorem.

Proof of Theorem II. We may assume that

$$A = \begin{pmatrix} L & A_1 \\ A_2 & M \end{pmatrix},$$

where

$$L = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix}, \qquad M = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_m \end{pmatrix},$$

and $A_1, A_2 \in M(m,m)$. Since $A \sim \rho_1 I_m \oplus \rho_2 I_{m-1} \oplus \rho_3$, we have

$$\operatorname{rank}(A - \rho_1 I_n) = \operatorname{rank} \begin{pmatrix} L - \rho_1 I_m & A_1 \\ A_2 & M - \rho_1 I_m \end{pmatrix} = m, \tag{3.4}$$

$$\operatorname{rank}(A - \rho_1 I_n)(A - \rho_2 I_n) = 1. \tag{3.5}$$

By the assumption $\lambda_i \neq \rho_1$ $(i = 1, \ldots, m)$, $L - \rho_1 I_m$ is nonsingular, and hence from (3.4) it follows that

$$M - \rho_1 I_m = A_2 (L - \rho_1 I_m)^{-1} A_1. \tag{3.6}$$

By the assumption $\mu_i \neq \rho_1$ $(i = 1, \ldots, m)$, $M - \rho_1 I_m$ is nonsingular, so that A_1 and A_2 are nonsingular. Using (3.6), we have

$$\begin{aligned}(A - \rho_1 I_n)(A - \rho_2 I_n) = &\begin{pmatrix} I_m & \\ & A_2 (L - \rho_1 I_m)^{-1} \end{pmatrix} \\ &\times \left[\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes \{(L - \rho_1 I_m)(L - \rho_2 I_m) + A_1 A_2\} \right] \\ &\times \begin{pmatrix} I_m & \\ & (L - \rho_1 I_m)^{-1} A_1 \end{pmatrix}.\end{aligned}$$

Set

$$U = (L - \rho_1 I_m)^{-1} A_1.$$

Then (3.5) holds if and only if

$$\operatorname{rank}[U(M - \rho_1 I_m) U^{-1} + (L - \rho_2 I_m)] = 1. \tag{3.7}$$

We can apply Proposition 3 to (3.7) to see that there are nonsingular diagonal matrices D_1, D_2 such that

$$U = D_1 C D_2,$$

where

$$C = \left(\frac{1}{(\rho_2 - \lambda_i) - (\mu_j - \rho_1)} \right)_{i,j}.$$

Define a nonsingular diagonal matrix P by

$$P = \begin{pmatrix} D_1 & \\ & D_2^{-1} \end{pmatrix} \begin{pmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{pmatrix},$$

where

$$f_i = -\prod_{\substack{1 \le k \le m \\ k \ne i}} (\lambda_i - \lambda_k), \quad f_{m+i} = \prod_{k=1}^{m} (\lambda_k + \mu_j - \rho_1 - \rho_2), \quad i = 1, \dots, m.$$

Then $P \in G_\Delta$, and we see that

$$P^{-1} A P = A_{\mathrm{II}},$$

where we have used Proposition 1 to calculate the entries of C^{-1}. This completes the proof.

Proof of Theorem II*. We may assume that

$$A = \begin{pmatrix} L & A_1 & \\ A_2 & M & a \\ & b & \nu \end{pmatrix},$$

where $A_1, A_2 \in M(m, m), a \in M(m-1, 1), b \in M(1, m-1)$,

$$L = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix}, \qquad M = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_{m-1} \end{pmatrix}.$$

Since $A \sim \rho_1 I_m \oplus \rho_2 I_m$, we obtain that

$$\operatorname{rank}(A - \rho_1 I_n) = m, \tag{3.8}$$

$$(A - \rho_1 I_n)(A - \rho_2 I_n) = O. \tag{3.9}$$

Noting the assumption $\lambda_i \neq \rho_1$ for $i = 1, \ldots, m$, we obtain from (3.8) that

$$\begin{pmatrix} M - \rho_1 I_{m-1} & a \\ b & \nu - \rho_1 \end{pmatrix} = A_2 (L - \rho_1 I_m)^{-1} A_1. \tag{3.10}$$

Then it follows from (3.9) and (3.10) that

$$(L - \rho_1 I_m)(L - \rho_2 I_m) + A_1 A_2 = O. \tag{3.11}$$

By the assumption (1.9) we see that A_1 and A_2 are nonsingular. We define $U \in \mathrm{GL}(m, \mathbf{C})$ by

$$A_2 = -U (L - \rho_2 I_m).$$

Then we obtain from (3.10) and (3.11)

$$U^{-1} \begin{pmatrix} M - \rho_1 I_{m-1} & a \\ b & \nu - \rho_1 \end{pmatrix} U = \rho_2 I_m - L.$$

In the proof of Theorem I we have determined unknowns a, b, and U of equations of this form. Applying the result, we obtain

$$a_i b_i = -\frac{\prod_{\ell=1}^{m} (\lambda_\ell + \mu_i - \rho_1 - \rho_2)}{\prod_{\substack{1 \le k \le m-1 \\ k \neq i}} (\mu_i - \mu_k)}, \qquad i = 1, \ldots, m-1,$$

$$U = HVD,$$

where we have set $a = {}^t(a_1, \ldots, a_{m-1}), b = (b_1, \ldots, b_{m-1})$ and

$$H = \begin{pmatrix} a_1 & & & \\ & \ddots & & \\ & & a_{m-1} & \\ & & & 1 \end{pmatrix}, \qquad V = \begin{pmatrix} \frac{1}{\rho_1+\rho_2-\lambda_1-\mu_1} & \cdots & \frac{1}{\rho_1+\rho_2-\lambda_m-\mu_1} \\ \vdots & & \vdots \\ \frac{1}{\rho_1+\rho_2-\lambda_1-\mu_{m-1}} & \cdots & \frac{1}{\rho_1+\rho_2-\lambda_m-\mu_{m-1}} \\ 1 & \cdots & 1 \end{pmatrix},$$

and D is a nonsingular diagonal matrix. Define a nonsingular diagonal matrix P by

$$P = \begin{pmatrix} D^{-1} & \\ & H \end{pmatrix} \begin{pmatrix} f_1 & & & \\ & \ddots & & \\ & & f_{n-1} & \\ & & & 1 \end{pmatrix},$$

where

$$f_i = \prod_{k=1}^{m-1} (\lambda_i + \mu_k - \rho_1 - \rho_2), \qquad i = 1, \ldots, m,$$

$$f_{m+j} = \prod_{\substack{1 \le k \le m-1 \\ k \neq j}} (\mu_j - \mu_k), \qquad j = 1, \ldots, m-1.$$

Then we see that

$$P^{-1} A P = A_{\mathrm{II}^*},$$

where we have used Proposition 2 to calculate the entries of V^{-1}. Since $P \in G_\Delta$, this completes the proof.

Proof of Theorem III. We may assume that

$$A = \begin{pmatrix} L & A_1 \\ A_2 & M \end{pmatrix},$$

where

$$L = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{m+1} \end{pmatrix}, \qquad M = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_m \end{pmatrix},$$

and $A_1 \in M(m+1, m), A_2 \in M(m, m+1)$. Since $A \sim \rho_1 I_m \oplus \rho_2 I_m \oplus \rho_3$, we have

$$\operatorname{rank}(A - \rho_1 I_n) = \operatorname{rank}(A - \rho_2 I_n) = m + 1.$$

Then, noting the assumption $\lambda_i \neq \rho_j$ for $i = 1, \ldots, m+1$ and $j = 1, 2$, we obtain

$$A_2 (L - \rho_1 I_{m+1})^{-1} A_1 = M - \rho_1 I_m, \tag{3.12}$$

$$A_2 (L - \rho_2 I_{m+1})^{-1} A_1 = M - \rho_2 I_m. \tag{3.13}$$

Set

$$A_1 = \begin{pmatrix} B_1 \\ b \end{pmatrix}, \qquad A_2 = (B_2 \quad a)$$

with $B_1, B_2 \in M(m, m), a = {}^t(a_1 \cdots a_m) \in M(m, 1)$ and $b = (b_1 \cdots b_m) \in M(1, m)$. Then from (3.12) and (3.13) it follows that

$$B_2 (L' - \rho_1 I_m)^{-1} B_1 + (\lambda_{m+1} - \rho_1)^{-1} ab = M - \rho_1 I_m, \tag{3.14}$$

$$B_2 (L' - \rho_2 I_m)^{-1} B_1 + (\lambda_{m+1} - \rho_2)^{-1} ab = M - \rho_2 I_m, \tag{3.15}$$

where we have set

$$L' = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix}.$$

Eliminating the terms containing ab from (3.14) and (3.15), we see that $\det B_1 \neq 0$ and obtain

(3.16)

$$B_2 = (M + (\lambda_{m+1} - \rho_1 - \rho_2) I_m) B_1^{-1} (L' - \rho_1 I_m) (L' - \lambda_{m+1} I_m)^{-1} (L' - \rho_2 I_m).$$

Put (3.16) into (3.14), then we have

$$B_1^{-1} Q_1 B_1 + E = Q_2, \tag{3.17}$$

where

$$Q_1 = (L' - \lambda_{m+1} I_m)^{-1} (L' - \rho_2 I_m),$$

$$Q_2 = (M + (\lambda_{m+1} - \rho_1 - \rho_2) I_m)^{-1} (M - \rho_1 I_m),$$

$$E = (\lambda_{m+1} - \rho_1)^{-1} (M + (\lambda_{m+1} - \rho_1 - \rho_2) I_m)^{-1} ab.$$

Applying Proposition 3 to (3.17) and using Proposition 1, we can calculate $a_i b_i$ $(i = 1, \dots, m)$ and the entries of B_1 and B_1^{-1}. In particular we have

$$
\begin{aligned}
B_1 &= D \cdot (u_{ij})_{i,j}\,, \\
u_{ij} &= (\lambda_{m+1} - \rho_1)(\rho_2 - \lambda_{m+1}) \\
&\quad \times \frac{\prod_{k=1}^{m} (\lambda_k + \mu_j - \rho_1 - \rho_2) \prod_{\substack{1 \le \ell \le m \\ \ell \ne j}} (\lambda_i + \mu_\ell - \rho_1 - \rho_2)}{\prod_{\substack{1 \le \ell \le m \\ \ell \ne i}} (\mu_j - \mu_\ell) \prod_{\substack{1 \le k \le m+1 \\ k \ne i}} (\lambda_i - \lambda_k)} \cdot \frac{1}{a_j}, \\
&\qquad i, j = 1, \dots, m,
\end{aligned}
$$

where D is a nonsingular diagonal matrix of size m. Define a nonsingular diagonal matrix P by

$$
P = \begin{pmatrix} D & \\ & I_{m+1} \end{pmatrix} \begin{pmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{pmatrix},
$$

where

$$
\begin{aligned}
f_i &= \frac{\lambda_i - \rho_1}{(\lambda_{m+1} - \rho_1) \prod_{\ell=1}^{m} (\lambda_i + \mu_\ell - \rho_1 - \rho_2)}, \qquad i = 1, \dots, m, \\
f_m &= -\prod_{\ell=1}^{m} (\lambda_{m+1} + \mu_\ell - \rho_1 - \rho_2)\,, \\
f_{m+1+i} &= \frac{(\lambda_{m+1} + \mu_i - \rho_1 - \rho_2) \prod_{\substack{1 \le \ell \le m \\ \ell \ne i}} (\mu_i - \mu_\ell)}{\rho_2 - \lambda_{m+1}}, \qquad i = 1, \dots, m.
\end{aligned}
$$

Then, $P \in G_\Delta$, and we see that

$$
P^{-1} A P = A_{\mathrm{III}},
$$

which completes the proof.

Proof of Theorem III*. We may assume that

$$
A = \begin{pmatrix} L & a & A_1 \\ b & \nu & c \\ A_2 & d & M \end{pmatrix},
$$

where

$$
L = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix}, \qquad M = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_m \end{pmatrix},
$$

and $A_1, A_2 \in M(m, m); a, d \in M(m, 1); b, c, \in M(1, m)$. Since $A \sim \rho_1 I_{m+1} \oplus \rho_2 I_m$, we have

$$
\operatorname{rank}(A - \rho_1 I_n) = m, \tag{3.18}
$$

$$
(A - \rho_1 I_n)(A - \rho_2 I_n) = 0. \tag{3.19}
$$

Noting the assumption $\lambda_i \neq \rho_1$ for $i = 1, \ldots, m$, from (3.18) we obtain

$$
(3.20) \qquad \begin{cases} \nu - \rho_1 = b\,(L - \rho_1 I_m)^{-1}\, a, \\ c = b\,(L - \rho_1 I_m)^{-1}\, A_1, \\ d = A_2\,(L - \rho_1 I_m)^{-1}\, a, \\ M - \rho_1 I_m = A_2\,(L - \rho_1 I_m)^{-1}\, A_1. \end{cases}
$$

Using these relations (3.20), we have

$$
\begin{aligned}
&(A - \rho_1 I_n)\,(A - \rho_2 I_n) \\
&\quad = \begin{pmatrix} I_m & & \\ & b\,(L - \rho_1 I_m)^{-1} & \\ & & A_2\,(L - \rho_1 I_m)^{-1} \end{pmatrix} \\
&\qquad \times \left[\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \otimes \{(L - \rho_1 I_m)\,(L - \rho_2 I_m) + ab + A_1 A_2\} \right] \\
&\qquad \times \begin{pmatrix} I_m & & \\ & (L - \rho_1 I_m)^{-1}\, a & \\ & & (L - \rho_1 I_m)^{-1}\, A_1 \end{pmatrix}.
\end{aligned}
$$

Then (3.19) holds if and only if

$$
(3.21) \qquad (L - \rho_1 I_m)\,(L - \rho_2 I_m) + ab + A_1 A_2 = O.
$$

Noting the assumption $\mu_i \neq \rho_1$ for $i = 1, \ldots, m$, from the fourth relation of (3.20) we obtain $\det A_1 \neq 0$, and hence

$$
(3.22) \qquad A_2 = (M - \rho_1 I_m)\, A_1^{-1}\, (L - \rho_1 I_m).
$$

Putting (3.22) into (3.21), we have

$$
(3.23) \qquad A_1\,(M - \rho_1 I_m)\, A_1^{-1} + ab\,(L - \rho_1 I_m)^{-1} = \rho_2 I_m - L.
$$

Applying Proposition 3 to (3.23) and using Proposition 1, we can calculate $a_j b_j$ $(j = 1, \ldots, m)$ and the entries of A_1 and A_1^{-1}, where we have set $a = {}^t(a_1, \ldots, a_m)$ and $b = (b_1, \ldots, b_m)$. In particular we have

$$
\begin{aligned}
&A_1 = (u_{ij})_{i,j} \cdot D^{-1}, \\
&u_{ij} = \frac{a_i}{\rho_1 + \rho_2 - \lambda_i - \mu_j}, \qquad i, j = 1, \ldots, m,
\end{aligned}
$$

where D is a nonsingular diagonal matrix of size m. Then using the second and the third relations of (3.20) and Lemma 1, we can calculate the entries of c and d. Define a nonsingular diagonal matrix P by

$$
P = \begin{pmatrix} I_{m+1} & \\ & D \end{pmatrix} \begin{pmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{pmatrix}
$$

where

$$f_i = \frac{\prod_{\substack{1\le k\le m\\ k\neq i}} (\lambda_k - \lambda_i)}{\lambda_i - \rho_1} \cdot a_i, \qquad i = 1, \dots, m,$$

$$f_m = 1,$$

$$f_{m+1+i} = -\prod_{\ell=1}^{m} (\lambda_\ell + \mu_i - \rho_1 - \rho_2), \qquad i = 1, \dots, m.$$

Then $P \in G_\Delta$, and we see that

$$P^{-1}AP = A_{\mathrm{III}^*},$$

which completes the proof.

Proof of Theorem IV. We may assume that

$$A = \begin{pmatrix} L & A_1 \\ A_2 & M \end{pmatrix},$$

where

$$L = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \lambda_3 & \\ & & & \lambda_4 \end{pmatrix}, \qquad M = \begin{pmatrix} \mu_1 & \\ & \mu_2 \end{pmatrix},$$

and $A_1 \in M(4,2), A_2 \in M(2,4)$. We set

$$A_1 = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \end{pmatrix}, \qquad A_2 = \begin{pmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \end{pmatrix}.$$

Since $A \sim \rho_1 I_2 \oplus \rho_2 I_2 \oplus \rho_3 I_2$, we have

$$\operatorname{rank}(A - \rho_i I_6) = 4 \tag{3.24}$$

for $i = 1, 2, 3$. Noting the assumption $\lambda_j \neq \rho_i$ for $j = 1, 2, 3, 4; i = 1, 2$, from (3.24) we obtain

$$A_2 (L - \rho_i I_4)^{-1} A_1 = M - \rho_i I_2 \tag{3.25}$$

for $i = 1, 2, 3$. By using the assumption (1.12) we see that none of u_{ij} and v_{ij} is 0. By solving (3.25), we can write the entries of A_1 and A_2 in terms of $u_{22}, u_{32}, u_{42}, v_{11}$ and v_{21}. Define a nonsingular diagonal matrix P by

$$P = \begin{pmatrix} 1 & & & & & \\ & f_2 & & & & \\ & & f_3 & & & \\ & & & f_4 & & \\ & & & & v_{11} & \\ & & & & & v_{21} \end{pmatrix},$$

where

$$f_2 = \frac{(\mu_2-\mu_1)\prod_{k=1,3,4}(\lambda_2-\lambda_k)}{\prod_{\ell=1}^{3}(\lambda_2-\rho_\ell)} \cdot \frac{u_{22}v_{21}}{[12;1]},$$
$$f_3 = \frac{(\mu_2-\mu_1)\prod_{k=1,2,4}(\lambda_3-\lambda_k)}{\prod_{\ell=1}^{3}(\lambda_3-\rho_\ell)} \cdot \frac{u_{32}v_{21}}{[13;1]},$$
$$f_4 = \frac{(\mu_2-\mu_1)\prod_{k=1,2,3}(\lambda_4-\lambda_k)}{\prod_{\ell=1}^{3}(\lambda_4-\rho_\ell)} \cdot \frac{u_{42}v_{21}}{[14;1]}.$$

Here we have set

$$[ij;k] = \lambda_i + \lambda_j + \mu_k - \rho_1 - \rho_2 - \rho_3$$

for $i,j = 1,2,3,4; k = 1,2$. Then $P \in G_\Delta$, and we see that

$$P^{-1}AP = A_{\mathrm{IV}},$$

which completes the proof.

Proof of Theorem IV*. We may assume that

$$A = \begin{pmatrix} L & A_1 & A_2 \\ A_3 & M & A_4 \\ A_5 & A_6 & N \end{pmatrix},$$

where

$$L = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}, \quad M = \begin{pmatrix} \mu_1 & \\ & \mu_2 \end{pmatrix}, \quad N = \begin{pmatrix} \nu_1 & \\ & \nu_2 \end{pmatrix},$$

and $A_1, A_2, A_3, A_4, A_5, A_6 \in M(2,2)$. Since $A \sim \rho_1 I_4 \oplus \rho_2 I_2$, we obtain

$$\operatorname{rank}(A - \rho_1 I_6) = 2, \tag{3.26}$$
$$(A-\rho_1 I_6)(A-\rho_2 I_6) = O. \tag{3.27}$$

Noting the assumption $\lambda_i \neq \rho_j$ for $i,j \in \{1,2\}$, from (3.26) we obtain

$$\begin{cases} M - \rho_1 I_2 = A_3 (L-\rho_1 I_2)^{-1} A_1, \\ A_6 = A_5 (L-\rho_1 I_2)^{-1} A_1, \\ A_4 = A_3 (L-\rho_1 I_2)^{-1} A_2, \\ N - \rho_1 I_2 = A_5 (L-\rho_1 I_2)^{-1} A_2. \end{cases} \tag{3.28}$$

Again noting the assumptions $\lambda_i \neq \rho_j, \mu_i \neq \rho_j$ and $\nu_i \neq \rho_j$ for $i,j \in \{1,2\}$, we see that every A_i is nonsingular for $i = 1,\dots,6$. Then we solve (3.28) to obtain

$$\begin{cases} A_3 = (M-\rho_1 I_2) A_1^{-1} (L-\rho_1 I_2), \\ A_4 = (M-\rho_1 I_2) A_1^{-1} A_2, \\ A_5 = (N-\rho_1 I_2) A_2^{-1} (L-\rho_1 I_2), \\ A_6 = (N-\rho_1 I_2) A_2^{-1} A_1. \end{cases} \tag{3.29}$$

By the help of (3.29) we have

$$
\begin{aligned}
&(A-\rho_1 I_6)(A-\rho_2 I_6)\\
&\quad=\begin{pmatrix} I_2 & & \\ & (M-\rho_1 I_2)A_1^{-1} & \\ & & (N-\rho_1 I_2)A_2^{-1}\end{pmatrix}\\
&\qquad\times\left[\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}\otimes\left\{(L-\rho_2 I_2)+A_1(M-\rho_1 I_2)A_1^{-1}+A_2(N-\rho_1 I_2)A_2^{-1}\right\}\right]\\
&\qquad\times\begin{pmatrix} L-\rho_1 I_2 & & \\ & A_1 & \\ & & A_2\end{pmatrix}.
\end{aligned}
$$

Then (3.27) holds if and only if

$$(L-\rho_2 I_2)+A_1(M-\rho_1 I_2)A_1^{-1}+A_2(N-\rho_1 I_2)A_2^{-1}=O. \tag{3.30}$$

We set

$$A_1=\begin{pmatrix}u_1&u_2\\u_3&u_4\end{pmatrix},\qquad A_2^{-1}=\begin{pmatrix}w_1&w_2\\w_3&w_4\end{pmatrix},$$

and determine the u_i's and the w_i's by solving (3.30). First we note that, on the assumption (1.13), none of the u_i's and the w_i's are 0. Then we introduce nonzero parameters $\theta_1,\theta_2,\tau_1,\tau_2$ and τ_3 by

$$
\begin{aligned}
u_1&=\frac{\lambda_1-\rho_1}{\lambda_1-\lambda_2}\cdot\theta_1, \qquad u_2=\frac{\lambda_1-\rho_1}{\lambda_1-\lambda_2}\cdot\theta_2,\\
w_1&=\frac{1}{(\lambda_1-\rho_1)(\nu_1-\nu_2)}\cdot\tau_1, \qquad w_2=\frac{1}{(\lambda_2-\rho_1)(\nu_1-\nu_2)}\cdot\tau_2,\\
w_4&=\frac{1}{(\lambda_2-\rho_1)(\nu_2-\nu_1)}\cdot\tau_3.
\end{aligned}
$$

Now we can write the entries of A_1 and A_2^{-1}, and hence the entries of A_i $(i=1,\dots,6)$ by (3.29), in terms of $\theta_1,\theta_2,\tau_1,\tau_2$ and τ_3. Define a nonsingular diagonal matrix P by

$$
P=\begin{pmatrix}
\frac{[122]}{\tau_1} & & & & & & \\
& \frac{[121]}{\tau_2} & & & & & \\
& & \frac{[121][122]}{\theta_1\tau_1} & & & & \\
& & & \frac{[112][122]}{\theta_2\tau_1} & & & \\
& & & & 1 & & \\
& & & & & \frac{\tau_3[121]}{\tau_2[112]} &
\end{pmatrix},
$$

where $[ijk]$ is defined in (1.14). Then $P\in G_\Delta$, and we see that

$$P^{-1}AP=A_{\mathrm{IV}^*},$$

which completes the proof.

REFERENCES

[1] E. GOURSAT, *Extension du problème de Riemann à des fonctions hypergéométriques de deux variables,* C. R. Acad. Sci. Paris, 95 (1882), pp. 903–906.

[2] Y. HARAOKA, *Finite monodromy of Pochhammer equation,* Ann. Inst. Fourier, to appear.
[3] H. KIMURA AND K. OKAMOTO, *On the polynomial Hamiltonian structure of the Garnier systems,* J. Math. Pures Appl., 63 (1984), pp. 129–146.
[4] A. H. M. LEVELT, *Hypergeometric functions,* thesis, University of Amsterdam, The Netherlands, 1961.
[5] K. OKUBO, *On the group of Fuchsian equations,* Seminar Reports of Tokyo Metropolitan University, Tokyo, Japan 1987.
[6] K. OKUBO, K. TAKANO, AND S. YOSHIDA, *A connection problem for the generalized hypergeometric equation,* Funkcial. Ekvac., 31 (1988), pp. 483–495.
[7] T. SASAI, *On a monodromy group and irreducibility conditions of a fourth order Fuchsian differential system of Okubo type,* J. Reine Angew. Math., 299/300 (1978), pp. 38–50.
[8] T. SASAI AND S. TSUCHIYA, *On a fourth order Fuchsian differential equation of Okubo type,* Funkcial Ekvac., 34 (1991), pp. 211–221.
[9] T. YOKOYAMA, *On an irreducibility condition for hypergeometric systems,* Funk. Ekvac., to appear.
[10] T. SASAI AND S. TSUCHIYA, *On a class of even order Fuchsian equations of Okubo type*, Funk. Ekvac., 35 (1992), pp. 505–514.

SIAM J. MATH. ANAL.
Vol. 25, No. 4, pp. 1227–1239, July 1994

013

ON MONOTONE SPLINE APPROXIMATION*

X. M. YU† AND S. P. ZHOU‡

Abstract. The present paper shows that one cannot expect the Jackson type estimates to hold for higher degree moduli of smoothness in monotone (or comonotone) spline approximation. This result gives a complete negative answer to a question raised by DeVore for $m \geq 2$ in monotone spline approximation. It also indicates that an equivalence between monotone polynomial approximation and monotone spline approximation claimed by Wang is wrong in one direction. There are corresponding results in L^p spaces.

Key words. monotone approximation, comonotone approximation, spline, Jackson type estimates

AMS subject classifications. 41A15, 41A29

1. Introduction.

Denote by $C^N_{[0,1]}$ the class of real functions which have N continuous derivatives on the interval $[0,1]$, $C_{[0,1]} = L^\infty_{[0,1]} = C^0_{[0,1]}$, and by $C^\infty_{[0,1]}$ the class of real functions which are infinitely differentiable on $[0,1]$. Let $L^p_{[0,1]}$ be the space of real integrable functions of power p on $[0,1]$, Π_n the class of algebraic polynomials of degree at most n, and for $k \geq 1$,

$$\Delta^k = \{f : \Delta^k_h f(x) \geq 0, \quad x \in [0, 1-kh], \quad h > 0\},$$

where

$$\Delta^k_h f(x) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(x+jh).$$

More generally, let $\Delta^1(r)$ denote the class of functions such that $f(x+h) - f(x)$ changes its sign exactly r times on the interval $[0, 1-h]$ for sufficiently small $h > 0$. Then $\Delta^1(0) = \Delta^1$.

For $f \in C_{[a,b]}$, let

$$\|f\|_{[a,b]} = \max_{a \leq x \leq b} |f(x)|,$$

and

$$\|f\| := \|f\|_{[0,1]}.$$

For $f \in L^p_{[a,b]}$ and $1 \leq p < \infty$,

$$\|f\|_{L^p_{[a,b]}} = \left(\int_a^b |f(x)|^p dx \right)^{1/p},$$

* Received by the editors March 24, 1992; accepted for publication (in revised form) April 30, 1993.

† Department of Mathematics, Southwest Missouri State University, 901 South National Avenue, Springfield, Missouri 65804-0094.

‡ Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5.

$$\|f\|_{L^p} = \|f\|_{L^p_{[0,1]}}.$$

As usual,

$$\omega_m(f,t)_{L^p} = \sup\left\{\|\Delta_h^m f(x)\|_{L^p_{[0,1-mh]}} : 0 < h \le t\right\},$$

$$\omega_m(f,t) = \omega_m(f,t)_{L^\infty}.$$

Let $\mathcal{S}(m+1,n)$, $m \ge 0$, denote the space of all splines of order $m+1$ on the $n+1$ equally spaced knots $\{i/n\}_{i=0}^n$, that is, $s \in \mathcal{S}(m+1,n)$ if s is a polynomial of degree m in each interval $[i/n,(i+1)/n]$ and $s^{(m-1)}$ is continuous on $[0,1]$ (if $m=0$ s is a piecewise constant with no continuity at the knots).

If $f \in L^p_{[0,1]} \bigcap \Delta^1(r)$, denote

$$E_{n,m}(f,r)_{L^p} = \inf\{\|f-s\|_{L^p}\},$$

where the infimum is taken over all $s \in \mathcal{S}(m+1,n)$, which are comonotone with $f(x)$, that is, $(s(x+h)-s(x))(f(x+h)-f(x)) \ge 0$ for $x \in [0,1-h]$ and sufficiently small $h>0$. Write

$$E_{n,m}(f,r) := E_{n,m}(f,r)_{L^\infty},$$

$$E_{n,m}(f)_{L^p} := E_{n,m}(f,0)_{L^p},$$

$$E_{n,m}(f) := E_{n,m}(f,0).$$

Throughout the paper, we will use $C(x)$ to denote a positive constant depending only upon x and C an absolute positive constant, which may, in general, vary in different relations.

For many years, comonotone approximations of functions by algebraic polynomials and by splines are very active fields in approximation theory. The special concern in these fields are focused on Jackson type estimates by many scholars. In monotone approximation by splines, DeVore [3] established the Jackson type estimate for $f \in C^j_{[0,1]} \cap \Delta^1$ to be

$$E_{n,m}(f) \le C(m)n^{-j}\omega(f^{(j)},n^{-1}) \tag{1.1}$$

for $0 \le j \le m$, and remarked in the same paper that it is preferable to get the Jackson type estimates of the form that

$$E_{n,m}(f) \le C(m)\omega_{m+1}(f,n^{-1}) \tag{1.2}$$

for $f \in C_{[0,1]} \cap \Delta^1$.

This estimate is true for $m=0,1$. If $m=0$, it is a special case of the above result (1.1). For $m=1$, it was actually proved in Newman [7], and we will give some brief remark on it in §2 of this paper.

Leviatan and Mhaskar [5] improves the above results to

$$E_{n,m}(f) \le C(m)n^{-1}\omega_m(f',n^{-1}) \tag{1.3}$$

for $f \in C^1_{[0,1]} \cap \Delta^1$.

It was not far from the preferable result (1.2). However, some surprising phenomena occured.

Wang claimed in [9] that for $f \in C_{[0,1]} \cap \Delta^1$,

$$E_{n,m}(f) \le C(m)\omega_{m+1}(f, n^{-1}) \tag{1.4}$$

is equivalent to

$$E_n^{(1)}(f) \le C(m)\omega_m(f, n^{-1}), \tag{1.5}$$

where $E_n^{(1)}(f)$ is the best monotone approximation of f by π_n in the uniform norm.

By having (1.4) imply (1.5), Wang proved that for $m \ge 3$,

$$\sup_{n\ge 1}\left\{\frac{E_{n,m}(f)}{\omega_{m+1}(f, n^{-1})} : f \in C_{[0,1]} \cap \Delta^1\right\} = +\infty. \tag{1.6}$$

Although Wang's result means that the estimates (1.2) cannot hold for a constant $C(m)$ depending upon m only for $m \ge 3$, it still might be possible that for every given function $f \in C_{[0,1]} \cap \Delta^1$, there is a constant $C(f, m)$ that is independent of n but may depend on f such that for $m \ge 3$,

$$E_{n,m}(f) \le C(f, m)\omega_{m+1}(f, n^{-1}). \tag{1.7}$$

(This is a weak form of the question (1.2).) Furthermore, taking the fact that (1.2) holds for $m = 0, 1$ into account, we may ask what happens in the case $m = 2$. The present paper will investigate those problems. We will establish the following result.

THEOREM 1. *Let $m \ge 0$. Then there exists a function $f \in C^1_{[0,1]} \bigcap \Delta^1$ such that*

$$\limsup_{n\to\infty} \frac{E_{n,m}(f)}{\omega_3(f, n^{-1})} = +\infty.$$

Theorem 1 follows as a particular case from the following, a slightly more general result.

THEOREM 2. *Let $m \ge 0$. Then there exists a function $f \in C^1_{[0,1]} \bigcap \Delta^1$ such that*

$$\limsup_{n\to\infty} \frac{E^*_{n,m}(f)}{\omega_3(f, n^{-1})} = +\infty,$$

where

$$E^*_{n,m}(f) = \inf\{\|f - s\|\},$$

and the infimum runs over all $s \in \mathcal{S}(m+1, n)$, which satisfies $s'(0) \ge 0$.

Theorem 1 completes the research in monotone spline approximation. It gives a complete negative answer to DeVore's question (1.2) as well as to (1.7) for all $m \ge 2$ and thus shows that Leviatan and Mhaskar's result (1.3) is the best possible Jackson type estimates in monotone spline approximation. In addition, Theorem 1 also indicates that Wang's claim of the equivalence between (1.4) and (1.5) is wrong in one direction, that is, (1.5) cannot guarantee (1.4) (we will strengthen the other direction in §3), since Theorem 1 holds and we have already that

$$E_n^{(1)}(f) \le C\omega_2(f, n^{-1})$$

for all $f \in C_{[0,1]} \cap \Delta^1$ (see DeVore [2]).

There are some essential differences between monotone spline approximation in continuous space and that in L^p space for $1 \leq p < \infty$.

THEOREM CSW (Chui, Smith, and Ward [1]). *If $f \in \Delta^1$ is a j-fold integral of an $L^p_{[0,1]}$ function, then for a nonnegative integer $0 \leq j \leq m$, one has*

$$E_{n,m}(f)_{L^p} \leq C(m)n^{-j}\omega(f^{(j)}, n^{-1})_{L^p}$$

for $1 \leq p < \infty$.

THEOREM LM1 (Leviatan and Mhaskar [5]). *Let $1 \leq p < \infty$, $f(x) \in \Delta^1$ with absolutely continuous $f'(x)$ in $[0,1]$ and $f'' \in L^p_{[0,1]}$, then*

$$E_{n,m}(f)_{L^p} \leq C(m)n^{-2}\omega_{m-1}(f'', n^{-1})_{L^p}. \tag{1.8}$$

In comonotone approximation by splines, Leviatan and Mhaskar [6] proved the following.

THEOREM LM2 (Leviatan and Mhaskar [6]). *Let $m \geq 0$, $f(x) \in C^j_{[0,1]} \cap \Delta^1(r)$, $0 \leq j \leq m$. Then*

$$E_{n,m}(f, r) \leq C(m, r)n^{-j}\omega(f^{(j)}, n^{-1}).$$

It is a very natural question to ask whether the estimates (1.3) for continuous function space can also hold for L^p spaces for $1 \leq p < \infty$, in other words, whether or not the estimate (1.8) can be improved. Although in most cases L^p spaces for $1 \leq p < \infty$ behave similarly to the continuous function space, in §4 we show that the estimate (1.3) is no longer valid in L^p for $1 < p < \infty$. That also indicates that the estimate (1.8) given by Leviatan and Mhaskar is actually the best possible Jackson type estimate in L^p for $1 < p < \infty$.

We will establish some general results in comonotone approximation case.

THEOREM 3. *Let $r \geq 0$, $1 < p < \infty$, $m \geq 0$. Then there exists a function $f \in C^1_{[0,1]} \bigcap \Delta^1(r)$ such that*

$$\limsup_{n\to\infty} \frac{E^*_{n,m}(f)_{L^p}}{n^{-1}\omega_2(f', n^{-1})_{L^p}} = +\infty,$$

where

$$E^*_{n,m}(f)_{L^p} = \inf\{\|f - s\|_{L^p}\},$$

and the infimum runs over all $s \in \mathcal{S}(m+1, n)$, which satisfies $s'(0) \geq 0$.

COROLLARY 1. *Let $r \geq 0$, $1 < p < \infty$, $m \geq 0$. Then there exists a function $f \in C^1_{[0,1]} \bigcap \Delta^1$ such that*

$$\limsup_{n\to\infty} \frac{E_{n,m}(f, r)_{L^p}}{n^{-1}\omega_2(f', n^{-1})_{L^p}} = +\infty.$$

From Theorem 3, another direct corollary that is a corresponding result in L^p, $1 < p < \infty$, to Theorem 1 follows.

COROLLARY 2. *Let* $1 < p < \infty$, $m \geq 0$. *Then there exists a function* $f \in C^1_{[0,1]} \bigcap \Delta^1$ *such that*

$$\limsup_{n\to\infty} \frac{E_{n,m}(f)_{L^p}}{\omega_3(f, n^{-1})_{L^p}} = +\infty.$$

We still do not know what *exactly* happens in L^1 space; however, we have the next theorem.

THEOREM 4. *Let* $r \geq 0$ *and* $m \geq 0$. *Then there exists a function* $f \in C^1_{[0,1]} \cap \Delta^1(r)$ *such that*

$$\limsup_{n\to\infty} \frac{E^*_{n,m}(f)_{L^1}}{n^{-1}\omega_3(f', n^{-1})_{L^1}} = +\infty.$$

COROLLARY 3. *Let* $r \geq 0$ *and* $m \geq 0$. *Then there exists a function* $f \in C^1_{[0,1]} \cap \Delta^1(r)$ *such that*

$$\limsup_{n\to\infty} \frac{E_{n,m}(f, r)_{L^1}}{n^{-1}\omega_3(f', n^{-1})_{L^1}} = +\infty.$$

By a quite similar way, in continuous function space, Theorems 1 and 2 can be generalized to the following

THEOREM 5. *Let* $r \geq 0$, $m \geq 0$. *Then there exists a function* $f \in C^1_{[0,1]} \bigcap \Delta^1(r)$ *such that*

$$\limsup_{n\to\infty} \frac{E^*_{n,m}(f)}{\omega_3(f, n^{-1})} = +\infty.$$

COROLLARY 4. *Let* $r \geq 0$, $m \geq 0$. *Then there exists a function* $f \in C^1_{[0,1]} \bigcap \Delta^1(r)$ *such that*

$$\limsup_{n\to\infty} \frac{E_{n,m}(f, r)}{\omega_3(f, n^{-1})} = +\infty.$$

2. DeVore's question (1.2) for m = 1. Newman [7] actually proved that for a function $f \in C_{[0,1]}$ there is a polygonal function $P_f(x)$ which agrees with $f(x)$ at $\{i/n\}_{i=0}^n$ so that

$$\|f - P_f\| \leq C\omega_2(f, n^{-1}).$$

We note that, in addition, if $f \in \Delta^1$, then clearly $P_f \in \Delta^1$, too. This observation leads to the positive answer of DeVore's question (1.2) for $m = 1$.

3. The relation between $E_{n,m}(f)$. and $E_n^{(1)}(f)$. In this section, we are going to prove the following.

THEOREM 6. *Let* $f \in C_{[0,1]} \cap \Delta^1$, *and* $m > 1$ *be a positive integer. Then*

$$E_n^{(1)}(f) \leq C(m) \left\{ E_{n,m}(f) + \omega_{m+1}(f, n^{-1}) \right\}.$$

First we show the following property for $s \in \mathcal{S}(m+1,n)$.

LEMMA 1. *Let m be a positive integer and $s \in \mathcal{S}(m+1,n)$. Then we have*

$$\omega_m\left(s', n^{-1}\right) \le C(m) n \omega_{m+1}\left(s, n^{-1}\right).$$

Proof. Let $0 < h \le 1/(n(m+1)^2) =: h_0$. Since s' is a polynomial of degree $m-1$ in each interval $I_i := [i/n, (i+1)/n]$, if $x \in I_i$, and $x + mh \in I_i$, then $\Delta_h^m s'(x) = 0$. Now suppose that there is some $1 \le j \le m$ such that x and $x + (j-1)h \in I_i$, and $x + jh$ and $x + mh \in I_{i+1}$. Denote $s(x) = p_i(x)$ for $x \in I_i$. We have

$$\begin{aligned}
\Delta_h^m s'(x) &= \sum_{k=0}^{j-1} (-1)^{m-k} \binom{m}{k} p_i'(x+kh) + \sum_{k=j}^{m} (-1)^{m-k} \binom{m}{k} p_{i+1}'(x+kh) \\
&= \sum_{k=j}^{m} (-1)^{m-k} \binom{m}{k} \left[p_{i+1}'(x+kh) - p_i'(x+kh)\right] + \Delta_h^m p_i'(x) \\
&= \sum_{k=j}^{m} (-1)^{m-k} \binom{m}{k} \left[p_{i+1}'(x+kh) - p_i'(x+kh)\right].
\end{aligned}$$

Hence

$$\begin{aligned}
|\Delta_h^m s'(x)| &\le C(m) \max\left\{|p_{i+1}'(x) - p_i'(x)| : \ x \in [(i+1)/n, (i+1)/n + mh_0]\right\} \\
&\le C(m) n \max\left\{|p_{i+1}(x) - p_i(x)| : \ x \in [(i+1)/n, (i+1)/n + mh_0]\right\}.
\end{aligned}$$

On the other hand, noticing that for $x \in [(i+1)/n, (i+1)/n + mh_0]$, both $x - (m+1)mh_0$ and $x - mh_0 \in I_i$, and we have

$$\Delta_{mh_0}^{m+1} s\left(x - (m+1)mh_0\right) = p_{i+1}(x) - p_i(x).$$

From this it follows that

$$\begin{aligned}
|\Delta_h^m s'(x)| &\le C(m) n \max\left\{\left|\Delta_{mh_0}^{m+1} s\left(x - (m+1)mh_0\right)\right| : \ x \in [(i+1)/n, (i+1)/n + mh_0]\right\} \\
&\le C(m) n \max_{x \in I_i} |\Delta_{mh_0}^{m+1} s(x)| \le C(m) n \omega_{m+1}(s, m/(n(m+1)^2)).
\end{aligned}$$

Then

$$\begin{aligned}
\omega_m(s', n^{-1}) &\le C(m) \omega_m\left(s', 1/(n(m+1)^2)\right) \le C(m) n \omega_{m+1}\left(s, m/(n(m+1)^2)\right) \\
&\le C(m) n \omega_{m+1}(s, n^{-1}). \qquad \square
\end{aligned}$$

LEMMA 2 (see Shevchuk [8]). *Let m be a positive integer. Then, if $f \in C_{[0,1]}^1 \cap \Delta^1$, we have*

$$E_n^{(1)}(f) \le C(m) n^{-1} \omega_m(f', n^{-1}).$$

Proof of Theorem 6. Let $s \in S(m+1,n) \cap \Delta^1$ and $\|f - s\| = E_{n,m}(f)$. Since s' is continuous, from Lemmas 1 and 2, we have

$$E_n^{(1)}(s) \le C(m) n^{-1} \omega_m(s', n^{-1}) \le C(m) \omega_{m+1}(s, n^{-1}).$$

Hence

$$E_n^{(1)}(f) \le E_{n,m}(f) + E_n^{(1)}(s) \le C(m)\left\{E_{n,m}(f) + \omega_{m+1}(f, n^{-1})\right\}.$$

This completes the proof. □

COROLLARY 5. *If* $f \in C_{[0,1]} \cap \Delta^1$, *and*

$$E_{n,m}(f) \le C(m)\omega_{m+1}(f, n^{-1}),$$

then we have

$$E_n^{(1)}(f) \le C(m)\omega_{m+1}(f, n^{-1}).$$

4. Proof of Theorem 3. In this section, we always assume that $1 < p < \infty$.

LEMMA 3. *Let* $x_j^r = \frac{1}{2} + \frac{j}{2r}$, $j = 1, 2, \ldots, r$ *for* $r \ge 1$. *Define*

$$f_1(x, r) = (-1)^r x^2 \prod_{j=1}^{r} (x - x_j^r)$$

for $r \ge 1$, *and*

$$f_1(x, 0) = x^2.$$

Then $f_1'(x, r)$ *has exactly* r *single zeros* $0 < \beta_1 < \beta_2 < \cdots < \beta_r < 1$. *Furthermore, there is a positive constant* A *independent of* x *so that for* $x \in [0, \beta_1/2]$ (*in case* $r = 0$, *we put* $\beta_1/2 = 1$),

$$f_1'(x, r) \ge Ax.$$

Proof. The argument is quite straightforward. □

LEMMA 4. *Given* $\epsilon_n > 0$ *with* $\epsilon_n \to 0$ *as* $n \to \infty$. *Let* $n \ge 2$,

$$h_n = -\epsilon_n x.$$

Then there exist functions $f_n(x) \in C_{[0,1]}^{\infty}$ *such that*

$$f_n'(0) = 0,$$

$$\|f_n - h_n\|_{L^p} \sim \epsilon_n^{2+1/p},$$

$$\|f_n' - h_n'\|_{L^p} \le C\epsilon_n^{1+1/p}.$$

Furthermore, there is a positive constant B *independent of* n *and* x *such that*

$$|f_n'(x)| \le Bx.$$

Proof. Let $n \ge 2$,

$$\alpha_n(x) = \frac{x}{x - \epsilon_n},$$

$$g_n(x) = \epsilon_n x e^{\alpha_n(x)} + h_n(x)$$

for $x \in [0, \epsilon_n)$, and

$$f_n(x) = \begin{cases} g_n(x), & x \in [0, \epsilon_n), \\ h_n(x), & x \in [\epsilon_n, 1]. \end{cases}$$

We have for $j \geq 1$ and $x \in [0, \epsilon_n)$,

$$\left| \frac{d^j}{dx^j} e^{\alpha_n(x)} \right| = \epsilon_n^{-j} \left| \frac{d^j}{dy^j} \exp\left(\frac{y}{y-1} \right) \right| \leq C(j) \epsilon_n^{-j}. \tag{4.1}$$

We only need to prove the last inequality; the argument for the others is quite straightforward. For $x \in [0, \epsilon_n/2)$, by (4.1),

$$\begin{aligned} |g_n'(x)| &\leq \epsilon_n |e^{\alpha_n(x)} - 1| + \epsilon_n x \left| \frac{d}{dx} e^{\alpha_n(x)} \right| \\ &\leq C\epsilon_n \frac{x}{|x - \epsilon_n|} + Cx \leq Cx, \end{aligned}$$

similarly for $x \in [\epsilon_n/2, \epsilon_n)$,

$$\frac{d}{dx} \epsilon_n x e^{\alpha_n(x)} \leq 2\epsilon_n + Cx \leq Cx,$$

and for $x \in [\epsilon_n/2, 1]$,

$$|h_n'(x)| = \epsilon_n \leq 2x.$$

All above estimates imply that there is a positive constant independent of n and x such that

$$|f_n'(x)| \leq Bx. \qquad \square$$

LEMMA 5. *Let $\theta = 1 - 1/p > 0$, $\epsilon_n = n^{-1-\theta/2}$. Define*

$$F_l(x) := \sum_{j=1}^{l} n_j^{-\theta/8} f_{n_j}(x),$$

$$Q_l(x) := q_l(x) + n_l^{-\theta/8} h_{n_l}(x),$$

where $h_n(x)$ and $f_n(x)$, $n \geq 2$ are the functions we defined in Lemma 4, $f_1(x) = f_1(x, r)$ appearing in Lemma 3, $q_l(x)$ is the algebraic polynomial of best approximation in the uniform norm of degree $m_l = [n_l^{\theta/(16p)}] + 1$ of $F_{l-1}(x)$, and $\{n_l\}$ is a sequence of natural numbers chosen by induction: Set $n_1 = 1$,

$$n_{l+1} \geq \|F_l^{(k)}\| \tag{4.2}$$

for $l = 1, 2, \ldots$, where $[x]$ is the greatest integer not exceeding x, $k = \left[\frac{96p}{\theta}\right] + 3$. Then the following estimates hold:

$$\|F_l - Q_l\|_{L^p} \sim n_l^{-\theta/8} \|f_{n_l} - h_{n_l}\|_{L^p} \sim n_l^{-\theta/8} \epsilon_{n_l}^{2+1/p}, \tag{4.3}$$

$$Q_l'(0) \le -Cn_l^{-\theta/8}\epsilon_{n_l}. \tag{4.4}$$

Proof. Applying Lemma 4, we get

$$\|F_l - Q_l\|_{L^p} = \|n_l^{-\theta/8}(f_{n_l} - h_{n_l}) + (F_{l-1} - q_l)\|_{L^p}. \tag{4.5}$$

Since, by (4.2),

$$\|F_{l-1} - q_l\| \le C\|F_{l-1}^{(k)}\|m_l^{-k} \le Cn_l^{-5},$$

and

$$n_l^{-\theta/8}\epsilon_{n_l}^{2+1/p} \ge n_l^{-3-13\theta/8},$$

hence from Lemma 4,

$$\|F_{l-1} - q_l\| \le Cn_l^{-\theta/8}\|f_{n_l} - h_{n_l}\|_{L^p}.$$

Together with (4.5), (4.3) follows. At the same time,

$$Q_l'(0) = q_l'(0) - n_l^{-\theta/8}\epsilon_{n_l}.$$

Applying a theorem on simultaneous approximation to continuous functions and their derivatives by Leviatan [4], together with Lemma 4 and (4.2), we have

$$|q_l'(0)| = |q_l'(0) - F_{l-1}'(0)| \le C\|F_{l-1}^{(k)}\|m_l^{-k+2} \le C\|F_{l-1}^{(k)}\|n_l^{-6} \le Cn_l^{-5},$$

therefore,

$$Q_l'(0) \le -n_l^{-\theta/8}\epsilon_{n_l} + Cn_l^{-5} \le -Cn_l^{-\theta/8}\epsilon_{n_l},$$

that is, (4.4) holds. □

LEMMA 6. *Let $0 < \beta_1 < \beta_2 < \cdots < \beta_r < 1$ be the r single zeros of $f_1'(x)$ in Lemma 3, write $\beta_{r+1} = 1$, and denote by γ_j the maximum point of $|f_1'(x)|$ within (β_j, β_{j+1}), $j = 1, 2, \ldots, r$. Set*

$$\delta = \min_{1\le j\le r}\left\{\frac{\beta_j}{2}, |f_1'(\gamma_j)|\right\}.$$

In addition to all conditions of Lemma 5, assume further that

$$n_{l+1} \ge \left[2^{8/\theta}\left(\max\{1, (B/A)^{8/\theta}\}n_l + m^{16p/\theta} + \delta^{-1} + 2^{8/\theta}\right)\right] + 1, \tag{4.6}$$

where A and B are positive constants appearing in Lemmas 3 and 4. Then

$$f(x) = \sum_{j=1}^{\infty} n_j^{-\theta/8} f_{n_j}(x) \in C_{[0,1]}^1 \cap \Delta^1(r),$$

and $f'(x) > 0$ for $x \in (0, \beta_1/2]$.

Proof. First of all, $f \in C_{[0,1]}^1$ is a clear fact by (4.6). From Lemmas 3 and 4, for $x \in [0, \beta_1/2]$,

$$f_1'(x) \ge Ax,$$

$$|f'_{n_l}(x)| \le Bx$$

for $l \ge 2$. Then by (4.6), for $x \in [0, \beta_1/2]$,

$$\begin{aligned} f'(x) &\ge f'_1(x) - \sum_{j=2}^{\infty} n_j^{-\theta/8} |f'_{n_j}(x)| \ge Ax - Bx \sum_{j=2}^{\infty} n_j^{-\theta/8} \\ &\ge Ax - \frac{A}{2} x \sum_{j=1}^{\infty} 2^{-j} = \frac{A}{2} x, \end{aligned}$$

that is, $f'(x) > 0$ for $x \in (0, \beta_1/2]$. It is also clear that

$$\operatorname{sgn}(f'_1(\gamma_j)) = (-1)^j, \qquad j = 1, 2, \ldots, r.$$

Due to (4.6), for all $l \ge 2$,

$$\epsilon_{n_l} \le n_l^{-1} \le \delta \le \frac{\beta_1}{2},$$

so for all $m \in \{n_l\}_{l=2}^{\infty}$ and $x \in [\beta_1/2, 1]$,

$$f_m(x) = -\epsilon_m x, \tag{4.7}$$

or

$$|f'_m(x)| = \epsilon_m < \delta.$$

Consequently, from (4.6) again,

$$\left| \sum_{j=2}^{\infty} n_j^{-\theta/8} f'_{n_j}(x) \right| \le \delta \sum_{j=2}^{\infty} n_j^{-\theta/8} \le \frac{\delta}{2}.$$

In view of that $|f'_1(\gamma_j)| \ge \delta$ for every $j = 1, 2, \ldots, r$,

$$\operatorname{sgn}(f'(\gamma_j)) = (-1)^j, \qquad j = 1, 2, \ldots, r,$$

which implies that $f'(x)$ has $r - 1$ zeros in $(\gamma_1, 1)$. We also note that $f'(\beta_1/2) > 0$, that is, $f'(x)$ has one zero in $(\beta_1/2, \gamma_1)$. Write

$$g(x) = f_1(x) - \sum_{j=2}^{\infty} n_j^{-\theta/8} \epsilon_{n_j} x,$$

then

$$g'(x) = f'_1(x) - \sum_{j=2}^{\infty} n_j^{-\theta/8} \epsilon_{n_j},$$

$$g'(0) = -\sum_{j=2}^{\infty} n_j^{-\theta/8} \epsilon_{n_j} < 0.$$

By considering (4.7), we see that $f(x) = g(x)$ for $x \in [\beta_1/2, 1]$. Now that $g'(0) < 0$ and $g'(\beta_1/2) = f'(\beta_1/2) > 0$ indicate that $g'(x)$ has one zero in $(0, \beta_1/2)$. Altogether,

$g'(x)$ has $r+1$ zeros in $[0,1]$. Since $g(x)$ is a polynomial of degree $r+2$, all $r+1$ zeros of $g'(x)$ must be single. Therefore, $f'(x)$ has exactly r single zeros in (0,1), and we have proved that $f \in \Delta^1(r)$. □

LEMMA 7 (Nikol'skii inequality). *Let $P(x) \in \Pi_M$; then for $1 \le p \le q \le \infty$,*

$$\|P'\|_{L^q_{[a,b]}} \le \frac{C}{(b-a)^{1+1/p-1/q}} M^{2+2/p-2/q} \|P\|_{L^p_{[a,b]}}.$$

LEMMA 8. *Under all conditions of Lemma 6, for any $s(x) \in \mathcal{S}(m+1, n_l)$ with $s'(0) \ge 0$ we have for large enough l,*

$$\|F_l - s\|_{L^p} \ge C\epsilon_{n_l} n_l^{-2+7\theta/8-\theta/(4p)} = C n_l^{-3+3\theta/8-\theta/(4p)} = C n_l^{-3+\theta/8+\theta^2/4}.$$

Proof. Applying Lemma 5, we get

$$\|F_l - Q_l\|_{L^p} \le C\epsilon_{n_l}^{1+1/p} |Q_l'(0)|.$$

Since $Q_l'(0) < 0$, $s(x) \in \mathcal{S}(m+1, n_l)$ satisfies $s'(0) \ge 0$, and $Q_l(x) - s(x)$ is a polynomial of degree m_l (since $m < m_l$ by (4.6)) on the interval $[0, 1/n_l]$. Applying Lemmas 5 and 7, together with (4.6), we obtain that

$$|Q_l'(0)| \le |Q_l'(0) - s'(0)| \le C n_l^{1+1/p} m_l^{2+2/p} \|Q_l - s\|_{L^p_{[0,1/n_l]}}$$

$$\le n_l^{1+1/p+\theta/(4p)} \left(\|Q_l - F_l\|_{L^p} + \|F_l - s\|_{L^p} \right).$$

Altogether,

$$\|F_l - Q_l\|_{L^p} \le C\epsilon_{n_l}^{1+1/p} n_l^{1+1/p+\theta/(4p)} \left(\|Q_l - F_l\|_{L^p} + \|F_l - s\|_{L^p} \right).$$

Because

$$\epsilon_{n_l}^{1+1/p} n_l^{1+1/p+\theta/(4p)} = n_l^{-(\theta/2)(1+1/p)+\theta/(4p)} = n_l^{-\theta/2-\theta/(4p)},$$

we then get, for large enough l,

$$\begin{aligned} \|F_l - s\|_{L^p} &\ge C\epsilon_{n_l}^{-1-1/p} n_l^{-1-1/p-\theta/(4p)} \|F_l - Q_l\|_{L^p} \\ &\ge C\epsilon_{n_l} n_l^{-1-1/p-\theta/8-\theta/(4p)} = C n_l^{-3+3\theta/8-\theta/(4p)} = C n_l^{-3+\theta/8+\theta^2/4}, \end{aligned}$$

which is the required result. □

Proof of Theorem 3. Select a sequence of natural numbers by induction. Set $n_1 = 1$,

(4.8)

$$n_{l+1} = \left[2^{8/\theta} \left(\max\{1, (B/A)^{8/\theta}\} n_l^{24/\theta} + \|F_l^{(k)}\| + \|F_l^{(3)}\|^{8/\theta} + m^{16p/\theta} + \delta^{-1} + 2^{8/\theta} \right) \right] + 1$$

for $l \ge 1$. Define

$$f(x) = \sum_{j=1}^{\infty} n_j^{-\theta/8} f_{n_j}(x).$$

Lemma 6 yields that $f \in C^1_{[0,1]} \cap \Delta^1(r)$ and $f'(x) > 0$ for $x \in (0, \beta_1/2]$. Write

$$\omega_2(f', n_l^{-1})_{L^p} \le \|F_{l-1}^{(3)}\| n_l^{-2} + n_l^{-\theta/8} \omega_2\left(f'_{n_l}, n_l^{-1}\right)_{L^p} + \sum_{j=l+1}^{\infty} n_j^{-\theta/8} \|f_{n_j}\| =: I_1 + I_2 + I_3.$$

From Lemma 4, it follows that

$$\begin{aligned} I_2 &= n_l^{-\theta/8} \omega_2(f'_{n_l} - h'_{n_l}, n_l^{-1})_{L^p} \le 4 n_l^{-\theta/8} \|f'_{n_l} - h'_{n_l}\|_{L^p} \\ &\le C \epsilon_{n_l}^{1+1/p} n_l^{-\theta/8} = C n_l^{-2+3\theta/8-\theta/(2p)}. \end{aligned} \tag{4.9}$$

Meanwhile, by (4.8),

$$I_1 \le C n_l^{-2+\theta/8}, \tag{4.10}$$

and obviously (4.8) also implies that

$$I_3 \le C n_{l+1}^{-\theta/8} \le C n_l^{-3}. \tag{4.11}$$

On the other hand, for any $s \in (m+1, n_l)$ with $s'(0) \ge 0$, by Lemma 8 and (4.11) we have for large enough l,

(4.12)

$$\begin{aligned} E^*_{n_l,m}(f)_{L^p} &= \min_s \|f(x) - s(x)\|_{L^p} \ge \min_s \|F_l - s\|_{L^p} - C \textstyle\sum_{j=l+1}^{\infty} n_j^{-\theta/8} \|f_{n_j}\| \\ &\ge C n_l^{-3+3\theta/8-\theta/(4p)} - C n_l^{-3} \ge C n_l^{-3+\theta/8+\theta^2/4}. \end{aligned}$$

All of the estimates (4.9)–(4.12) give

$$\frac{E^*_{n_l,m}(f)_{L^p}}{n_l^{-1}\omega_2(f', n_l^{-1})_{L^p}} \ge C \min\left\{n_l^{\theta^2/4}, (\epsilon_{n_l} n_l)^{-1/p} n_l^{-\theta/(4p)}\right\} = C \min\left\{n_l^{\theta^2/4}, n_l^{\theta/(4p)}\right\},$$

or

$$\limsup_{n\to\infty} \frac{E^*_{n,m}(f)_{L^p}}{n^{-1}\omega_2(f', n^{-1})_{L^p}} = +\infty,$$

and Theorem 3 is thus proved. □

5. Remarks.

Remark 1. From the proof of Theorem 3, we can see that $\theta = 1 - 1/p = 0$ when $p = 1$, so that it does not help us to achieve the same result as Theorem 3. This is the reason why $\omega_3(f', n^{-1})_{L^1}$ appears in Theorem 4 instead of $\omega_2(f', n^{-1})_{L^1}$. Since the technique is similar, we leave the proof of Theorem 4 to readers.

Remark 2. If $\theta = 1$, $m_l = [n_l^{1/16}]+1$, $k = 99$, $\epsilon_n = n^{-9/8}$, with other modifications, we can construct a function $f \in C^1_{[0,1]} \cap \Delta^1(r)$, by an argument similar to that of Theorem 3, such that

$$E^*_{n_l,m}(f) \ge C n_l^{-9/4},$$

and

$$\omega_3(f, n_l^{-1}) = O(n_l^{-19/8}).$$

This completes the proof of Theorem 5. We omit the details here.

Acknowledgment. The authors thank the referee for his helpful comments and suggestions for improvement of this paper.

REFERENCES

[1] C. K. CHUI, P. W. SMITH, AND J. D. WARD, *Degree of L_p approximation by monotone splines*, SIAM J. Math. Anal., 11(1980), pp. 436–447.

[2] R. A. DEVORE, *Degree of approximation*, in Approximation Theory II (Proc. Internat. Sympos., Univ. Texas, 1976), Academic Press, New York, 1976, pp. 117–162.

[3] ———, *Monotone approximation by splines*, SIAM J. Math. Anal., 8(1977), pp. 891–905.

[4] D. LEVIATAN, *The behavior of the derivatives of the algebraic polynomials of best approximation*, J. Approx. Theory, 35(1982), pp. 167–176.

[5] D. LEVIATAN AND H. N. MHASKAR, *The rate of monotone spline approximation in the L_p norm*, SIAM J. Math. Anal., 13(1982), pp. 866–874.

[6] ———, *Comonotone approximation by splines of piesewise monotone functions*, J. Approx. Theory, 35(1982), pp. 364–369.

[7] D. J. NEWMAN, *The Zygmund condition for polygonal approximation*, Proc. Amer. Math. Soc., 45(1974), pp. 303–304.

[8] I. A. SHEVCHUK, *On coapproximation of monotone functions*, Soviet Math. Dokl., 40(1990), pp. 349–354.

[9] X. WANG, *On a conjecture of DeVore*, J. Zhejiang Univ., 5(1986), pp. 106–108. (In Chinese.)

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1241–1258, September 1994

THE INFLUENCE OF DOMAIN AND DIFFUSIVITY PERTURBATIONS ON THE DECAY OF END EFFECTS IN HEAT CONDUCTION*

CHANGHAO LIN† AND L. E. PAYNE‡

Abstract. For a standard heat conduction problem in a semi-infinite cylinder the authors investigate the influence of domain perturbation and of the variation of the diffusivity coefficient on the decay of Saint Venant end effects. The particular problem investigated is one in which the lateral surface of the cylinder is maintained at zero temperature and a nonzero temperature is prescribed on the near end. Energy methods are used to assess the influence of the perturbations. A Phragmén–Lindelöf type alternative and explicit decay bounds are derived.

Key words. domain and diffusivity perturbations, continuous dependence, Saint Venant principle, Phragmén–Lindelöf theorem, heat conduction equation

AMS subject classifications. 35K05, 35B20, 35B30, 35B40

1. Introduction. It was shown by Knowles [8] that the temperature in a semi-infinite cylinder which is at zero temperature initially and exposed to some time varying temperature distribution on the finite end (the lateral surface being maintained at zero temperature) decays exponentially with distance from the finite end. Whereas Knowles derived an energy decay estimate, pointwise decay was established by Horgan, Payne, and Wheeler [7]. A natural question is the following: Suppose we wish to compare the solution of one temperature problem with that of another problem whose diffusivity coefficient is close to that of the first problem. If the data in the two problems are identical, is it possible to derive an explicit decay estimate which not only exhibits the known exponential decay of some appropriate norm of the difference of the two solutions, but also contains an amplitude term which for fixed time tends to zero as the difference between the diffusivity coefficients tends to zero? Knowles' result of course implies the exponential decay, but it does not imply that the two solutions remain near to one another. An analogous question for a related static problem of finite anti-plane shear deformation was considered by Horgan and Payne [5].

Alternatively we may wish to compare the solution of one heat conduction problem with that of a related problem for a semi-infinite cylinder whose cross section is a perturbation of that of the first cylinder. Then if the perturbation is small we would like to determine an explicit decay bound for some norm of the difference of the solutions which again not only exhibits the exponential decay but also contains an amplitude term which for fixed t tends to zero as the perturbation tends to zero. A related question for the elliptic system of linear isotropic elasticity was investigated by Horgan and Payne [4].

Specifically, we are concerned in this present paper with the heat equation defined on a semi-infinite cylinder in $\mathbb{R}^3$ with zero temperature on the lateral surface and prescribed temperature on the near end. In §2 we show that if the solution is bounded in an energy norm then it must decay exponentially in energy norm as the distance

Received by the editors June 30, 1992; accepted for publication (in revised form) October 7, 1992. This research was carried out while the first author held a visiting appointment at Cornell University and was partially supported by National Science Foundation grant DMS-9100786.

† Department of Mathematics, South China Normal University, Guangzhou 510631 China.
‡ Department of Mathematics, Cornell University, Ithaca, New York, 14853.

from the near end tends to infinity. In §3 we compare the solutions of two heat equations with different diffusivity coefficients and establish an explicit inequality which displays continuous dependence on this coefficient. In §4 we compare solutions of heat equations defined in two cylinders with different cross sections (one cross section may be regarded as a perturbation of the other) and derive estimates which show the influence of the perturbation of cross-sectional domain on the decay of solution.

Let R denote the cylindrical region given by

$$R = \{(x_1, x_2, x_3) \mid (x_1, x_2) \in D, x_3 > 0\}, \tag{1.1}$$

where D is a bounded simply-connected domain in $\mathbb{R}^2$ with smooth boundary ∂D. Both perturbation problems involve a basic problem whose solution $u(x,t)$ satisfies the equation

$$\Delta u = \nu u_{,t} \tag{1.2}$$

in the space-time region $R \times (0, \infty)$. In addition, $u(x,t)$ is required to satisfy the initial and boundary conditions

$$u(x_1, x_2, x_3, 0) = 0, \qquad (x_1, x_2, x_3) \in R, \tag{1.3}$$

$$u(x_1, x_2, x_3, t) = 0, \qquad (x_1, x_2) \in \partial D, \quad x_3 \geq 0, \ t \geq 0, \tag{1.4}$$

$$u(x_1, x_2, 0, t) = g(x_1, x_2, t), \qquad (x_1, x_2) \in D, \ t \geq 0, \tag{1.5}$$

$$\sup_t \left\{ \int_0^t \int_R u_{,i}\, u_{,i}\, dx d\tau + \frac{1}{2} \int_R u^2 dx \right\} \text{ bounded.} \tag{1.6}$$

In (1.2), Δ is the Laplace operator, ν is the reciprocal of the diffusivity coefficient, and a comma has been used to denote differentiation. The function $g(x_1, x_2, t)$ in (1.5) is assumed to be differentiable in t and to vanish on ∂D. For the diffusivity perturbation problem it would have been possible to treat the case of nonzero initial data, since the difference of the two solutions would satisfy homogeneous initial data. However, in this case it would have been somewhat more difficult to make the decay estimate explicit.

In what follows we adopt the convention of summing over repeated indices, Latin indices running from 1 to 3 and Greek indices from 1 to 2. A subscript preceded by a comma denotes partial differentiation with respect to the corresponding coordinate.

For the temperature field satisfying (1.2)–(1.6) Knowles [8] has shown that if $E(z,t)$ is defined as

$$E(z,t) = \int_0^t \int_{R_z} u_{,i}\, u_{,i}\, dx d\tau, \qquad z \geq 0, \ t \geq 0, \tag{1.7}$$

then

$$E(z,t) \leq E(0,t) e^{-2kz}. \tag{1.8}$$

In (1.7), R_z denotes the portion of the cylindrical domain R for which $x_3 > z$. We will subsequently use the symbol D_z to designate the intersection of R with the plane $x_3 = z$. The constant k in (1.8) is defined as

$$k = \sqrt{\lambda_1}, \tag{1.9}$$

where λ_1 is the first eigenvalue in the (fixed membrane) problem

$$\begin{aligned} \Delta\widehat{\varphi} + \lambda\widehat{\varphi} &= 0 \quad \text{in } D, \\ \widehat{\varphi} &= 0 \quad \text{on } \partial D, \end{aligned} \tag{1.10}$$

the constant λ_1 depends only on the geometry of D and numerous lower bounds for λ_1 can be found in the literature. (See, e.g., [1].)

In §2 we establish (1.8) without the decay assumptions imposed by Knowles. In fact, we derive a Phragmén–Lindelöf type alternative which shows that for each t either

$$\lim_{z\to\infty}\left[\int_0^t\int_{R/R_z} u_{,i}\,u_{,i}\,dxd\tau + \frac{1}{2}\int_{R/R_z} u^2dx\right]e^{-2kz} \geq M(t) \tag{1.11}$$

or

$$\begin{aligned} &\left[\int_0^t\int_{R_z} u_{,i}\,u_{,i}\,dxd\tau + \frac{1}{2}\int_{R_z} u^2dx\right] \\ &\qquad \leq \left[\int_0^t\int_R u_{,i}\,u_{,i}\,dxd\tau + \frac{1}{2}\int_R u^2dx\right]e^{-2kz}, \end{aligned} \tag{1.12}$$

where M is a positive function of t, and k is given by (1.9).

We shall say that our solution has bounded energy if $\sup_t\left[\int_0^t\int_R u_{,i}\,u_{,i}\,dxd\tau + \frac{1}{2}\int_R u^2dx\right]$ is less than some positive constant. We show in §2 that if for all $t \geq 0$

$$\lim_{z\to\infty}\left[\left\{\int_0^t\int_{D_z} u_{,i}\,u_{,i}\,dAd\tau + \frac{1}{2}\int_{D_z} u^2dA\right\}e^{-2kz}\right] = 0 \tag{1.13}$$

then the solution has bounded energy. Here dA is the element of area in D_z. In this paper we shall assume that the solution of various heat conduction problems in question satisfy (1.13) and thus possess bounded energy. This will eliminate the necessity of prescribing the uniform decay of solution as $z \to \infty$.

We observe that if we are interested only in the range $0 \leq t \leq T$ for some constant T, then in our definition of bounded energy, instead of $\sup_t$ we need only $\max_{t\in[0,T]}$, and (1.13) need hold only for $t \in [0,T]$.

We would like to derive continuous dependence inequalities, in the form of upper bounds for the square of the L^2 norm of the difference of solutions over a space-time region, which not only exhibit the desired continuous dependence but also decay at least as fast as the rate predicted by the work of Knowles. Our results fall slightly short of this goal in that the decay rates we obtain contain Knowles' decaying exponential e^{-2kz} multiplied by polynomial functions of z.

2. A Phragmén–Lindelöf type alternative. In this section we show that a solution of (1.2)–(1.5) must either grow exponentially in some measure or decay exponentially. To see this we set (following Horgan and Payne [6]; see also Flavin, Knops, and Payne [3])

$$F(z,t) = -\int_0^t\int_{D_z} uu_{,3}\,dAd\tau, \qquad z \geq 0,\ t \geq 0, \tag{2.1}$$

where, as mentioned earlier, the notation D_z is used to indicate that the integration

is to be taken over D in the plane $x_3 = z$. Note that $F(z,t)$ may also be written as

$$F(z,t) = F(0,t) - \int_0^t \int_{R/R_z} u_{,i}\, u_{,i}\, dx d\tau - \frac{1}{2}\int_{R/R_z} u^2 dx. \tag{2.2}$$

Thus if we differentiate (2.2) with respect to z we find

$$F'(z,t) = -\int_0^t \int_{D_z} u_{,i}\, u_{,i}\, dA d\tau - \frac{1}{2}\int_{D_z} u^2 dA, \tag{2.3}$$

where the prime indicates partial differentiation with respect to z. But

$$\begin{aligned}\left|\int_0^t \int_{D_z} u u_{,3}\, dA d\tau\right| &\le \left\{\int_0^t \int_{D_z} u^2 dA d\tau \cdot \int_0^t \int_{D_z} u_{,3}^2\, dA d\tau\right\}^{\frac{1}{2}} \\ &\le \frac{1}{k}\left\{\int_0^t \int_{D_z} u_{,\alpha}\, u_{,\alpha}\, dA d\tau \cdot \int_0^t \int_{D_z} u_{,3}^2\, dA d\tau\right\}^{\frac{1}{2}} \\ &\le \frac{1}{2k}\int_0^t \int_{D_z} u_{,i}\, u_{,i}\, dA d\tau.\end{aligned} \tag{2.4}$$

This leads to the inequality

$$|F| \le -\frac{1}{2k} F'(z,t), \tag{2.5}$$

or the two inequalities

$$F'(z,t) + 2kF(z,t) \le 0 \tag{2.6}$$

and

$$F'(z,t) - 2kF(z,t) \le 0. \tag{2.7}$$

We remark first that if F ever becomes negative for fixed t and some value of z, say $z = z_0$, then for that value of t, since F' is nonpositive, F must remain negative for all $z \ge z_0$. Then from (2.7) we have

$$-F(z,t) \ge -F(z_0,t) e^{2k(z-z_0)}. \tag{2.8}$$

Thus for that value of t, either $-F(z,t)$ eventually grows exponentially or $F(z,t) \ge 0$ for all values of z. But if $F(z,t) \ge 0$ for all z, then it follows from (2.6) that

$$F(z,t) \le F(0,t) e^{-2kz}. \tag{2.9}$$

Assuming $F(0,t) = -\int_0^t \int_{D_0} u u_{,3}\, dA d\tau$ is bounded we conclude that $F(z,t)$ decays (for fixed t) exponentially in z. Then from (2.2) we have

$$F(z,t) = \int_0^t \int_{R_z} u_{,i}\, u_{,i}\, dx d\tau + \frac{1}{2}\int_{R_z} u^2 dx \tag{2.10}$$

and

$$F(0,t) = \int_0^t \int_{R} u_{,i}\, u_{,i}\, dx d\tau + \frac{1}{2}\int_{R} u^2 dx. \tag{2.11}$$

It follows then that for fixed t, either

$$\lim_{z\to\infty}\left[\left\{\int_0^t \int_{R/R_z} u_{,i}\, u_{,i}\, dx d\tau + \frac{1}{2}\int_{R/R_z} u^2 dx\right\} e^{-2kz}\right] \ge M(t) \tag{2.12}$$

or

$$\int_0^t \int_{R_z} u_{,i}\, u_{,i}\, dx d\tau + \frac{1}{2}\int_{R_z} u^2 dx \leq \left[\int_0^t \int_R u_{,i}\, u_{,i}\, dx d\tau + \frac{1}{2}\int_R u^2 dx\right] e^{-2kz}. \tag{2.13}$$

Clearly any finite energy solution will violate (2.12) and hence for finite energy solutions (2.13) holds. Of course (2.13) will hold under a much weaker hypothesis than that of finite energy. In fact, if

$$\lim_{z\to\infty}\left[\left\{\int_0^t \int_{D_z} u_{,i}\, u_{,i}\, dA d\tau + \frac{1}{2}\int_{D_z} u^2 dA\right\} e^{-2kz}\right] = 0 \tag{2.14}$$

it follows from an application of L'Hopital's theorem that, since

$$\begin{aligned}\lim_{z\to\infty}&\left[\left\{\int_0^t \int_{R/R_z} u_{,i}\, u_{,i}\, dx d\tau + \frac{1}{2}\int_{R/R_z} u^2 dx\right\} e^{-2kz}\right\} \\ &= \lim_{z\to\infty}\frac{1}{2k}\left[\left\{\int_0^t \int_{D_z} u_{,i}\, u_{,i}\, dA d\tau + \frac{1}{2}\int_{D_z} u^2 dA\right\} e^{-2kz}\right],\end{aligned} \tag{2.15}$$

(2.12) is violated and inequality (2.13) holds. Thus instead of imposing the hypothesis of finite energy to conclude (2.13) it would actually be sufficient to impose hypothesis (2.14).

3. Continuous dependence on the diffusivity coefficient. We denote by $v(x,t)$ the solution of (1.2)–(1.5) with ν replaced by the constant $\tilde{\nu}$. We could, in fact, treat the case in which v differs from u on the end $x_3 = 0$, but since the problems are linear we could decompose the problems and investigate separately the perturbation in ν and the perturbation in g. Since the influence of the perturbation in g follows directly from the results of Knowles [8], we restrict attention in this paper to the case in which u and v satisfy the same boundary and initial conditions.

If we now set

$$w = u - v, \tag{3.1}$$

we note that w satisfies

$$\Delta w = \nu w_{,t} + (\nu - \tilde{\nu}) v_{,t} \quad \text{in } R \times (0,\infty), \tag{3.2}$$

$$w(x_1, x_2, x_3, 0) = 0 \quad \text{in } R, \tag{3.3}$$

$$w(x_1, x_2, x_3, t) = 0 \quad \text{on } \partial D \times [0,\infty), \tag{3.4}$$

$$w(x_1, x_2, 0, t) = 0, \qquad (x_1, x_2) \in D,\ t \geq 0. \tag{3.5}$$

We now define for $z \geq 0$, $t \geq 0$,

$$\Phi(z,t) = \int_0^t \int_{D_z} w^2 dA d\tau. \tag{3.6}$$

Differentiating Φ we obtain

$$\begin{aligned}\Phi'(z,t) &= 2\int_0^t \int_{D_z} ww_{,3}\, dA d\tau \\ &= -2\int_0^t \int_{R_z} [w\Delta w + w_{,i}\, w_{,i}]dx d\tau \\ &= -\nu \int_{R_z} w^2 dx - 2(\nu-\tilde{\nu})\int_0^t \int_{R_z} wv_{,\tau}\, dx d\tau \\ &\qquad - 2\int_0^t \int_{R_z} w_{,i}\, w_{,i}\, dx d\tau.\end{aligned} \tag{3.7}$$

Dropping the first term on the right of (3.7) and making use of Schwarz's inequality we conclude that

$$\begin{aligned}\Phi'(z,t) \le 2|\nu-\tilde{\nu}| &\left[\int_0^t \int_{R_z} w^2 dx d\tau \int_0^t \int_{R_z} v_{,\tau}^2\, dx d\tau\right]^{\frac{1}{2}} \\ &- 2\int_0^t \int_{R_z} w_{,i}\, w_{,i}\, dx d\tau.\end{aligned} \tag{3.8}$$

Since w vanishes on ∂D we make use of the inequality

$$\int_{D_z} w^2 dA \le \frac{1}{k^2}\int_{D_z} w_{,\alpha}\, w_{,\alpha}\, dA \tag{3.9}$$

and the arithmetic-geometric mean inequality to obtain

$$\Phi'(z,t) \le \frac{1}{\alpha(z)}(\nu-\tilde{\nu})^2 \int_0^t \int_{R_z} v_{,\tau}^2\, dx d\tau - \left[2 - \frac{\alpha(z)}{k^2}\right]\int_0^t \int_{R_z} w_{,i}\, w_{,i}\, dx d\tau \tag{3.10}$$

for some positive function $\alpha(z)$. Actually, we could allow α to also depend on t, but there is no advantage to doing so.

To express the last integral in (3.10) in terms of $\Phi(z,t)$, we note that

$$\begin{aligned}\int_{D_z} w^2 dA &= -2\int_{R_z} ww_{,3}\, dx \\ &\le 2\left[\int_{R_z} w^2 dx \int_{R_z} w_{,3}^2\, dx\right]^{\frac{1}{2}} \\ &\le \frac{1}{k}\int_{R_z} w_{,i}\, w_{,i}\, dx.\end{aligned} \tag{3.11}$$

In the last step we have used (3.9) and the arithmetic-geometric mean inequality. Integrating (3.11) with respect to t and inserting into (3.10) we obtain the differential inequality (provided $\alpha(z) \le 2\lambda_1$)

$$\Phi'(z,t) + k\left[2 - \frac{\alpha(z)}{k^2}\right]\Phi(z,t) \le \frac{1}{\alpha(z)}(\nu-\tilde{\nu})^2 \int_0^t \int_{R_z} v_{,\tau}\, dx d\tau. \tag{3.12}$$

To derive an explicit bound for the right-hand side of (3.12), we apply the same arguments as those used on $\Phi(z,t)$ to conclude that

$$\frac{d}{dz}\left\{\int_0^t \int_{D_z} v_{,\tau}^2\, dA d\tau\right\} \le -2k\int_0^t \int_{D_z} v_{,\tau}^2\, dA d\tau \tag{3.13}$$

which integrates to give

$$\int_0^t \int_{D_z} v_{,\tau}^2 \, dA d\tau \leq Q_0(t) e^{-2kz}, \tag{3.14}$$

where

$$Q_0(t) = \int_0^t \int_{D_0} v_{,\tau}^2 \, dA d\tau = \int_0^t \int_D g_{,\tau}^2 \, dA d\tau. \tag{3.15}$$

An integration of (3.14) with respect to z leads to

$$\int_0^t \int_{R_z} v_{,\tau}^2 \, dx d\tau \leq \frac{Q_0(t)}{2k} e^{-2kz}. \tag{3.16}$$

Thus inserting (3.16) into (3.12) we are led to

$$\Phi'(z,t) + k\left[2 - \frac{\alpha(z)}{k^2}\right]\Phi(z,t) \leq \frac{(\nu - \tilde{\nu})^2}{2k\alpha(z)} Q_0(t) e^{-2kz}. \tag{3.17}$$

We now make the choice

$$\alpha(z) = \frac{k}{z + \frac{1}{2k}}. \tag{3.18}$$

But (3.17) may then be rewritten as

$$[\Phi(z,t) e^{2kz} (2kz+1)^{-1}]' \leq \frac{(\nu - \tilde{\nu})^2}{4k^3} Q_0(t). \tag{3.19}$$

An integration then leads to

$$\Phi(z,t) \leq \frac{(\nu - \tilde{\nu})^2}{4k^3} Q_0(t) z(2kz+1) e^{-2kz}. \tag{3.20}$$

This is the desired continuous dependence inequality.

We remark again that the decay rate is not quite as fast as we would like since we know by Knowles' result (1.8) that $\Phi(z,t)$ should decay at least of order e^{-2kz}. However the factor $(\nu - \tilde{\nu})^2$ in the amplitude term gives a measure of the closeness of u and v in energy measure even for small value of z.

We may of course integrate (3.20) to find a bound for $\int_0^t \int_{R_z} w^2 dx d\tau$, i.e.,

$$\int_0^t \int_{R_z} w^2 dx d\tau \leq \frac{(\nu - \tilde{\nu})^2}{16k^5} Q_0(t)[4k^2z^2 + 6kz + 3] e^{-2kz}. \tag{3.21}$$

We have thus established the following theorem.

THEOREM 1. *Let u be the solution of problem* (1.2)–(1.5) *and let v be the solution of the same problem with ν replaced by $\tilde{\nu}$. Then for arbitrary $z \geq 0$, $t \geq 0$, the quantity $u - v$ satisfies the following inequalities*:

$$\int_0^t \int_{D_z} (u-v)^2 ds d\tau \leq \frac{(\nu - \tilde{\nu})^2}{4k^3} Q_0(t) z(2kz+1) e^{-2kz}, \tag{3.22}$$

and

$$\int_0^t \int_{R_z} (u-v)^2 dx d\tau \leq \frac{(\nu - \tilde{\nu})^2}{16k^5} Q_0(t)[4k^2z^2 + 6kz + 3] e^{-2kz}, \tag{3.23}$$

where k is given by (1.9), (1.10) *and* $Q_0(t)$ *by* (3.15).

The choice of $\alpha(z)$ in (3.18) was made with two ideas in mind, i.e., easy computation and a resulting decay rate which is close to that found by Knowles. Choosing α to be a constant would make the computations simpler at the expense of a resulting decay bound in which the decay rate is not as sharp. On the other hand the decay rate could be improved slightly by a somewhat more complicated choice for $\alpha(z)$—a choice which would lead to more involved computations.

4. The influence of geometric perturbation. In this section we wish to compare the solution $u(x,t)$ of (1.2)–(1.5) defined on $\hat{R} \times (0,\infty)$ with the solution $v(x,t)$ of the analogous problem defined on $\tilde{R} \times (0,\infty)$, where

$$\hat{R} = \{(x_1,x_2,x_3) \mid (x_1,x_2) \in \hat{D}, x_3 \geq 0\},$$
$$\tilde{R} = \{(x_1,x_2,x_3) \mid (x_1,x_2) \in \tilde{D}, x_3 \geq 0\}.$$

Here we regard $\tilde{D}$ as a perturbation of $\hat{D}$ and assume that $\hat{D}$ and $\tilde{D}$ have smooth boundaries $\partial\hat{D}$ and $\partial\tilde{D}$ respectively. For simplicity we assume further that $\hat{D}$ and $\tilde{D}$ are both star-shaped with respect to a point $P \in \hat{D} \cap \tilde{D}$ which we take as the origin in the plane. Specifically u is a solution of (1.2)–(1.5) with $g = \hat{g}$, $D = \hat{D}$ and $R = \hat{R}$ respectively and $v(x,t)$ is a solution of

$$\Delta v = \nu v_{,t} \quad \text{in } \tilde{R} \times (0,\infty) \tag{4.1}$$

subject to the conditions

$$v(x_1,x_2,x_3,0) = 0, \qquad (x_1,x_2,x_3) \in \tilde{R}, \tag{4.2}$$

$$v(x_1,x_2,x_3,t) = 0, \qquad (x_1,x_2) \in \partial\tilde{D}, \quad x_3 \geq 0, \ t \geq 0, \tag{4.3}$$

$$v(x_1,x_2,0,t) = \tilde{g}(x_1,x_2,t), \qquad (x_1,x_2) \in \tilde{D}, \ t \geq 0. \tag{4.4}$$

In this section we assume that $\hat{g}(x_1,x_2,t)$ is piecewise $C^{1,2}(\hat{D} \times [0,\infty))$ and vanishes on $\partial\hat{D}$ while $\hat{g}(x_1,x_2,t)$ is piecewise $C^{1,2}(\tilde{D} \times [0,\infty))$ and vanishes on $\partial\tilde{D}$. We compare solution $u(x,t)$ of (1.2)–(1.5) with $v(x,t)$ of (4.1)–(4.4) over $\hat{R} \cap \tilde{R} \times (0,\infty)$. To this end we employ in this section the notation

$$\begin{aligned} D &= \hat{D} \cap \tilde{D} \neq \phi, \\ D' &= \hat{D} \cup \tilde{D}. \end{aligned} \tag{4.5}$$

If we now set

$$w = u - v \tag{4.6}$$

in $R \times (0,\infty)$, where $R = \{(x_1,x_2,x_3) \mid (x_1,x_2) \in D, \ x_3 \geq 0\}$, then w satisfies

$$\Delta w = \nu w_{,t} \quad \text{in } R \times (0,\infty) \tag{4.7}$$

with

$$w(x_1,x_2,x_3,0) = 0, \qquad (x_1,x_2,x_3) \in R, \tag{4.8}$$

$$w(x_1,x_2,x_3,t) = u - v \quad \text{on } \partial D \times [0,\infty), \tag{4.9}$$

$$w(x_1,x_2,0,t) = \hat{g} - \tilde{g}, \qquad (x_1,x_2) \in D, \ t \geq 0. \tag{4.10}$$

Again this problem may be decomposed into one with zero boundary conditions on the lateral surface and one with zero conditions on the end $x_3 = 0$. The first problem has already been considered by Knowles [8] so we consider only the problem in which $\hat{g} - \tilde{g} = 0$ in $D \times [0, \infty)$. Also the constant ν may be scaled out by redefining the time variable. Thus we assume in this section that $\nu = 1$.

Since it is difficult to deal with the inhomogeneous data on the lateral surface we introduce an auxiliary function $H(x,t)$ which is a solution, for each t, of

$$\Delta H = 0 \quad \text{in } R, \tag{4.11}$$

$$H = w - v \quad \text{on } \partial D,\ x_3 \geq 0, \tag{4.12}$$

$$H = 0 \quad \text{in } D,\ x_3 = 0, \tag{4.13}$$

$$H \to 0 \text{ (uniformly in } x_1, x_2) \text{ as } x_3 \to \infty. \tag{4.14}$$

From the triangle inequality we then have

$$\|w\| \leq \|w - H\| + \|H\|, \tag{4.15}$$

where the norm is the L^2 norm over $R_z \times (0,t)$, i.e.,

$$\|\varphi\|^2 = \int_0^t \int_{R_z} \varphi^2 dx d\tau. \tag{4.16}$$

To find a decay bound for $\|w\|$ we will then derive decay bounds for $\|w - H\|$ and for $\|H\|$.

We first investigate the term $\|H\|$ in (4.15), since the results of this computation will be needed in deriving the decay estimate for $\|w - H\|$. We write

$$H = H_1 + H_2 \tag{4.17}$$

where for each t,

$$\Delta H_1 = 0 \quad \text{in } R_z, \tag{4.18}$$

$$H_1 = 0 \quad \text{on } \partial D \times [z, \infty), \tag{4.19}$$

$$H_1 = H \quad \text{on } D_z, \tag{4.20}$$

$$H_1 \to 0 \text{ (uniformly in } x_1, x_2) \quad \text{as } x_3 \to \infty, \tag{4.21}$$

and

$$\Delta H_2 = 0 \quad \text{in } R_z, \tag{4.22}$$

$$H_2 = u - v \quad \text{on } \partial D \times [z, \infty), \tag{4.23}$$

$$H_2 = 0 \quad \text{on } D_z, \tag{4.24}$$

$$H_2 \to 0 \text{ (uniformly in } x_1, x_2) \text{ as } x_3 \to \infty. \tag{4.25}$$

From the triangle inequality we then have

$$\|H\| \leq \|H_1\| + \|H_2\|. \tag{4.26}$$

We now derive a bound for $\|H_1\|$. It is well known (see, e.g., Payne [9]) that

$$\int_{R_z} H_1^2 dx \leq \frac{1}{p} \int_{D_z} H^2 dA, \tag{4.27}$$

where p is the first eigenvalue in the problem

$$\Delta^2 B = 0 \quad \text{in } R_z, \tag{4.28}$$

$$B, \Delta B = 0 \quad \text{on } \partial D \times (z, \infty), \tag{4.29}$$

$$B, \Delta B + p\frac{\partial B}{\partial x_3} = 0 \quad \text{on } D_z, \tag{4.30}$$

$$B \to 0 \text{ (uniformly in } x_1, x_2) \quad \text{as } x_3 \to \infty. \tag{4.31}$$

For the cylindrical domain, R_z, the first eigenfunction B_1 is easily seen to be

$$B_1 = \text{const}(x_3 - z)e^{-\sqrt{\lambda_1}(x_3 - z)}\hat{\varphi}_1(x_1, x_2), \tag{4.32}$$

where $\hat{\varphi}_1$ and λ_1 are given by (1.10). An easy computation leads to

$$p = 2\sqrt{\lambda_1} = 2k. \tag{4.33}$$

Thus we find from (4.27) with an integration in t

$$\|H_1\|^2 \leq \frac{1}{2k} \int_0^t \int_{D_z} H^2 dA d\tau. \tag{4.34}$$

To bound $\|H_2\|$ we introduce the auxiliary function φ which, for each t, is a solution of

$$\Delta\varphi = -H_2 \quad \text{in } R_z, \tag{4.35}$$

$$\varphi = 0 \quad \text{on } \partial R_z, \tag{4.36}$$

$$\varphi \to 0 \text{ (uniformly in } x_1, x_2) \quad \text{as } x_3 \to \infty. \tag{4.37}$$

Then

$$\begin{aligned} \int_{R_z} H_2^2 dx &= -\int_{R_z} H_2 \Delta\varphi dx \\ &= -\int_z^\infty \oint_{\partial D_\eta} [u - v]\varphi_{,\alpha} n_\alpha ds d\eta \\ &\leq \left\{ \int_z^\infty \oint_{\partial D_\eta} [u - v]^2 ds d\eta \int_z^\infty \oint_{\partial D_\eta} [\varphi_{,\alpha} n_\alpha]^2 ds d\eta \right\}^{\frac{1}{2}} \end{aligned} \tag{4.38}$$

where n_α $(\alpha = 1, 2)$ denotes the component of outward unit normal vector on ∂D.

To compute a bound for the integral of $(\varphi_{,\alpha} n_\alpha)^2$ over the lateral surface we use a Rellich identity (see, e.g., [10]) obtained by an integration of the identity

$$\int_{R_z} x_\beta \varphi_{,\beta} [\Delta\varphi + H_2] dx = 0. \tag{4.39}$$

An integration by parts yields

$$\int_z^\infty \oint_{\partial D_\eta} (\varphi_{,\alpha}\, n_\alpha)^2 x_\beta n_\beta ds d\eta + \int_{R_z} \varphi_{,3}^2\, dx = -\int_{R_z} x_\beta \varphi_{,\beta}\, H_2 dx, \tag{4.40}$$

which leads to

$$\int_z^\infty \oint_{\partial D_\eta} (\varphi_{,\alpha}\, n_\alpha)^2 ds d\eta \le \frac{2\gamma_M}{h_0} \left\{ \int_{R_z} \varphi_{,\beta}\, \varphi_{,\beta}\, dx \int_{R_z} H_2^2 dx \right\}^{\frac{1}{2}} \tag{4.41}$$

where

$$h_0 = \max_{\partial D} x_\alpha n_\alpha;\ \gamma_M = \max_D [x_\alpha x_\alpha]^{\frac{1}{2}}. \tag{4.42}$$

But

$$\begin{aligned} \int_{R_z} \varphi_{,\beta}\, \varphi_{,\beta}\, dx &\le \int_{R_z} \varphi_{,i}\, \varphi_{,i}\, dx = \int_{R_z} \varphi H_2 dx \\ &\le \frac{1}{k} \left\{ \int_{R_z} \varphi_{,\beta}\, \varphi_{,\beta}\, dx \int_{R_z} H_2^2 dx \right\}^{\frac{1}{2}}. \end{aligned} \tag{4.43}$$

Thus from (4.43) and (4.41) we have

$$\int_z^\infty \oint_{\partial D_z} (\varphi_{,\alpha}\, n_\alpha)^2 ds d\eta \le \frac{2\gamma_M}{h_0 k} \int_{R_z} H_2^2 dx. \tag{4.44}$$

Inserting (4.44) into (4.38), integrating with respect to t, and using Schwarz's inequality, we find

$$\|H_2\|^2 \le \frac{2\gamma_M}{h_0 k} \int_0^t \int_z^\infty \oint_{\partial D_\eta} (u - v)^2 ds d\eta d\tau. \tag{4.45}$$

Thus substituting (4.34) and (4.45) into (4.26), we find

$$\|H\| \le \left\{ \frac{1}{2k} \int_0^t \int_{D_z} H^2 dA \right\}^{\frac{1}{2}} + \left\{ \frac{2\gamma_M}{h_0 k} \int_0^t \int_z^\infty \oint_{\partial D_\eta} (u - v)^2 ds d\eta d\tau \right\}^{\frac{1}{2}}. \tag{4.46}$$

We next derive a bound for the second integral on the right of (4.46) using arguments similar to those used in [2]. We introduce, for fixed x_3 and t, the notation

$$u^* = \begin{cases} u & \text{in } \hat{D}, \\ 0 & \text{in } \mathbb{R}^2/\hat{D}, \end{cases}$$

$$v^* = \begin{cases} v & \text{in } \tilde{D}, \\ 0 & \text{in } \mathbb{R}^2/\tilde{D}, \end{cases}$$

and set

$$w^* = u^* - v^*. \tag{4.47}$$

Now if (r_0, θ_0) is a point on ∂D, and (r_1, θ_0) is the point on $\partial D'$ intersected by the ray through (r_0, θ_0), then we have

$$|w^*(r_0, \theta_0)| = \left| \int_{r_0}^{r_1} \frac{\partial w^*}{\partial \rho} d\rho \right| \le \left[\delta(\theta) \int_{r_0}^{r_1} \left[\frac{\partial w^*}{\partial \rho} \right]^2 d\rho \right]^{\frac{1}{2}}. \tag{4.48}$$

Using (4.48) we may then write

$$\begin{aligned}\oint_{\partial D}(u-v)^2 ds &= \oint_{\partial D}[w^*]^2 ds = \int_0^{2\pi}[w^*]^2\frac{\rho}{n_\rho}d\theta \\ &\le \int_0^{2\pi}\int_{r_0}^{r_1}\delta(\theta)\left[\frac{\partial w^*}{\partial \rho}\right]\frac{\rho}{n_\rho}d\rho d\theta \\ &\le c\delta\int_{D'/D} w^*_{,\alpha}\, w^*_{,\alpha}\, dA,\end{aligned} \tag{4.49}$$

where n_ρ is the radial component of the unit normal vector on ∂D and

$$\begin{aligned}\delta &= \max_{0\le\theta\le 2\pi}\delta(\theta), \\ c &= \max_{\partial D}[n_\rho^{-1}].\end{aligned} \tag{4.50}$$

Clearly δ is the maximum distance along a ray between $\partial\hat{D}$ and $\partial\tilde{D}$.

Thus we have

(4.51)

$$\begin{aligned}&\int_0^t\int_z^\infty \oint_{\partial D_\eta}(u-v)^2 ds d\eta d\tau \\ &\quad\le c\delta\int_0^t\int_{R'_z/R_z} w^*_{,\alpha}\, w^*_{,\alpha}\, dx d\tau \\ &\quad\le 2c\delta\left[\int_0^t\int_{\hat{R}_z/R_z} u_{,i}\, u_{,i}\, dx d\tau + \int_0^t\int_{\tilde{R}_z/R_z} v_{,i}\, v_{,i}\, dx d\tau\right] \\ &\quad\le 2c\delta[\hat{E}_1(z,t)+\tilde{E}_1(z,t)],\end{aligned}$$

where

$$\hat{E}_1(z,t) = \int_0^t\int_{\hat{R}_z} u_{,i}\, u_{,i}\, dx d\tau, \tag{4.52}$$

$$\tilde{E}_1(z,t) = \int_0^t\int_{\tilde{R}_z} v_{,i}\, v_{,i}\, dx d\tau. \tag{4.53}$$

From the results of Knowles [8] we know that

$$\hat{E}_1(z,t) \le \hat{E}_1(0,t)e^{-2\hat{k}z}, \tag{4.54}$$

$$\tilde{E}_1(z,t) \le \tilde{E}_1(0,t)e^{-2\tilde{k}z}, \tag{4.55}$$

where

$$\hat{k} = \sqrt{\lambda_1(\hat{D})}; \qquad \tilde{k} = \sqrt{\lambda_1(\tilde{D})}. \tag{4.56}$$

Let $\overline{k}$ be defined by

$$\overline{k} = \min[\hat{k}, \tilde{k}]. \tag{4.57}$$

Then (4.46) becomes

(4.58)
$$\|H\| \le \left\{\frac{4\gamma_M c\delta}{h_0 k}[\hat{E}_1(0,t)+\tilde{E}_1(0,t)]\right\}^{\frac{1}{2}} e^{-\bar{k}z} + \left\{\frac{1}{2k}\int_0^t\int_{D_z} H^2 dA d\tau\right\}^{\frac{1}{2}}$$
$$= \left\{S_1(t)\delta e^{-2\bar{k}z}\right\}^{\frac{1}{2}} + \left\{\frac{1}{2k}\int_0^t\int_{D_z} H^2 dA d\tau\right\}^{\frac{1}{2}},$$

where

(4.59)
$$S_1 = \frac{4\gamma_M c}{h_0 k}[\hat{E}_1(0,t)+\tilde{E}_1(0,t)].$$

Using the arithmetic-geometric mean inequality for an arbitrary positive $\beta(z)$, we have

(4.60)
$$\|H\|^2 \le \frac{1+\beta(z)}{\beta(z)} S_1\delta e^{-2\bar{k}z} + \frac{1+\beta(z)}{2k}\int_0^t\int_{D_z} H^2 dA d\tau.$$

Choosing

(4.61)
$$\beta(z) = \frac{1}{2kz}$$

and setting

(4.62)
$$\chi(z,t) = \|H\|^2,$$

we find

(4.63)
$$\chi'(z,t) + \left(2k - \frac{1}{z+1/(2k)}\right)\chi(z,t) \le 4k^2 z S_1\delta e^{-2\bar{k}z}$$

or

(4.64)
$$\left\{\chi(z,t)e^{2kz}\left[z+\frac{1}{2k}\right]^{-1}\right\}' \le \frac{4k^2 z S_1\delta}{z+1/(2k)} e^{2(k-\bar{k})z}$$
$$\le 4k^2 S_1\delta e^{2(k-\bar{k})z}.$$

Since $D \subset \hat{D}$ and $D \subset \tilde{D}$ we know from the monotonicity of λ_1 that $k \ge \bar{k}$ with equality iff $\hat{D}$ and $\tilde{D}$ are identical. Inequality (4.64) integrates to give

(4.65)
$$\|H\|^2 \le \frac{kS_1\delta(2kz+1)}{k-\bar{k}}\left[e^{-2\bar{k}z} - e^{-2kz}\right]$$
$$\le 2kS_1\delta(2kz+1)ze^{-2\bar{k}z}.$$

In the last step of (4.65), we have used the fact that for arbitrary constants x and x_0

(4.66)
$$e^x \ge e^{x_0} + (x-x_0)e^{x_0},$$

or

(4.67)
$$e^{-2kz} \ge e^{-2\bar{k}z} - 2(k-\bar{k})ze^{-2\bar{k}z}.$$

Since $H_{,t}$ satisfies the same equation as H we also have the inequality

$$\begin{aligned} \|H_{,t}\|^2 &\leq \frac{kS_2\delta(2kz+1)}{(k-\bar{k})}\left[e^{-2\bar{k}z}-e^{-2kz}\right] \\ &\leq 2kS_2\delta(2kz+1)ze^{-2\bar{k}z}, \end{aligned} \tag{4.68}$$

where

$$S_2 = \frac{4\gamma_M c}{h_0 k}[\hat{E}_2(0,t)+\tilde{E}_2(0,t)] \tag{4.69}$$

and

$$\hat{E}_2(0,t) = \int_0^t \int_{\hat{R}} u_{,i\tau}\, u_{,i\tau}\, dxd\tau \tag{4.70}$$

$$\tilde{E}_2(0,t) = \int_0^t \int_{\tilde{R}} v_{,i\tau}\, v_{,i\tau}\, dxd\tau. \tag{4.71}$$

We require (4.68) in computing the bound for $\|w-H\|$. Bounds for $\hat{E}_\alpha(0,t)$ and $\tilde{E}_\alpha(0,t)$ $(\alpha=1,2)$ are given in the Appendix.

We turn now to the derivation of the bound for $\|w-H\|$. Setting

$$G(z,t) = \int_0^t \int_{D_z} [w-H]^2 dAd\tau \tag{4.72}$$

we have

$$\begin{aligned} G'(z,t) &= 2\int_0^t \int_{D_z} [w-H][w-H]_{,3}\, dAd\tau \\ &= -2\int_0^t \int_{R_z} (w-H)\Delta w dxd\tau \\ &\quad -2\int_0^t \int_{R_z} (w-H)_{,i}\,(w-H)_{,i}\, dxd\tau \\ &= -\int_{R_z} (w-H)^2 dx - 2\int_0^t \int_{R_z} (w-H)H_{,\tau}\, dxd\tau \\ &\quad -2\int_0^t \int_{R_z} (w-H)_{,i}\,(w-H)_{,i}\, dxd\tau. \end{aligned} \tag{4.73}$$

Dropping the first term on the right and making use of Schwarz's inequality we have for arbitrary $\gamma_1(z)$

$$\begin{aligned} G'(z,t) &\leq 2\left\{\int_0^t \int_{R_z} (w-H)^2 dxd\tau \int_0^t \int_{R_z} H_{,\tau}^2\, dxd\tau\right\}^{\frac{1}{2}} \\ &\quad -2\int_0^t \int_{R_z} (w-H)_{,i}\,(w-H)_{,i}\, dxd\tau \\ &\leq -\left[2-\frac{\gamma_1(z)}{k}\right]\int_0^t \int_{R_z} (w-H)_{,i}\,(w-H)_{,i}\, dxd\tau \\ &\quad +\frac{2S_2\delta(2kz+1)z}{\gamma_1(z)}e^{-2\bar{k}z}. \end{aligned} \tag{4.74}$$

We have made use of the arithmetic-geometric mean inequality and (4.68) in the last step. Choosing

$$\gamma_1(z) = \frac{1}{z + 1/(2k)}, \tag{4.75}$$

we find (using (3.11) with w replaced by $w - h$)

$$G'(z,t) + \left(2k - \frac{1}{z + 1/(2k)}\right) G(z,t) \leq \frac{1}{k} S_2 \delta (2kz+1)^2 z e^{-2\bar{k}z}. \tag{4.76}$$

Now (4.76) may be rewritten as

$$\left[G e^{2kz} \left(z + \frac{1}{2k}\right)^{-1} \right]' \leq 2 S_2 \delta (2kz+1) z e^{2(k-\bar{k})z} \tag{4.77}$$

and (4.77) integrates to give

(4.78)

$$\begin{aligned} \int_0^t \int_{D_z} (w-H)^2 dA d\tau &\leq 2 S_1 \delta \left(z + \frac{1}{2k}\right) \left[\int_0^z \eta (2k\eta + 1) e^{2(k-\bar{k})\eta} d\eta \right] e^{-2kz} \\ &\leq \frac{S_2 \delta (2kz+1)^2 z}{k} \cdot \frac{e^{-2\bar{k}z} - e^{-2kz}}{2(k-\bar{k})} \\ &\leq \frac{S_2 \delta (2kz+1)^2 z^2}{k} e^{-2\bar{k}z}. \end{aligned}$$

Here, for convenience, we have factored out a maximum of $\eta(2k\eta + 1)$ in the second step, and used the inequality (4.67) in the last step.

To obtain the desired bound for $\|w - H\|$ we must integrate (4.78) with respect to x_3 from z to infinity. This leads to an inequality of the form

$$\|w - H\|^2 \leq \frac{1}{k} S_2 \delta Q_1(z) e^{-2\bar{k}z}, \tag{4.79}$$

where $Q_1(z)$ is an easily computable quartic function of z.

We list here the inequalities derived in the Appendix,

$$\hat{E}_1(0,t) \leq \int_0^t \int_{\hat{D}} \left[4\hat{k}\hat{g}^2 + \frac{1}{2\hat{k}} \hat{g}_{,\alpha}\, \hat{g}_{,\alpha} + \frac{1}{16\hat{k}^3} \hat{g}_{,\tau}^2 \right] dA d\tau = B_1(t), \tag{4.80}$$

$$\tilde{E}_1(0,t) \leq \int_0^t \int_{\tilde{D}} \left[4\tilde{k}\tilde{g}^2 + \frac{1}{2\tilde{k}} \tilde{g}_{,\alpha}\, \tilde{g}_{,\alpha} + \frac{1}{16\tilde{k}^3} \hat{g}_{,\tau}^2 \right] dA d\tau = B_2(t), \tag{4.81}$$

$$\hat{E}_2(0,t) \leq \int_0^t \int_{\hat{D}} \left[4\hat{k}\hat{g}_{,\tau}^2 + \frac{1}{2\hat{k}} \hat{g}_{,\alpha\tau}\, \hat{g}_{,\alpha\tau} + \frac{1}{16\hat{k}^3} \hat{g}_{,\tau\tau}^2 \right] dA d\tau = B_3(t), \tag{4.82}$$

$$\tilde{E}_2(0,t) \leq \int_0^t \int_{\tilde{D}} \left[4\tilde{k}\tilde{g}_{,\tau}^2 + \frac{1}{2\tilde{k}} \tilde{g}_{,\alpha\tau}\, \tilde{g}_{,\alpha\tau} + \frac{1}{16\tilde{k}^3} \tilde{g}_{,\tau\tau}^2 \right] dA d\tau = B_4(t). \tag{4.83}$$

We have thus established the following theorem.

THEOREM 2. *Let u be a solution of* (1.2)–(1.5) *in $\hat{R}\times(0,\infty)$ and v be a solution of* (4.1)–(4.4) *in $\tilde{R}\times(0,\infty)$. Then if $\hat{g}=\tilde{g}$ in $\hat{D}\cap\tilde{D}$ it follows that for arbitrary $z\geq 0$, $t\geq 0$ and $\overline{k}$ given by* (4.57),

$$\begin{aligned}\|u-v\| \leq \delta^{\frac{1}{2}}\Big\{&A_1[(2kz+1)z]^{\frac{1}{2}}[B_1(t)+B_2(t)]^{\frac{1}{2}}\\ &+A_2[Q_1(z)]^{\frac{1}{2}}[B_3(t)+B_4(t)]^{\frac{1}{2}}\Big\}e^{-\overline{k}z},\end{aligned}\tag{4.84}$$

where A_1, A_2 are constants defined as

$$A_1=\frac{8\gamma_M c}{h_0},\qquad A_2=\frac{4\gamma_M c}{h_0 k^2},\tag{4.85}$$

the B_i's are given by (4.80)–(4.83), *δ and c by* (4.50), *h_0 and γ_M by* (4.42), *and $Q_1(z)$ is a computable quartic function of z.*

The decay rate indicated by (4.84) is essentially the decay rate from the results of Knowles [8]. However the factor $\delta^{1/2}$ in (4.84) displays the continuous dependence on the geometry. With increased smoothness assumptions on $\hat{g}$ and $\tilde{g}$, we could bound higher norm of the solution and thus obtain a bound of type (4.76) which contains a larger exponent on the δ. Likewise, we clearly can relax the smoothness assumptions on $\partial\hat{D}$ and $\partial\tilde{D}$. However, we do not pursue these questions in this paper.

Appendix. We derive here the inequalities (4.80)–(4.83). In fact it suffices to derive (4.80). The other three inequalities follow from a similar argument. Now

$$\begin{aligned}\hat{E}_1(0,t)&=\int_0^t\int_{\hat{R}}u_{,i}\,u_{,i}\,dxd\tau\\ &=-\int_0^t\int_{\hat{D}_0}uu_{,3}\,dAd\tau-\int_0^t\int_{\hat{R}}uu_{,\tau}\,dxd\tau\\ &\leq\left\{\int_0^t\int_{\hat{D}_0}\hat{g}^2dAd\tau\int_0^t\int_{\hat{D}_0}u_{,3}^2\,dAd\tau\right\}^{\frac{1}{2}}.\end{aligned}\tag{A.1}$$

But

$$\begin{aligned}\int_0^t\int_{\hat{D}_0}u_{,3}^2\,dAd\tau&=-2\int_0^t\int_{\hat{R}}u_{,3}\,u_{,33}\,dxd\tau\\ &=-2\int_0^t\int_{\hat{R}}u_{,3}\,[\Delta u-u_{,\alpha\alpha}]dxd\tau\\ &=\int_0^t\int_{\hat{D}_0}\hat{g}_{,\alpha}\,\hat{g}_{,\alpha}\,dAd\tau-2\int_0^t\int_{\hat{R}}u_{,3}\,u_{,\tau}\,dxd\tau.\end{aligned}\tag{A.2}$$

From (A.2) we have for arbitrary positive constant σ

$$\begin{aligned}\int_0^t\int_{\hat{D}_0}u_{,3}^2\,dAd\tau\leq&\int_0^t\int_{\hat{D}_0}\hat{g}_{,\alpha}\,\hat{g}_{,\alpha}\,dAd\tau\\ &+\sigma^{-1}\int_0^t\int_{\hat{R}}u_{,3}^2\,dxd\tau+\sigma\int_0^t\int_{\hat{R}}u_{,\tau}^2\,dxd\tau.\end{aligned}\tag{A.3}$$

But

$$\begin{aligned}\int_0^t \int_{\hat{R}} u_{,\tau}^2 \, dx d\tau &= \int_0^t \int_{\hat{R}} u_{,\tau} \, \Delta u dx d\tau \\ &\leq \left\{ \int_0^t \int_{\hat{D}_0} \hat{g}_{,\tau}^2 \, dA d\tau \int_0^t \int_{\hat{D}_0} u_{,3}^2 \, dA d\tau \right\}^{\frac{1}{2}} \\ &\leq \frac{1}{2\sigma_1} \int_0^t \int_{\hat{D}_0} \hat{g}_{,\tau}^2 \, dA d\tau + \frac{\sigma_1}{2} \int_0^t \int_{\hat{D}_0} u_{,3}^2 \, dA d\tau\end{aligned} \tag{A.4}$$

for some positive constant σ_1. Inserting (A.4) into (A.3) we have

$$\begin{aligned}&\left(1 - \frac{\sigma\sigma_1}{2}\right) \int_0^t \int_{\hat{D}_0} u_{,3}^2 \, dA d\eta \\ &\quad \leq \int_0^t \int_{\hat{D}_0} \hat{g}_{,\alpha} \, \hat{g}_{,\alpha} \, dA d\tau \\ &\qquad + \sigma^{-1} \int_0^t \int_{\hat{R}} u_{,3}^2 \, dx d\eta + \frac{\sigma}{2\sigma_1} \int_0^t \int_{\hat{D}_0} \hat{g}_{,\tau}^2 \, dA d\tau \\ &\quad \leq \int_0^t \int_{\hat{D}_0} \hat{g}_{,\alpha} \, \hat{g}_{,\alpha} \, dA d\tau + \sigma^{-1} \hat{E}_1(0,t) + \frac{\sigma}{2\sigma_1} \int_0^t \int_{\hat{D}_0} \hat{g}_{,\tau}^2 \, dA d\tau.\end{aligned} \tag{A.5}$$

Choosing

$$\sigma = (2\hat{k})^{-1}, \qquad \sigma_1 = 2\hat{k},$$

we find

$$\int_0^t \int_{\hat{D}_0} u_{,3}^2 \, dA d\tau \leq 2 \int_0^t \int_{\hat{D}_0} \hat{g}_{,\alpha} \, \hat{g}_{,\alpha} \, dA d\tau + 4\hat{k} \hat{E}_1(0,t) + \frac{1}{4\hat{k}^2} \int_0^t \int_{\hat{D}_0} \hat{g}_{,\tau}^2 \, dA d\tau. \tag{A.6}$$

On inserting (A.6) back into (A.1), after using the arithmetic-geometric mean inequality on the right of (A.1), inequality (4.80) follows. Inequalities (4.81)–(4.83) are obtained analogously.

Acknowledgment. The authors would like to thank the referee for suggestions for improving the manuscript.

REFERENCES

[1] C. BANDLE, *Isoperimetric inequalities and applications*, Pitman Lecture Series 7, Pitman, Boston, MA, 1980.

[2] P. S. CROOKE AND L. E. PAYNE, *Continuous dependence on geometry for the backward heat equation*, Math. Methods Appl. Sci., 6 (1984), pp. 433–448.

[3] J. N. FLAVIN, R. J. KNOPS, AND L. E. PAYNE, *Decay estimates for the constrained elastic cylinder of variable cross section*, Quart. Appl. Math., 47 (1989), pp. 325–350.

[4] C. O. HORGAN AND L. E. PAYNE, *The influence of geometric perturbation on the decay of Saint Venant end effects in linear isotropic elasticity*, in Partial Differential Equations with Real Analysis, Pitman Research Notes in Mathematics 263, H. Begehr and A. Jeffrey, eds., Longman, 1992, pp. 187–218.

[5] ———, *The effect of constitutive law perturbation on finite anti-plane shear deformation of a semi-infinite strip*, Quart. Appl. Math., 51 (1993), pp. 441–465.

[6] C. O. HORGAN AND L. E. PAYNE, *Decay estimates for a class of nonlinear boundary value problems in two dimensions*, SIAM, J. Math. Anal., 20 (1989), pp. 782–785.

[7] 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.

[8] J. K. KNOWLES, *On the spatial decay of solutions of the heat equation*, J. Appl. Math. Phys., 22 (1971), pp. 1050–1056.

[9] L. E. PAYNE, *Some isoperimetric inequalities for harmonic functions*, SIAM J. Math. Anal., 1 (1970), pp. 354–359.

[10] V. G. SIGILLITO, *Explicit a priori inequalities with applications to boundary value problems*, Res. Notes Math. Ser., 13, Pitman, Boston, MA, 1977.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1259–1268, September 1994

002

ELLIPTIC EQUATIONS IN DIVERGENCE FORM, GEOMETRIC CRITICAL POINTS OF SOLUTIONS, AND STEKLOFF EIGENFUNCTIONS *

G. ALESSANDRINI† AND R. MAGNANINI‡

Abstract. The Stekloff eigenvalue problem (1.1) has a countable number of eigenvalues $\{p_n\}_{n=1,2,\ldots}$, each of finite multiplicity. In this paper the authors give an upper estimate, in terms of the integer n, of the multiplicity of p_n, and the number of critical points and of nodal domains of the eigenfunctions corresponding to p_n.

In view of a possible application to inverse conductivity problems, the result for the general case of elliptic equations with discontinuous coefficients in divergence form is proven by replacing the classical concept of critical point with the more suitable notion of geometric critical point.

Key words. eigenvalue problems, geometric properties of elliptic equations, critical points, inverse conductivity problems

AMS subject classifications. 35J25, 35P99, 35C62, 30C15

1. Introduction.

In this paper we are concerned with weak solutions of the elliptic equation in divergence form:

$$\operatorname{div}(A\nabla u) = 0 \quad \text{in } \Omega, \tag{1.1a}$$

and especially with Stekloff eigenfunctions, that is, those nontrivial solutions that, for some constant p, the Stekloff eigenvalue, satisfy in the weak sense the boundary condition

$$A\nabla u \cdot \nu = pu \quad \text{on } \partial\Omega. \tag{1.1b}$$

Here Ω is a simply connected bounded domain in the plane, with Lipschitz boundary $\partial\Omega$, ν is the exterior unit normal to $\partial\Omega$, and $A = (a_{ij})$ is a 2×2 symmetric matrix of $L^\infty(\Omega)$ coefficients satisfying, for some constant $\lambda, 0 < \lambda \leq 1$, the uniform ellipticity condition

$$\lambda |\xi|^2 \leq \sum_{i,j=1}^{2} a_{ij}(z)\,\xi_i\xi_j \leq \lambda^{-1}|\xi|^2, \quad \text{for every } z \in \Omega, \xi \in \mathbb{R}^2. \tag{1.2}$$

Here and in what follows, we use the complex coordinate $z = x + iy$ in the plane.

The study of this eigenvalue problem was started by Stekloff [St] in 1902. In §3, we recall the definitions of Stekloff eigenvalues and eigenfunctions; a review of their known properties can be found in Bandle [B]. Research on this subject has been mainly devoted to estimates on eigenvalues (see, for instance, [H-P-S] and also [B] for further references). Let us mention in passing that, in connection with applied problems in fluid mechanics, mixed type problems also have been considered (see [F-K]). Typically,

* Received by the editors February 2, 1993; accepted for publication (in revised form) June 18, 1993.

† Dipartimento di Scienze Matematiche, Università di Trieste, Italy.

‡ Dipartimento di Matematica, Università di Firenze, Italy.

in such problems, we assume that (1.1b) holds only on one portion of $\partial\Omega$, whereas on the rest of $\partial\Omega$, $A\nabla u \cdot \nu = 0$ is assumed.

Our main results, which are summarized in Theorem 3.2, consist of estimates on the multiplicities of Stekloff eigenvalues and the numbers of nodal domains and of critical points of Stekloff eigenfunctions.

One up-to-date motivation of our study of the Stekloff eigenvalue problem comes from the so-called inverse conductivity problem: suppose that the coefficient matrix A (*the conductivity*) is unknown; we wish to determine it, or, more generally, to recover partial information about it from the knowledge of the so-called Dirichlet to Neumann map Λ_A. Here, $\Lambda_A : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ is defined as the operator that maps the Dirichlet data $u_{|\partial\Omega}$ for (1.1a) into the corresponding Neumann data $A\nabla u \cdot \nu_{|\partial\Omega}$ (see, for instance, [Sy-U], [Sy]). Then it is evident that the Stekloff eigenvalues and the boundary traces of the Stekloff eigenfunctions are exactly the eigenvalues and eigenfunctions of Λ_A. Such traces and eigenvalues can be approximately measured by experiments and could be effectively used as the data of the inversion procedure. For a discussion of a related spectral approach to the inverse conductivity problem, see Gisser, Isaacson, and Newell [G-I-N]. A deeper understanding of the geometrical features of Stekloff eigenfunctions would then be helpful in assessing uniqueness and continuous dependence questions for the inverse conductivity problem and perhaps also in the design of reconstruction algorithms.

In addition to the possible applications to inverse problems, the authors have been inspired by the work of Payne and Philippin [P-P] on questions of symmetry related to the Stekloff eigenvalue problem (see also [A-M2]). In this respect, we mention that some of our results, when restricted to Laplace's equation and to the second Stekloff eigenvalue, have already been presented in [P-P].

The flavor of our results is similar to those of Cheng [C] for the eigenfunctions of the Laplacian on surfaces; however, the methods in the proofs and the specific results are different because of the intrinsic differences between the two eigenvalue problems.

To mention the most apparent peculiarity of Stekloff eigenfunctions, observe that, by the maximum principle, every solution of (1.1a), and thus every Stekloff eigenfunction, can have neither interior maxima nor minima, and each of its level sets must reach the boundary. Due to this observation, the geometric-topologic properties of level lines and level sets of Stekloff eigenfunctions are quite different from those of the vibrating membrane problems. In fact, in some respects, the study of such properties is perhaps easier for equations like (1.1a), for which we have theorems that permit us to estimate the number of interior critical points in terms of the boundary data [A], [A-M1]. Theorems like this will be our basic tool, along with simple variations of the fundamental Courant's nodal domain theorem.

Still, in view of the application to the inverse conductivity problem, we have chosen to treat the Stekloff eigenvalue problem when no smoothness assumption is imposed on the coefficients in (1.1a). In fact, there are practical cases when the conductivity coefficients are discontinuous and we are especially interested in determining the discontinuities (see, for instance, [B-F-I], [P], [Su-U]). This generality causes additional technical difficulties: solutions of (1.1a) need not to be differentiable in the classical sense and thus the notion of critical point must be adapted to this nonsmooth setting. For this reason, we introduce the concept of *geometric critical point* (see Definition 2.3). Roughly speaking, a point will be called a geometric critical point for a solution u of (1.1a) if it is a critical point with respect to an appropriate (possibly nonsmooth) change of variables that makes u become smooth. We also give the definition of *geometric index*, which generalizes the notion of multiplicity of a critical

point (see Definition 2.4).

These definitions are based on the crucial remark that the unique continuation property and a representation theorem also hold for solutions of (1.1a). We have not been able to find in the literature theorems of such a kind for equations like (1.1a), whereas they are well known for equations with smoother coefficients (see [Sch]) and for equations not in divergence form (see [B-N]). Although the method of proof of such results may sound familiar to the experts in quasi-conformal mappings and complex analytic methods, we include our own proofs of the unique continuation property and of representation formulas for solutions of (1.1a) (Theorem 2.1 and Corollary 2.2). We trust that these results might be of some independent interest and that they can find useful applications, especially in the field of inverse problems.

We conclude this introduction by pointing out the following consequence of such results. In [A], [A-M1], when the coefficients in (1.1a) are smooth, an estimate on the maximum number of interior critical points of a solution $u \in C^1(\bar{\Omega}) \cap C^2(\Omega)$ of (1.1a) was given in terms of the number of times some boundary data of u changes its sign on $\partial\Omega$ (see [A, Thm. 1.1], [A-M1, Thms. 2.1, 2.2], for details). Theorems 2.7 and 2.8 in this paper generalize the above results to equation (1.1a) with nonsmooth coefficients.

2. Geometric critical points. Throughout this section, $B_R(0)$ will denote the disk with radius R centered at $z = 0$.

THEOREM 2.1 (representation formula). *Let $u \in W^{1,2}(\Omega)$ be a nonconstant solution of* (1.1a).

There exists a quasi-conformal mapping $\chi : \Omega \to B_1(0)$ and a real-valued harmonic function h on $B_1(0)$ such that

$$u = h \circ \chi \quad \text{in } \Omega. \tag{2.1}$$

The dilatation coefficient $\chi_{\bar{z}}/\chi_z$ of χ is bounded by the constant $(1-\lambda)/(1+\lambda)$.

Proof. By (1.1a), the 1-form $\omega = -(a_{12}u_x + a_{22}u_y)dx + (a_{11}u_x + a_{12}u_y)dy$ is closed in Ω. Therefore, we can find $v \in W^{1,2}(\Omega)$ such that $dv = \omega$. The function v will be called the *stream function* associated with u, in analogy with the theory of gas dynamics (see, for instance, [B-S]).

In other words, u and v satisfy the following elliptic system:

$$\nabla v = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} A \nabla u, \tag{2.2}$$

almost everywhere in Ω. Note that (2.2) is just a special case of the elliptic systems studied by Bers and Nirenberg in [B-N].

By setting $f = u + iv$ and using the standard notation for complex derivatives, (2.2) takes the form

$$f_{\bar{z}} = \mu f_z + \nu \overline{f_z}, \tag{2.2'}$$

where

$$\mu = \frac{a_{22} - a_{11} - 2ia_{12}}{1 + a_{11} + a_{22} + a_{11}a_{22} - a_{12}^2}, \qquad \nu = \frac{1 - a_{11}a_{22} + a_{12}^2}{1 + a_{11} + a_{22} + a_{11}a_{22} - a_{12}^2},$$

and the following estimate can be easily deduced from (1.2)

$$|\mu| + |\nu| \leq \frac{1-\lambda}{1+\lambda} < 1 \quad \text{in } \Omega.$$

Consequently, since $f \in W^{1,2}(\Omega, \mathbb{C})$, then it is a quasi-regular mapping with dilatation bounded by $(1-\lambda)/(1+\lambda)$. Since f is nonconstant, it can be factored as

$$f = F \circ \chi \quad \text{in } \Omega, \tag{2.3}$$

where F is a holomorphic function on the disk $B_1(0)$ and $\chi : \Omega \to B_1(0)$ is a quasi-conformal homeomorphism (see [L-V]). Finally,

$$\left|\frac{\chi_{\bar{z}}}{\chi_z}\right| = \left|\frac{f_{\bar{z}}}{f_z}\right| \leq \frac{1-\lambda}{1+\lambda}.$$

COROLLARY 2.2 (Unique continuation principle). *If there exists $z_0 \in \Omega$ and positive constants $C_1, C_2, \dots, C_N, \cdots$ such that*

$$\int_{\Omega \cap B_R(z_0)} |\nabla u|^2 \, dx\, dy \leq C_N R^N \quad \forall R > 0, \quad \forall N = 1, 2, \dots, \tag{2.4}$$

then u is constant in Ω.

Proof. Let us suppose by contradiction that u is nonconstant. Without loss of generality, we may set $z_0 = 0$ and $u(0) = v(0) = 0$.

Note that, by (2.2), the stream function v associated with u satisfies the following equation:

$$\operatorname{div}\left(\frac{1}{\det A} A \nabla v\right) = 0 \quad \text{in } \Omega, \tag{2.5}$$

in the weak sense. Such an equation satisfies the ellipticity condition (1.2) as well.

The local boundedness estimate [G-T, Thm. 8.17] is applicable to both (1.1a) and (2.5). Thus, by this estimate, the Poincaré inequality, and (2.4), we have

$$\begin{aligned} &\max_{B_{R/2}(0)} |u - u_R|^2 \leq K C_N R^N, \\ &\max_{B_{R/2}(0)} |v - v_R|^2 \leq K C_N R^N, \end{aligned} \qquad \forall N = 1, 2, \cdots \quad \text{and} \quad \forall R,\ 0 < R < R_0.$$

Here $R_0 = \operatorname{dist}(0, \partial\Omega)$, K is a positive constant depending on λ only, and u_R, v_R denote the mean values on the disk $B_R(0)$ of u, v, respectively. Since $u(0) = v(0) = 0$, we also have $u_R^2, v_R^2 \leq K C_N R^N$, and hence

$$\max_{B_{R/2}(0)} (u^2 + v^2) \leq 8 K C_N R^N, \quad \forall R,\ 0 < R < R_0, \quad \forall N = 1, 2, \dots. \tag{2.6}$$

Now, we claim that there exist $\rho, 0 < \rho < R_0$, a quasi-conformal homeomorphism $\tilde{\chi}$ of $B_\rho(0)$ in itself, and a positive integer M such that the quasi-regular mapping $f = u + iv$ can be factored as

$$f(z) = [\tilde{\chi}(z)]^M, \qquad |z| < \rho, \tag{2.7}$$

where $\tilde{\chi}$ has dilatation bounded by $(1-\lambda)/(1+\lambda)$ and $\tilde{\chi}(0) = 0$. This factorization is readily obtained from (2.3), first, by choosing M as the order of the first nontrivial term in the Taylor series for $F - F(\chi(0))$ at $\chi(0)$, and, second, by noticing the local invertibility of the branches of the multivalued function $[F - F(\chi(0))]^{1/M}$.

Since $\tilde{\chi}$ is quasi-conformal, we have that $\tilde{\chi}^{-1}$ is Hölder continuous at zero with some exponent $0 < \alpha \leq 1$. From (2.6), (2.7), we have that

$$Q|z|^{2M/\alpha} \leq |f(z)|^2 \leq KC_N|z|^N, \quad \forall z, |z| < \rho, \quad \forall N = 1, 2, \ldots,$$

for some positive constant Q, and this is impossible.

DEFINITION 2.3. *We say that $z_0 \in \Omega$ is a* geometric critical point *for u, if we have $\nabla h(\chi(z_0)) = 0$, where h and χ are, respectively, the harmonic function and the quasi-conformal mapping appearing in* (2.1).

Remark. It is an obvious, but essential, consequence of this definition that geometric critical points of nonconstant solutions of (1.1a) are isolated.

We now recall a classical definition of the index of a smooth function (see [Mi]). For a C^1 function h with isolated critical points in the disk $B_1(0)$ and a subdomain $D \subset\subset B_1(0)$ such that ∂D is smooth and contains no critical point of h, the index of h in D is

$$I(D,h) = -\frac{1}{2\pi}\int_{\partial D} d\arg(\nabla h).$$

With such a choice of the sign, if h is harmonic, $I(D,h)$ gives the number of critical points of h in D, when counted according to their multiplicities. Moreover, $I(D,h)$ is constant under perturbations of D that contain the same critical points, and its definition can be extended to the case when ∂D is nonsmooth.

We generalize this notion to nonconstant solutions of (1.1a).

DEFINITION 2.4. *Let $D \subset\subset \Omega$ be an open set. If u has no geometric critical points on ∂D, we define the* geometric index of u in D *as*

$$I(D,u) = I(\chi(D), h),$$

where h and χ are as in Theorem 2.1.

Moreover, we define the geometric index of u *at $z_0 \in \Omega$ as*

$$I(z_0,u) = \lim_{r\to 0} I(B_r(z_0), u).$$

Such a limit exists, since the geometric critical points of a solution of (1.1a) *are isolated.*

The next lemma gives a sort of justification for the term "geometric" in the previous definitions and shows that these do not depend on the particular choice of the representation (2.1).

LEMMA 2.5. *Let u be a nonconstant solution of* (1.1a). *If $z_0 \in \Omega$ is a geometric critical point for u with geometric index $I = I(z_0,u)$, then there exists a neighborhood $U \subset \Omega$ of z_0 such that the level line $\{z \in U : u(z) = u(z_0)\}$ is made of $I+1$ simple arcs, whose pairwise intersection consists of $\{z_0\}$ only.*

Proof. By the representation (2.1), since χ is a quasi-conformal homeomorphism, it is enough to look at the level line $\{\zeta \in B_1(0) : h = h(\chi(z_0))\}$ near $\chi(z_0)$.

Since $I(\chi(z_0), h) = I(z_0,u) = I$, then $h - h(\chi(z_0))$ is asymptotic to a homogeneous harmonic polynomial of degree $I+1$ near $\chi(z_0)$.

Remark. Observe that, if u is C^1 in a neighborhood of z_0 (which happens, for instance, when A is Hölder continuous; see [Sch]), then z_0 is a geometric critical point with geometric index I if and only if $\nabla u(z_0) = 0$ with standard index I. This is a consequence of Lemma 2.5 above and Lemma 3.1 in [A-M1]. Note that in [A-M1] the opposite sign is chosen in the definition of the index.

PROPOSITION 2.6 (Continuity of the geometric index). *Let $\{A_m\}_{m=1,2,\dots}$ be a sequence of symmetric matrices with $L^\infty(\Omega)$ entries satisfying* (1.2). *Let $u_m \in W^{1,2}(\Omega)$ be weak solutions of*

$$\operatorname{div}(A_m \nabla u_m) = 0 \quad \text{in } \Omega,$$

which converge to u in $W^{1,2}_{\text{loc}}(\Omega)$.

If $D \subset\subset \Omega$ is such that u has no geometric critical point on ∂D, then we have

$$\lim_{m\to\infty} I(D, u_m) = I(D, u). \tag{2.8}$$

Proof. By the proof of Theorem 2.1, for each u_m we may construct a stream function v_m such that $f_m = u_m + iv_m$ are quasi-regular mappings with dilatation coefficients uniformly bounded by $(1-\lambda)/(1+\lambda)$. By (2.1), we also have $u_m = h_m \circ \chi_m$, and the dilatation coefficients of the quasi-conformal mappings χ_m are also uniformly bounded. Since $u_m \to u$ in $W^{1,2}_{\text{loc}}(\Omega)$, by using the uniform interior bounds for f_m and χ_m in C^α (see [G-T, Thm. 8.24]), we have that, possibly passing to subsequences, h_m and χ_m converge, respectively, in $C^\infty_{\text{loc}}(B_1(0))$ and $C^0_{\text{loc}}(\Omega) \cap W^{1,2}_{\text{loc}}(\Omega)$, to the functions h and χ corresponding to u in the representation (2.1).

By definition, $I(D,u) = I(\chi(D), h)$ and $I(D, u_m) = I(\chi_m(D), h_m)$. Furthermore, by our hypothesis, ∇h does not vanish on $\partial\chi(D)$, so that $|\nabla h_m|$ is uniformly bounded away from zero on $\partial\chi(D)$, for m large enough. Thus,

$$I(\chi(D), h) = \lim_{m\to\infty} I(\chi(D), h_m),$$

and hence, by the $C^0_{\text{loc}}(\Omega)$ convergence of χ_m, we arrive at (2.8). Observe now that, since the very beginning of our argument, we could have replaced the sequence $\{u_m\}$ with any of its subsequences. Therefore, the limit in (2.8) exists and the stated equality holds.

THEOREM 2.7. *Let $g \in H^{1/2}(\partial\Omega)$ be of bounded variation on $\partial\Omega$ and such that $\partial\Omega$ can be split into $2M$ arcs on which alternatively g is a nondecreasing and nonincreasing function of the arclength parameter.*

Let $u \in W^{1,2}(\Omega)$ be the unique solution of (1.1a) *satisfying the Dirichlet condition $u = g$ on $\partial\Omega$.*

Then the geometric critical points of u in Ω, when counted according to their indices, are at most $M-1$.

Proof. In view of Lemma 2.5 above, this is just a rephrasing of Theorem 1.1 in [A1]. We omit the details.

THEOREM 2.8. *Let $g \in H^{-1/2}(\partial\Omega)$ be such that $\partial\Omega$ can be split into $2M$ closed arcs $\Gamma_1, \dots, \Gamma_{2M}$ such that $(-1)^j g \geq 0$ on Γ_j, $j = 1, \dots, 2M$, in the sense of distributions.*

Let $u \in W^{1,2}(\Omega)$ be a solution of (1.1a) *satisfying the Neumann condition $A\nabla u \cdot \nu = g$ on $\partial\Omega$.*

Then, the geometric critical points of u in Ω, when counted according to their indices, are at most $M-1$.

Proof. We may suppose that $\partial\Omega$ is C^∞. If it were not so, we could construct a Lipschitz mapping that transforms Ω into a disk. Such a mapping does not alter the nature of the equation nor the sign conditions on the Neumann data g.

Let us choose a sequence $\{A_m\}$ of $C^\infty(\bar\Omega)$ symmetric matrices satisfying (1.2) and such that $A_m \to A$ in $L^p(\Omega)$, for some $1 \leq p < \infty$. It is a straightforward exercise now to construct a sequence $\{g_m\} \subset C^\infty(\partial\Omega)$ converging to g in $H^{-1/2}(\partial\Omega)$ and such

that $\int_{\partial\Omega} g_m \, ds = 0$ and $(-1)^j g_m > 0$ in the interior of each $\Gamma_j, j = 1, \dots, 2M$, for all $m = 1, 2, \dots$.

For any integer m, let $u_m \in C^\infty(\bar{\Omega})$ be the unique solution of the following problem:

$$\operatorname{div}(A_m \nabla u_m) = 0 \quad \text{in } \Omega, \qquad A_m \nabla u_m \cdot \nu = g_m \quad \text{on } \partial\Omega,$$

such that $\int_\Omega u_m \, dx \, dy = \int_\Omega u \, dx \, dy$. Since u_m is smooth on $\partial\Omega$, we can apply Theorem 2.2 in [A-M1] and obtain $I(D, u_m) \leq M - 1$ for every integer m.

Moreover, we can easily see that, by possibly passing to subsequences, $u_m \to u$ in $W^{1,2}_{\text{loc}}(\Omega)$, and hence Proposition 2.6 is applicable. Therefore, for any $D \subset\subset \Omega$ such that ∂D does not contain any geometric critical point of u, we have $I(D, u) = \lim_{m\to\infty} I(D, u_m)$, and hence $I(D, u) \leq M - 1$. By the arbitrariness of D in Ω, we obtain our thesis.

3. Multiplicity of Stekloff eigenvalues and geometric critical points of Stekloff eigenfunctions. As is well known, by observing that the trace imbedding $W^{1,2}(\Omega) \hookrightarrow L^2(\partial\Omega)$ is compact, the Stekloff eigenfunctions and eigenvalues in $W^{1,2}(\Omega)$ are characterized as the critical points and critical values of the Rayleigh quotient

$$R(u) = \frac{\int_\Omega A\nabla u \cdot \nabla u \, dx \, dy}{\int_{\partial\Omega} u^2 \, ds}; \tag{3.1}$$

here ds denotes the arclength element on $\partial\Omega$ (see [St]). The nth Stekloff eigenvalue p_n can be recursively defined as the minimum of the quotient (3.1) over all functions of class $W^{1,2}(\Omega)$ that are orthogonal in $L^2(\partial\Omega)$ to the subspaces $V_k, k = 1, \dots, n-1$, where

$$V_k = \{u \in W^{1,2}(\Omega) : u \text{ is a weak solution of (1.1) with } p = p_k\}. \tag{3.2}$$

In this way, we can form a divergent sequence $0 = p_1 < p_2 < \cdots < p_n < \cdots$ of eigenvalues, each of them of finite multiplicity.

DEFINITION 3.1. *Let p_n be the nth Stekloff eigenvalue; we denote by μ_n its multiplicity, that is,*

$$\mu_n = \dim V_n. \tag{3.3}$$

For $n \geq 2$, we will also set

$$\kappa_n = \max_{u \in V_n \setminus \{0\}} \#\left\{ \begin{array}{c} \text{geometric critical points of } u \text{ in } \Omega, \\ \text{counted according to their index} \end{array} \right\}. \tag{3.4}$$

A nodal domain Ω_k *of* $u \in V_n$ *is a connected component of the set* $\{z \in \Omega : u(z) \neq 0\}$, *while a connected component of a set* $\partial\Omega_k \cap \partial\Omega$ *will be referred to as a* boundary nodal domain of u.

We then define

$$\Delta_n = \max_{u \in V_n \setminus \{0\}} \#\{\text{nodal domains of } u\}, \tag{3.5}$$

$$\delta_n = \max_{u \in V_n \setminus \{0\}} \#\{\text{boundary nodal domains of } u\}. \tag{3.6}$$

THEOREM 3.2. *The following inequalities hold*:

$$\Delta_{n+1} \leq 1 + \sum_{k=1}^{n} \mu_k, \qquad n = 1, 2, \ldots, \tag{3.7}$$

$$\kappa_n \leq \Delta_n - 2, \qquad n = 2, 3, \ldots, \tag{3.8}$$

$$\mu_n \leq 2(\kappa_n + 1), \qquad n = 2, 3, \ldots. \tag{3.9}$$

COROLLARY 3.3. *For every integer* $n \geq 2$, *we have*

$$\mu_n \leq 2 \cdot 3^{n-2}, \tag{3.10}$$

$$\kappa_n \leq 3^{n-2} - 1, \tag{3.11}$$

$$\Delta_n \leq 3^{n-2} + 1. \tag{3.12}$$

Proof. By applying (3.7)–(3.9) we obtain the recurrence relation $\mu_{n+1} \leq 2 \times \sum_{k=1}^{n} \mu_k, n = 1, 2, \ldots$. Since $\mu_1 = 1$, we obtain (3.10); (3.11) and (3.12) then easily follow from (3.8) and (3.7).

Remark. When $n = 2$, (3.11) gives $\kappa_2 = 0$. This provides a different proof of Lemma 3 in [P-P].

The proof of Theorem 3.2 requires the following lemma, which will be proved at the end of this section.

LEMMA 3.4. *Let* Δ_n *and* δ_n *be defined as in Definition* 3.1. *Then*

$$\delta_n \leq 2(\Delta_n - 1), \qquad n = 2, 3, \ldots. \tag{3.13}$$

Proof of Theorem 3.2. *Step* 1. We prove (3.7) by contradiction. This argument has been used already in [K-S] for the case of the Laplace operator. Suppose there exists a nontrivial eigenfunction $u \in V_{n+1}$ with Δ nodal domains $\Omega_1, \Omega_2, \ldots, \Omega_\Delta$, and $\Delta \geq 2 + \sum_{k=1}^{n} \mu_k$. Let us denote by $u_1^{(k)}, \ldots, u_{\mu_k}^{(k)}$ a basis of the vector space $V_k, 1 \leq k \leq n$.

Now consider the function $v = \sum_{j=1}^{\Delta-1} \alpha_j (u \mathbf{1}_{\Omega_j})$; here $\mathbf{1}_{\Omega_j}$ denotes the characteristic function of the set Ω_j. The real numbers $\alpha_1, \ldots, \alpha_{\Delta-1}$ can be chosen not all zero and such that

$$\int_{\partial\Omega} v u_\ell^{(k)} \, ds = 0, \quad \text{for all } \ell = 1, \ldots, \mu_k, \qquad k = 1, 2, \ldots, n; \tag{3.14}$$

in fact (3.14) provides $\sum_{k=1}^{n} \mu_k \leq \Delta - 2$ linear homogeneous conditions on $\Delta - 1$ parameters.

In view of (3.14), the function v is admissible for the variational characterization (3.1) of p_{n+1}. From the definition of v, we have

$$\begin{aligned} \int_{\Omega_j} A \nabla v \cdot \nabla v \, dx \, dy &= \int_{\partial\Omega_j} (A \nabla v \cdot \nu) \, v \, ds \\ &= \alpha_j^2 \int_{\partial\Omega_j} (A \nabla u \cdot \nu) \, u \, ds, \qquad j = 1, \ldots, \Delta - 1. \end{aligned}$$

Hence, (1.1b) implies

$$\int_{\Omega_j} A \nabla v \cdot \nabla v \, dx \, dy = p_{n+1} \alpha_j^2 \int_{\partial\Omega_j} u^2 \, ds, \qquad j = 1, \ldots, \Delta - 1$$

and, by adding the above $\Delta - 1$ relations, we obtain

$$\int_\Omega A\nabla v \cdot \nabla v \, dx\, dy = p_{n+1} \int_{\partial\Omega} v^2 \, ds.$$

Therefore, v is a nontrivial Stekloff eigenfunction corresponding to the eigenvalue p_{n+1}. Now, since $v \equiv 0$ on Ω_Δ, by the unique continuation property, we have $v \equiv 0$ on Ω, which is a contradiction.

Step 2. Let $u_n \in V_n$; by (1.1b) and by Lemma 3.4, $A\nabla u_n \cdot \nu$ satisfies the hypotheses of Theorem 2.7 with $M \leq \Delta_n - 1$; thus, (3.8) follows easily.

Step 3. By contradiction, suppose $\mu_n \geq 2(\kappa_n + 1) + 1$.

Let $u^{(j)}, j = 1, \ldots, \mu_n$, be a basis of V_n and fix $\kappa_n + 1$ distinct points $z_1, \ldots, z_{\kappa_n+1}$ in Ω. As we did in the proof of Theorem 2.8, we approximate A by a sequence of matrices $\{A_m\}_{m=1,2,\cdots}$ with $C^\infty(\Omega)$-coefficients satisfying (1.2). For each $j, 1 \leq j \leq \mu_n$, the weak solutions $u_m^{(j)}$ of the Dirichlet problem

$$\operatorname{div}\left(A_m \nabla u_m^{(j)}\right) = 0 \quad \text{in } \Omega, \qquad u_m^{(j)} - u^{(j)} \in W_0^{1,2}(\Omega),$$

are $C^\infty(\Omega)$-functions and form a sequence $u_m^{(j)}$ that converges to $u^{(j)}$ in $W^{1,2}_{\text{loc}}(\Omega)$. By our hypothesis on μ_n, for each $j, 1 \leq j \leq \mu_n$, we can find real numbers $\alpha_m^{(j)}$ such that $\sum_{j=1}^{\mu_n} (\alpha_m^{(j)})^2 = 1, m = 1, 2, \cdots$ and, also

$$\sum_{j=1}^{\mu_n} \alpha_m^{(j)} \nabla u_m^{(j)}(z_\ell) = 0, \quad \text{for all } \ell = 1, \ldots, \kappa_n + 1, \qquad m = 1, 2, \ldots. \tag{3.15}$$

For each $j = 1, \ldots, \mu_n$, the sequence $\alpha_m^{(j)}$ can be chosen to converge to some number $\alpha^{(j)}$ so that we have

$$\sum_{j=1}^{\mu_n} \left(\alpha^{(j)}\right)^2 = 1.$$

Now, let D be an open set with $\bar{D} \subset \Omega, z_1, \ldots, z_{\kappa_n+1} \in D$, and such that ∂D does not contain any geometric critical point of the function $v = \sum_{j=1}^{\mu_n} \alpha^{(j)} u^{(j)}$. The sequence of functions $v_m = \sum_{j=1}^{\mu_m} \alpha_m^{(j)} u_m^{(j)}$ is such that $I(v_m, D) \geq \kappa_n + 1$, by (3.15). Moreover, by possibly passing to a subsequence, $v_m \to v$ in $W^{1,2}_{\text{loc}}(\Omega)$, as $m \to \infty$, so that Proposition 2.6 implies that $I(v, D) \geq \kappa_n + 1$, that is, v is a nontrivial eigenfunction in V_n with at least $\kappa_n + 1$ geometric critical points in $D \subset \Omega$. This is a contradiction.

We conclude by giving a sketch of the proof of Lemma 3.4. To this end, we introduce the following definitions.

DEFINITION 3.5. *We say that a simply connected open subset A of Ω is a* cap, *if $\partial\Omega \cap \partial A$ is connected and nonempty.*

Let $\Omega_1, \ldots, \Omega_K$ be open subsets of Ω; we say that $\{\Omega_k\}_{k=1,\ldots,K}$ is an *admissible covering of* Ω, if $\Omega_1, \ldots, \Omega_K$ are pairwise disjoint, $\Omega \subset \bigcup_{k=1}^K \bar{\Omega}_k$, and $\partial\Omega_k \cap \partial\Omega \neq \emptyset$, for every $k = 1, \ldots, K$.

Proof of Lemma 3.4 (Sketch). Let $u \in V_n, u$ nontrivial, and let $\Omega_1, \ldots, \Omega_K$ be the nodal domains of u in Ω; these sets form an admissible covering of Ω.

Now, let N_Ω be the number of boundary nodal domains of u in $\partial\Omega$. Then (3.13) is implied by $N_\Omega \leq 2(K - 1)$. This inequality is readily proved by induction on the

number K of nodal domains and by using the following facts, the proofs of which are straightforward:

(i) Every nodal domain Ω_k is simply connected.

(ii) The covering $\{\Omega_k\}_{k=1,\dots,K}$ contains at least one cap.

(iii) If Ω_K is a cap, then $\tilde{\Omega} = \Omega \setminus \bar{\Omega}_K$ is a simply connected open set and $\{\Omega_k\}_{k=1,\dots,K-1}$ is an admissible covering of $\tilde{\Omega}$. Moreover, $N_\Omega \leq N_{\tilde{\Omega}} + 2$.

REFERENCES

[A] G. ALESSANDRINI, *Critical points of solutions of elliptic equations in two variables*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 14 (1987), pp. 229–256.

[A-M1] G. ALESSANDRINI AND R. MAGNANINI, *The index of isolated critical points and solutions of elliptic equations in the plane*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 567–589.

[A-M2] ———, *Symmetry and non-symmetry for the overdetermined Stekloff eigenvalue problem*, Z. Angew. Math. Phys., 45 (1994), pp. 44–52.

[B] C. BANDLE, *Isoperimetric Inequalities and Applications*, Pitman, London, 1980.

[B-F-I] H. BELLOUT, A. FRIEDMAN, AND V. ISAKOV, *Stability for an inverse problem in potential theory*, Trans. Amer. Math. Soc., 332 (1992), pp. 271–296.

[B-S] S. BERGMAN AND M. SCHIFFER, *Kernel Functions and Differential Equations in Mathematical Physics*, Academic Press, New York, 1953.

[B-N] L. BERS AND L. NIRENBERG, *On a representation theorem for linear elliptic systems with discontinuous coefficients and its applications*, in Convegno Internazionale sulle Equazioni Lineari alle Derivate Parziali, Cremonese, Roma 1955.

[C] S-Y. CHENG, *Eigenfunctions and nodal sets*, Comment. Math. Helv., 51 (1976), pp. 43–55.

[F-K] D. W. FOX AND J. R. KUTTLER, *Sloshing frequencies*, Z. Angew. Math. Phys., 34 (1983), pp. 668–696.

[G-I-N] D. G. GISSER, E. ISAACSON, AND J. C. NEWELL, *Electric current computed tomography and eigenvalues*, SIAM J. Appl. Math., 50 (1990), pp. 1623–1634.

[G-T] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[H-P-S] J. HERSCH, L. E. PAYNE, AND M. M. SCHIFFER, *Some inequalities for Stekloff eigenvalues*, Arch. Rational Mech. Anal., 57 (1974–75), pp. 99–114.

[I] V. ISAKOV, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.

[K-S] J. R. KUTTLER AND V. G. SIGILLITO, *An inequality for a Stekloff eigenvalue by the method of defect*, Proc. Amer. Math. Soc., 20 (1969), pp. 357–360.

[L-V] O. LEHTO AND K. VIRTANEN, *Quasiconformal Mappings in the Plane*, Springer-Verlag, Berlin, 1973.

[Mi] J. MILNOR, *Differential Topology*, Princeton University Press, Princeton, NJ, 1958.

[P-P] L. E. PAYNE AND G. A. PHILIPPIN, *Some overdetermined boundary value problems for harmonic functions*, Z. Angew. Math. Phys., 42 (1991), pp. 864–873.

[P] J. POWELL, *On a small perturbation in the two dimensional inverse conductivity problem*, J. Math. Anal. Appl., to appear.

[Sch] F. SCHULZ, *Regularity Theory for Quasilinear Elliptic Systems and Monge-Ampère Equations in Two Dimensions*, Springer-Verlag, New York, 1990.

[St] M. W. STEKLOFF, *Sur les problèmes fondamentaux en physique mathématique*, Ann. Sci. Ecole Norm. Sup., 19 (1902), pp. 455–490.

[Su-U] Z. SUN AND G. UHLMANN, *Recovery of singularities for formally determined inverse problems*, Comm. Math. Phys., 153 (1993), pp. 431–451.

[Sy] J. SYLVESTER, *An anisotropic inverse boundary value problem*, Comm. Pure Appl. Math., 43 (1990), pp. 201–233.

[Sy-U] J. SYLVESTER AND G. UHLMANN, *A uniqueness theorem for an inverse boundary value problem*, Ann. of Math., 125 (1987), pp. 153–169.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1269–1302, September 1994

003

APPROXIMATION OF ATTRACTORS BY ALGEBRAIC OR ANALYTIC SETS*

C. FOIAS† AND R. TEMAM†‡

Abstract. In this work the authors introduce a method for approximating the global attractor of dissipative differential equations (including the two-dimensional Navier–Stokes equations) based on the time analyticity (in a fixed infinite band) of all solutions lying on the attractor. In particular, three families of polynomial maps are constructed with the property that, under some suitable conditions, their zeros approximate the attractor.

Key words. attractor, absorbing sets, hypergeometric function, polynomial map

AMS subject classifications. 34D45, 35Q30, 47H99, 41A05, 33C05

Introduction. The long-term behavior of the solutions of dissipative evolution equations is described by a compact attractor that attracts all bounded sets. At our present level of understanding of dynamical systems, little is known about the geometry of these attractors which may be fractal sets [10], [16], [21]–[23]. We know, thanks to general theorems which apply to diverse dissipative systems, that the attractor has a finite dimension in the sense of the Hausdorff dimension or of capacity (fractal dimension); see, for instance, [15], [17], [7], [2], [25].

Our object in this article is to present a general theory for the approximation of these possibly fractal attractors by smooth sets, more precisely by algebraic or analytic sets. Furthermore, the approximating sets that we construct can be defined by simple explicit relations using the equation itself. Other approximations of attractors by finite-dimensional algebraic or analytic sets were constructed elsewhere [26], but the corresponding sets were graphs above a finite-dimensional space. This restriction is removed here allowing, we believe, for more flexibility and more structure. Besides the intrinsic independent interest of the results presented here we hope that this new tool can help us understand the geometry of attractors. Our methods apply to both ordinary differential equations (i.e., finite-dimensional case) and partial differential equations (i.e., infinite-dimensional case). However, they are better suited for the latter, because of the existence of large eigenvalues of the linear part of the equations. Our approach is as follows. Each point on the attractor belongs to a complete orbit defined for all real times. Also due to the time analyticity of solutions these orbits are defined for complex time in a strip around the real line. By utilizing the Taylor series expansion of the solution around the origin and the relations obtained by repeated differentiations of the equation, we derive the equations of the approximate manifolds. Three different methods of approximations are proposed. They yield algebraic or analytic sets that can approximate the attractor at an arbitrarily high level of accuracy. The first method is not the most efficient, but it is a simple one for which it is easy

*Received by the editors December 2, 1991; accepted for publication (in revised form) May 25, 1993. This work was supported in part by Department of Energy grants DOE DE-FG02-86ER25020 and DOE DE-FG02-92ER25120, National Science Foundation grants NSF DMS-8802596 and NSF DMS-9007802, Office of Naval Research grant NAVY N00014-91-J-1140, and by the Research Fund of Indiana University.

†Department of Mathematics and the Institute for Applied Mathematics and Scientific Computing, Indiana University, Bloomington, Indiana 47405.

‡Laboratoire d'Analyse Numérique, Université Paris-Sud, Bat. 425, 91405 Orsay, France.

to introduce certain notation and concepts. The second method is an oversimplified version of the first one in the infinite-dimensional case. The third one is more involved: it is based on a conformal mapping of the strip of analyticity on the unit circle and the utilization of various interpolation formulas and of hypergeometric functions. This method is the best suited for both finite- and infinite-dimensional cases.

For each type of approximation the results that we prove are the following: we derive the equations of the manifold, and show how an appropriate neighborhood (semi-algebraic or semi-analytic set) of the manifold contains the attractor; we show also that this neighborhood attracts all orbits in finite time at an (explicitly given) exponential rate.

This article is organized as follows. In §§1 and 2 we present the equations and the standing hypotheses and recall a few known results on global attractors, and on the time analyticity of solutions. In §3 we present the first two approximation methods, the J and I-manifolds. Section 4 contains an interpolation formula used in §5 to derive the third approximation method corresponding to the K-manifolds. Section 6 contains an extension of the previous results to more regular spaces. Finally in §7, using in particular the results of §6, we show the analytic structure of the approximating manifolds.

Most of the results presented here were announced and summarized in [8] and [6]; the finite-dimensional case was already considered in [9].

1. Preliminaries on the equations. We are given a Hilbert space H (scalar product $(\cdot,\cdot)$, norm $|\cdot|$) a linear unbounded self-adjoint operator A with domain $\mathcal{D}(A) \subset H$; we assume that A is closed, strictly positive, and has a compact inverse. It is then possible to define all the powers of A, A^s, $s \in \mathbb{R}$ which operate in the domain $\mathcal{D}(A^s)$ of A^s. It is also well known that there exists an orthonormal basis of H consisting of the eigenvectors of A:

$$\begin{cases} Aw_j = \lambda_j w_j \quad \forall j \in \mathbb{N}, \\ 0 < \lambda_1 \leq \lambda_1 \leq \ldots, \qquad \lambda_j \to \infty \quad \text{as} \quad j \to \infty. \end{cases} \tag{1.1}$$

We consider here an evolution equation of the form

$$\frac{du}{dt} + Au + R(u) = 0, \tag{1.2}$$

where u is a function defined on $\mathbb{R}$ (or on some interval of $\mathbb{R}$) with values in $\mathcal{D}(A)$ and R is a nonlinear operator from $\mathcal{D}(A)$ into H. More precisely we assume that

$$R(u) = \sum_{j=0}^{\nu} R_j(u) \tag{1.3}$$

with $R_j(u)$ of the form

$$R_j(u) = R_j(u, \ldots, u), \tag{1.4}$$

$R_j(.,\ldots,.)$ being a j-multilinear continuous operator from both $\mathcal{D}(A)^j$ into H and from $\mathcal{D}(A^{1/2})^j$ into $\mathcal{D}(A^{-1/2})$ and satisfying further appropriate hypotheses. In particular we assume that

$$R_0 \in \mathcal{D}(A) \tag{1.5}$$

and, for $j = 1, \dots, \nu$ and for some γ , $0 < \gamma < 1$:

$$R_j \text{ is } j\text{-multilinear continuous from } \mathcal{D}(A)^j \text{ into } \mathcal{D}(A^{1-\gamma}). \tag{1.6}$$

Consequently there exist constants $c_1', \dots, c_\nu'$ such that

$$|A^{1-\gamma} R_j(u)| \le c_j' |Au|^j, \qquad j = 1, \dots, \nu. \tag{1.7}$$

Here the c_i', like all the c_i and other c_i' appearing in the sequel denote various constants. It follows from (1.6) that R is compact from $\mathcal{D}(A)$ into H and bounded from $\mathcal{D}(A)$ into $\mathcal{D}(A^{1-\gamma})$:

$$|A^{1-\gamma} R(u)| \le |A^{1-\gamma} R_0| + \sum_{j=1}^{\nu} c_j' |Au|^j. \tag{1.8}$$

At the price of some slight technical complications we can consider more general operators R, analytic from $D(A)$ into H. However, the class of equations above already contains many equations arising in mathematical physics.

We are interested in the initial value problem consisting of (1.2) and

$$u(0) = u_0, \tag{1.9}$$

where u_0 is given in H. The above hypotheses do not ensure that (1.2) is a dissipative equation nor that the initial value problem (1.2), (1.9) is well posed. We assume that for every u_0 in H, the problem (1.2), (1.9) possesses a unique solution u that belongs both to the space $\mathcal{C}_b(\mathbb{R}_+; \mathbb{H})$ of continuous and bounded functions from $\mathbb{R}_+$ into H and to $L^2(0, T; \mathcal{D}(A^{1/2}))$, for every $T > 0$:

$$u \in \mathcal{C}_b(\mathbb{R}_+; H) \cap L^2(0, T; \mathcal{D}(A^{1/2})) \quad \forall T > 0; \tag{1.10}$$

furthermore, if $u_0 \in \mathcal{D}(A^{1/2})$ then

$$u \in \mathcal{C}_b(\mathbb{R}; \mathcal{D}(A^{1/2})) \cap L^2(0, T; \mathcal{D}(A)) \quad \forall T > 0. \tag{1.11}$$

We denote by $S(t)$ the corresponding operator,

$$S(t) : u_0 \in H \to u(t) \in H;$$

the operators $S(t)$, $t \ge 0$, constitute a semigroup of operators satisfying the usual relations

$$S(0) = I, \ S(t + \tau) = S(t) \cdot S(\tau) \quad \forall t, \tau \ge 0.$$

The analysis that we develop below hinges upon the complexification of equation (1.2), i.e., the passage to complex time $t = \zeta$, and the use of the time analyticity of the solutions of (1.2) (see [14], [20], and [7]); hereafter we follow [7].

We assume that the solution u of (1.2), (1.9), (1.10) is analytic from $(0, \infty)$ into $\mathcal{D}(A)$ and that, if $|A^{1/2} u_0| \le r$, the domain of analyticity of u comprises the region

$$\begin{aligned} &\Delta(r) = \Delta_1(r) \cup \Delta_2(r), \\ &\Delta_1(r) = \{\zeta \in \mathcal{C}, \ \mathrm{Re}\ \zeta \ge \delta_0, |\ \mathrm{Im}\ \zeta| \le \delta_0\}, \\ &\Delta_2(r) = \{\zeta \in \mathcal{C}, |\ \mathrm{Im}\ \zeta| \le \ \mathrm{Re}\ \zeta, 0 < \ \mathrm{Re}\ \zeta \le \delta_0\}, \end{aligned} \tag{1.12}$$

where $\delta_0 = \delta_0(r) > 0$ is a number depending on r and on (1.2). Furthermore u is bounded as a $\mathcal{D}(A^{1/2})$-valued function in the region $\Delta(r)$, and u is bounded as a $\mathcal{D}(A)$-valued function in $\Delta_1(r)$. If $u_0 \in \mathcal{D}(A), u$ is bounded as a $\mathcal{D}(A)$-valued function in $\Delta(r)$. In all cases we denote by $\mu_0(r)$, $\mu_1(r)$, $\mu_2(r)$ the following least upper bounds:

$$
\begin{aligned}
&|u(\zeta)| \leq \mu_0(r), |A^{1/2}u(\zeta)| \leq \mu_1(r) \quad \text{for } \zeta \text{ in } \Delta(r),\\
&\text{and } |Au(\zeta)| \leq \mu_2(r) \text{ for } \zeta \text{ in } \Delta_1(r).
\end{aligned} \tag{1.13}
$$

When needed we shall write the series expansion of $u = u(\zeta)$ and $f = f(\zeta) = R(u(\zeta))$; for instance if u is analytic around $\zeta = 0$, then

$$
u(\zeta) = \sum_{n=0}^{\infty} \frac{\zeta^n}{n!} u^{(n)}(0), \tag{1.14}
$$

$$
f(\zeta) = R(u(\zeta)) = \sum_{n=0}^{\infty} \frac{\zeta^n}{n!} f^{(n)}(0). \tag{1.15}
$$

For the sake of simplicity we shall write

$$
u_n = u^{(n)}(0), \qquad f_n = f^{(n)}(0). \tag{1.16}
$$

By successive differentiations of (1.2) and by using the chain differentiation rule, we can express all the u_n's, $n \geq 1$, and all the f_n's, $n \geq 0$, as functions of u_0:

$$
\begin{aligned}
u_1 &= -Au_0 - f_0,\\
f_0 &= R_0 + \sum_{j=1}^{\nu} R_j(u_0),\\
u_2 &= -Au_1 - f_1,\\
f_1 &= R_1(u_1) + R_2(u_1, u_0) + R_2(u_0, u_1) + \cdots.
\end{aligned} \tag{1.17}
$$

If, for instance, we can consider the case where the number ν in (1.3) is equal to 2,

$$
R(u) = R_0 + R_1(u) + R_2(u, u), \tag{1.18}
$$

then we find

$$
\begin{cases}
f_0 = R_0 + R_1(u_0) + R_2(u_0, u_0),\\
f_1 = R_1(u_1) + R_2(u_0, u_1) + R_2(u_1, u_0),\\
\cdots\\
f_n = R_1(u_n) + \sum_{j=0}^{n} \binom{n}{j} R_2(u_j, u_{n-j}),\\
\cdots.
\end{cases} \tag{1.19}
$$

The u_n's are themselves expressed in terms of u_0 by successive differentiation of (1.2):

$$
\begin{cases}
u_1 = -Au_0 - R_0 - R_1(u_0) - R_2(u_0, u_0),\\
u_2 = -Au_1 - R_1(u_1) - R_2(u_0, u_1) - R_2(u_1, u_0),\\
u_3 = -Au_2 - R_1(u_2) - R_2(u_0, u_2) - 2R_2(u_1, u_1) - R_2(u_2, u_0),\\
\cdots\\
u_{n+1} = -Au_n - R_1(u_n) - \sum_{j=0}^{n} \binom{n}{j} R_2(u_j, u_{n-j}),\\
\cdots.
\end{cases} \tag{1.20}
$$

Remark 1.1. There are many relevant equations in mathematical physics, of type (1.2)–(1.4), that satisfy the above hypotheses, namely (1.5), (1.6), and (1.10)–(1.12).

For example, for $\nu = 2$, it suffices that R_1 and R_2 satisfy

$$|R_1 u| \leq c_1 |A^{1/2} u| \quad \forall u \in \mathcal{D}(A^{1/2}), \tag{1.21}$$

$$|R_2(u,v)| \leq \begin{cases} c_2 |u|^{1/2} |Au|^{1/2} |A^{1/2} v|, \\ c_2 |u|^{1/2} |A^{1/2} u|^{1/2} |A^{1/2} v| |Av|^{1/2} \end{cases} \quad \forall u, v \in \mathcal{D}(A), \tag{1.22}$$

$$|A^{1/2} R_2(u,v)| \leq c_3 |u|^{1/2} |A^{1/2} u|^{1/2} |v|^{1/2} |A^{1/2} v|^{1/2} \quad \forall u, v \in \mathcal{D}(A), \tag{1.23}$$

$$(R_2(u,u), u) = 0 \quad \forall u \in \mathcal{D}(A). \tag{1.24}$$

Conditions (1.21)–(1.24) are fulfilled by the Navier–Stokes equations in space dimension two, and related equations. Other equations satisfying the hypotheses of §1 can be found in [25]. Specific equations will not be considered in this article. The reader is referred to subsequent articles for the study of the Navier–Stokes equations and other specific equations. However, in all these cases one can consider (1.2) in the spaces $\mathcal{D}(A^k)$ (= domain of A^k normed by $|A^k u|$) for $k \geq 1$.

Higher regularity. Concerning the mapping R in (1.2), we can assume that for every $k \geq 1$,

$$\begin{cases} R_j \text{ is } j\text{-multilinear continuous from } \mathcal{D}(A^k)^j, \\ \text{into } \mathcal{D}(A^{k-\gamma}) \text{ for } j = 1, \ldots, \nu, \end{cases} \tag{1.25}$$

and $R_0 \in \mathcal{D}(A^k)$. Consequently, there exist constants $c'_{j,k}$ such that

$$|A^{k-\gamma} R_j(u)| \leq c'_{j,k} |A^k u|^j, \qquad j = 1, \ldots, \nu \quad \forall k \geq 1. \tag{1.26}$$

We extend to $\mathcal{D}(A^k)$ all the hypotheses made concerning (1.2). We assume that, for every $u_0 \in \mathcal{D}(A^k)$, or $u_0 \in \mathcal{D}(A^{k+1/2})$, the initial value problem possesses a unique solution,

$$u \in \mathcal{C}_b(\mathbb{R}; \mathcal{D}(A^k)) \cap L^2(0,T; \mathcal{D}(A^{k+1/2})) \quad \forall T > 0\,, \tag{1.27}$$

or

$$u \in \mathcal{C}_b(\mathbb{R}; \mathcal{D}(A^{k+1/2})) \cap L^2(0,T; \mathcal{D}(A^{k+1/2})) \quad \forall T > 0. \tag{1.28}$$

We assume that $S(t)u_0$ is continuous from $(0,\infty) \times H$ into $\mathcal{D}(A^k)$ for every k; moreover that, for $u_0 \in \mathcal{D}(A^{1/2})$ with $|A^{1/2} u_0| \leq r$, u is analytic in $\Delta(r)$ as a $\mathcal{D}(A^{k+1})$-valued function, bounded as a $\mathcal{D}(A^{k+1/2})$-valued function in $\Delta(r)$ and bounded as a $\mathcal{D}(A^{k+1})$-valued function in $\Delta_1(r)$.

As mentioned above, the Navier–Stokes equations in space dimension two, or more generally, all dissipative equations considered in [25] satisfy these conditions. Our first interest in these more regular spaces resides in the following.

PROPOSITION 1.2. *The vector-valued coefficients* u_n *and* f_n *defined in* (1.14), (1.15), *and* (1.16) *are, as functions of* u_0, *analytic polynomial maps from* $\mathcal{D}(A^n)$ *and* $\mathcal{D}(A^{n+1})$, *respectively, into* H.

We recall that an analytic polynomial map from X into Y, with X, Y some Banach spaces, is a finite sum of functions $u \mapsto \varphi(u)$ of the form $\varphi(u) = \phi(u, u, \dots, u)$ where ϕ is a continuous multilinear map from X into Y. (See [12, Ch. 26].) With this definition, from the computations of the type (1.17), (1.20) and from the continuity conditions (1.25), we readily infer by induction (in n), from that each u_n is an analytic map from $\mathcal{D}(A^{n+k})$ into $\mathcal{D}(A^k)$ for all $k = 0, 1, \dots$ and $n = 0, 1, 2, \dots$. Using now the relations of the type (1.17), (1.19) as well as (1.25), we obtain that f_n is an analytic polynomial map from $\mathcal{D}(A^{n+1+k})$ into $\mathcal{D}(A^{k+1-\gamma})$ for all $k = 0, 1, \dots$ and $n = 0, 1, \dots$. This concludes the proof of Proposition 1.2.

Remark 1.3. Since, as shown above, f_n is a continuous map from $\mathcal{D}(A^{n+1+k})$ into $\mathcal{D}(A^{k+1-\gamma})$ and the identity map from $\mathcal{D}(A^{k+1-\gamma})$ into $\mathcal{D}(A^k)$ is compact, it follows that f_n is a compact map from $\mathcal{D}(A^{n+1+k})$ into $\mathcal{D}(A^k)$ for all $k = 0, 1, 2, \dots$ and $n = 0, 1, 2, \dots$.

2. Attractors and their integral representation. We shall work with equations of the form (1.2) that are *dissipative*. One characterization of dissipativity is the existence of an absorbing set (see [1], [25]): This is a bounded set $\mathcal{B}_0 \subset \mathcal{D}(A)$ such that

$$\text{(2.1)} \qquad \begin{cases} \text{For every bounded set } \mathcal{B} \subset \mathcal{D}(A)\,, \text{ there exists } t_0 = t_0(\mathcal{B}) \\ \text{such that } S(t)\mathcal{B} \subset \mathcal{B}_0 \quad \forall t \geq t_0. \end{cases}$$

The existence of an absorbing set implies that the orbits do not wander in the whole space as it happens with hamiltonian systems but rather concentrate in the region $\mathcal{B}_0$ or even in part of it.

A common aspect of dissipative systems is the existence of a global attractor $\mathcal{A}$, that is, the maximal compact connected set $\mathcal{A}$ in $\mathcal{D}(A)$ enjoying the following properties

$$\text{(2.2)} \qquad S(t)\mathcal{A} = \mathcal{A}, \qquad t \geq 0,$$

$$\text{(2.3)} \qquad \mathcal{A} \text{ attracts the bounded sets of } H,$$

i.e., for every bounded set $\mathcal{B} \subset H$

$$\operatorname{dist}\,(S(t)\mathcal{B}, \mathcal{A}) = \sup_{x \in \mathcal{B}} \inf_{y \in \mathcal{A}} |S(t)x - y| \to 0 \quad \text{as } t \to \infty.$$

Like (1.10)–(1.13), properties (2.1)–(2.3) do not follow from previous hypotheses and are assumed here. Note that $\mathcal{A} \subset \mathcal{B}_0$ and, since $\mathcal{B}_0$ is bounded in $\mathcal{D}(A)$ we have

$$\text{(2.4)} \qquad |u_0| \leq r_0, \quad |A^{1/2}u_0| \leq r_1, \quad |Au_0| \leq r_2 \quad \forall u_0 \in \mathcal{B}_0\,,$$

where r_0, r_1, r_2 are adequate constants depending only on the equation. We refer the reader to [11], [25] for numerous classes of equations for which (1.10)–(1.13) and (2.1)–(2.4) are fulfilled.

A transcendental equation for $\mathcal{A}$ is easily derived; indeed u_0 belongs to $\mathcal{A}$ if and only if u_0 belongs to a complete orbit $\{u(t),\ t \in \mathbb{R}\}$ bounded in H. In this case we may

assume without loss of generality that $u(0) = u_0$ and by integration of (1.2) between $t < 0$ and 0 we find

$$u(0) = e^{tA}u(t) - \int_t^0 e^{\tau A}R(u(\tau))d\tau.$$

Upon letting $t \to -\infty$ and remembering that $u(\cdot)$ is bounded, we find

$$u(0) = -\int_{-\infty}^0 e^{\tau A}R(u(\tau))d\tau;$$

alternatively with a slight change of notation we find an equation for $\mathcal{A}$:

$$u_0 = -\int_{-\infty}^0 e^{\tau A}R(S(\tau)u_0)d\tau. \tag{2.5}$$

Here $S(\tau) = (S(-\tau))^{-1}$ for $\tau < 0$, this operator being defined on $S(-\tau)H$; note that $S(-\tau)$ is injective due to the analyticity in time of the solutions.

If u_0 is a point on the attractor then as observed above we may asume that $u_0 = u(0)$, where $\{u(\tau)\ ,\ \tau \in \mathbb{R}\}$ is a complete orbit. Due to (1.12) and since $\mathcal{A}$ is bounded in $\mathcal{D}(A^{1/2})$, u is a $\mathcal{D}(A)$-valued analytic function in a region of $\mathbb{C}$ which comprises the strip

$$\Delta = \{\zeta,\ |\mathrm{Im}\zeta| \le \delta_0\}. \tag{2.6}$$

The width $2\delta_0$ of this strip can be chosen the same for all points u_0 of $\mathcal{A}$, namely $\delta_0(r_1)$ (see (1.13)). Furthermore u is bounded in $\mathcal{D}(A)$ in the region (2.6) and we shall write as in (1.13) (with $\mu_j = \mu_j(r_1)\ ,\ j = 0, 1, 2$)

$$|u(\zeta)| \le \mu_0, \quad |A^{1/2}u(\zeta)| \le \mu_1, \quad |Au(\zeta)| \le \mu_2 \quad \forall \zeta \in \Delta. \tag{2.7}$$

It is useful to remark now that if $u_0 \in \mathcal{B}_0$ but not necessarily in $\mathcal{A}$, then the analytic extension $u(\zeta)$ of $u(t) = S(t)u_0$ satisfies

$$|u(\zeta)| \le \mu_0, \qquad |A^{1/2}u(\zeta)| \le \mu_1 \text{ in } \Delta(r_1) \text{ and } |Au(\zeta)| \le \mu_2 \text{ in } \Delta_1(r_1). \tag{2.8}$$

We will make considerable use of a conformal mapping of the strip Δ in (2.6) onto the unit disk of $\mathbb{C}$,

$$\mathbb{D} = \{z \in \mathbb{C}\ ,\ |z| < 1\}. \tag{2.9}$$

This is given by

$$z = \varphi(\zeta) = \frac{e^{\zeta/\delta'} - 1}{e^{\zeta/\delta'} + 1} \tag{2.10}$$

where $\delta' = 2\delta_0/\pi$. We observe that

$$\varphi((-\infty, 0]) = (-1, 0], \qquad \varphi(0) = 0,$$

and the inverse mapping is

$$\zeta = \psi(z) = \delta' \log \frac{1+z}{1-z} = 2\delta'\left(z + \frac{z^3}{3} + \frac{z^5}{5} + \cdots\right). \tag{2.11}$$

Using (2.5), (2.10) and the notation $f(\zeta) = R(u(\zeta))$ (see (1.14)), we rewrite (2.5) in the form

$$u_0 = -\int_{-1}^{0} e^{\psi(\sigma)A} f(\psi(\sigma)) d\psi(\sigma)$$
$$= -\int_{-1}^{0} e^{\psi(\sigma)A} f(\psi(\sigma)) \frac{2\delta' d\sigma}{1-\sigma^2},$$

or

$$u_0 = -2\delta' \int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta' A} g(\sigma) \frac{d\sigma}{1-\sigma^2}, \tag{2.12}$$

where

$$g(\sigma) = f(\psi(\sigma)). \tag{2.13}$$

3. Approximation of attractors by analytic sets (I). Our aim in this section and in §5 is to define some (simple) analytic manifolds that approximate the attractor $\mathcal{A}$ in the sense that $\mathcal{A}$ lies in a (thin) neighborhood of this manifold. The construction of this approximating manifold $\mathcal{M}$ follows a general methodology that we shall now describe; of course the dimension and complexity of $\mathcal{M}$ depend on ϵ and on the equation.

Our procedure depends on the choice of a cut-off value λ_m in the spectrum of A. We set $\Lambda = \lambda_{m+1}$ and denote by $Q = Q_m$ the orthogonal projection onto the space spanned by the eigenvectors corresponding to the eigenvalues larger than or equal to λ_{m+1}. To avoid unnecessary technicalities, we choose m such that $\lambda_m < \lambda_{m+1}$. Of course $P = P_m = I - Q_m$ is the projector onto the space spanned by the first m eigenvectors of A, $w_1, \ldots, w_m$ (see (1.1)).

We set $p_m = P_m u, q_m = Q_m u$ or dropping the indices m for the sake of simplicity

$$p = Pu, \qquad q = Qu.$$

By projecting equation (1.2) onto the spaces PH and QH we obtain a system of equations satisfied by p and q that is equivalent to (1.2). We recall that P and Q commute with A and hence the projected equations read

$$\frac{dp}{dt} + Ap + PR(p+q) = 0, \tag{3.1}$$

$$\frac{dq}{dt} + Aq + QR(p+q) = 0. \tag{3.2}$$

By applying Q to (2.5) we obtain

$$q(0) = -\int_{-\infty}^{0} e^{tA} QR(u(\tau)) d\tau. \tag{3.3}$$

The approximation procedures that we develop here hinge on (3.3). They consist in showing that the right-hand side of (3.3) is the sum of an analytic function of $u(0)$ and of a term which can be made as small as desired; furthermore this analytic function

of $u(0)$ is of polynomial type, i.e., a finite sum of multilinear continuous functions of $u(0)$. In this section we derive two polynomial approximations of the right-hand side of (3.3), that we denote $-I_N(u(0))$ and $-J_N(u(0))$. In §5 we derive a third more involved approximation denoted $-K_N(u(0))$; N is an integer which appears in the course of the construction.

Truncated series. Let u_0 be a point on the attractor $\mathcal{A}$. As mentioned in §2, u_0 necessarily belongs to a complete orbit $\{u(t), t \in \mathbb{R}\}$ bounded in H and, without loss of generality we can assume that $u_0 = u(0)$. The function $t \to u(t)$ is in fact $\mathcal{D}(A)$-analytic in a neighborhood of the region Δ given in (2.6).

We consider the series expansion of $u = u(\tau)$

$$u(\tau) = \sum_{n=0}^{\infty} u_n \frac{\tau^n}{n!}, \qquad u_n = u^{(n)}(0), \tag{3.4}$$

and for a fixed integer N, we consider

$$u_N(\tau) = \sum_{n=0}^{N} u_n \frac{\tau^n}{n!}. \tag{3.5}$$

As is well known u_N is a uniform approximation of u in a ball centered at $\tau = 0$ of radius smaller than $\delta_0 = 2\delta$, say 2δ. More precisely, due to Cauchy's formula and (2.7),

$$|u_n| = |u^{(n)}(0)| \leq \frac{\mu_0 n!}{(2\delta)^n}, \tag{3.6}$$

where $\mu_0 = \mu_0(r_1)$; thus for $|\tau| \leq \delta$

$$\begin{aligned} |u(\tau) - u_N(\tau)| &= \left| \sum_{n=N+1}^{\infty} u_n \frac{\tau^n}{n!} \right| \\ &\leq \sum_{n=N+1}^{\infty} \frac{\mu_0}{2^n} = \frac{\mu_0}{2^N}, \end{aligned}$$

that is,

$$|u(\tau) - u_N(\tau)| \leq \frac{\mu_0}{2^N} \text{ for } |\tau| \leq \delta. \tag{3.7}$$

Consider similarly the function $\tau \to f(\tau) = R(u(\tau))$. This function is analytic in a neighborhood of Δ; owing to (2.7) and the hypotheses on R, f is bounded as follows on Δ:

$$\begin{aligned} |f(\tau)| = |R(u(\tau))| &\leq |R_0| + \sum_{j=1}^{\nu} |R_j(u(\tau))| \\ &\leq \lambda_1^{\gamma-1} |A^{1-\gamma} R(u(\tau))| \leq \lambda_1^{\gamma-1} \left\{ |A^{1-\gamma} R_0| + \sum_{j=1}^{\nu} c_j' (\mu_2)^j \right\}, \end{aligned}$$

where we used (2.7). Thus

$$\begin{cases} |f(\tau)| \le \rho_0, \qquad \tau \in \Delta, \text{ with} \\ \rho_0 = \lambda_1^{\gamma-1}\{|A^{1-\gamma}R_0| + \sum_{j=1}^{\nu} c_j'(\mu_2)^j\}. \end{cases} \tag{3.8}$$

The series expansion of f is written as (3.4):

$$f(\tau) = \sum_{n=0}^{\infty} f_n \frac{\tau^n}{n!}, \qquad f_n = f^{(n)}(0),$$

with

$$|f_n| = |f^{(n)}(0)| \le \frac{\rho_0 n!}{(2\delta)^n}.$$

We set

$$f_N(\tau) = \sum_{n=0}^{N} f_n \frac{\tau^n}{n!};$$

then as for u_N, we have

$$|f(\tau) - f_N(\tau)| \le \frac{\rho_0}{2^N} \quad \text{for } |\tau| \le \delta. \tag{3.9}$$

The approximation J_N. For $\delta = \delta_0/2$, we decompose the integral in the right-hand side of (3.3) as

$$\int_{-\infty}^{0} = \int_{-\infty}^{-\delta} + \int_{-\delta}^{0}.$$

The first integral is easily majorized:

$$\left| \int_{-\infty}^{-\delta} e^{\tau A} QR(u(\tau))d\tau \right| \le \rho_0 \int_{-\infty}^{-\delta} e^{\tau\Lambda} d\tau \le \frac{\rho_0}{\Lambda} e^{-\delta\Lambda}. \tag{3.10}$$

For the second integral we write

$$\begin{aligned} \int_{-\delta}^{0} e^{\tau A} QR(u(\tau))d\tau &= \int_{-\delta}^{0} e^{\tau A} Qf(\tau)d\tau \\ &= \int_{-\delta}^{0} e^{\tau A} Q(f(\tau) - f_N(\tau))d\tau + \int_{-\delta}^{0} e^{\tau A} Qf_N(\tau)d\tau. \end{aligned}$$

Due to (3.9),

$$\left| \int_{-\delta}^{0} e^{\tau A} Q(f(\tau) - f_N(\tau))d\tau \right| \le \frac{\rho_0}{2^N} \int_{-\delta}^{0} e^{\tau\Lambda} d\tau \le \frac{\rho_0 \delta}{2^N}. \tag{3.11}$$

Finally, by virtue of Proposition 1.2, the last integral as function of $u(0) = u_0$ is an analytic polynomial map from $\mathcal{D}(A^{N+1})$ into H; we denote it $QJ_N(u_0)$, where

$$\begin{aligned} J_N(u_0) &= \int_{-\delta}^{0} e^{\tau A} f_N(\tau)d\tau \\ &= \int_{-\delta}^{0} e^{\tau A} \sum_{n=0}^{N} f_n \frac{\tau^n}{n!} d\tau \\ &= \sum_{n=0}^{N} S_n f_n, \end{aligned} \tag{3.12}$$

with

$$S_n = \frac{1}{n!}\int_{-\delta}^{0} e^{\tau A}\tau^n d\tau \tag{3.13}$$

(after integrating by parts n times)

$$= (-1)^n A^{-n-1} + (-1)^{n+1} e^{-\delta A} \sum_{j=0}^{n} \frac{\delta^{n-j}}{(n-j)!} A^{-j-1}.$$

We recall that the f_n's are expressed in term of $u_0, \dots, u_n$, through formulas like (1.17) and (1.19), and $u_1, \dots, u_n$ are expressed in terms of u_0 through formula like (1.17), (1.20). Therefore the notation $J_N(u_0)$ is justified as J_N is indeed a function of u_0. Actually, referring again to Proposition 1.2, J_N is a analytic polonomial map from $\mathcal{D}(A^{N+1})$ into H; moreover, by virtue of Remark 1.3, this map is also compact.

From the preceding, we infer that for any $u_0 \in \mathcal{A}$,

$$|Q(u_0 + J_N(u_0))| \leq \frac{\rho_0}{\Lambda} e^{-\delta\Lambda} + \frac{\rho_0 \delta}{2^N} \tag{3.14}$$

and we conclude with the following.

THEOREM 3.1. *We consider the dissipative equation* (1.2) *and its compact attractor $\mathcal{A}$, the hypotheses on A and R of* §1 *being satisfied. We are also given a spectral projector $Q = Q_m$ on H as indicated above. Also let J_N be the function defined in* (3.12) *and* (3.13). *Then J_N it is a compact analytic polynomial map from $\mathcal{D}(A^{N+1})$ into H and for every $\epsilon > 0$, we have*

$$|Q_m(u_0 + J_N(u_0))| \leq \epsilon \quad \forall u_0 \in \mathcal{A}, \tag{3.15}$$

where $|\cdot|$ is the norm in H, provided

$$\lambda_{m+1} = \Lambda \geq \frac{2}{\delta_0} \log \frac{2\rho_0}{\lambda_1 \epsilon}, \qquad N \geq \frac{1}{\log 2} \log \frac{\rho_0 \delta_0}{\epsilon}. \tag{3.16}$$

Remark 3.2. Theorem 3.1 expresses in particular the fact that the attractor lies in a neighborhood of the set $\mathcal{M}_J$ consisting of the roots of the equation

$$Q_m(u_0 + J_N(u_0)) = 0, \tag{3.17}$$

more precisely in the set

$$\{u_0 \in H,\ |Q_m(u_0 + J_N(u_0))| \leq \epsilon\}. \tag{3.18}$$

For Theorem 3.1 to be of interest, it is desirable that m and N are not too large and that the set (3.19) is not too thick. We already observed that N and $\delta\lambda_{m+1}$ should be of order of $\log(1/\epsilon)$ (see (3.16)). The question of the thickness of (3.18) will be addressed in §7, where it will also be shown that $\mathcal{M}_J$ is an analytic set in H.

Absorbing property. Since (3.18) is a neighborhood of the attractor $\mathcal{A}$, any orbit starting from a point $u(0) = u_0$ in the space, eventually enters the set (3.18). Our aim is now to estimate the time of absorption into this set. In fact we shall show that orbits enter into (3.18) in a finite time which, provided m and N are large enough, depends only on $|u_0|$, the norm of u_0 in H.

Let $\mathcal{M} = \mathcal{M}_\epsilon$ be the set (3.17), where m and N will be chosen essentially as in Theorem 3.1, precisely satisfying (3.16) with ϵ replaced by $\epsilon/2$. Consider an orbit $u(t) = S(t)u_0$ (for $t \geq 0$) that may or may not lie on the attractor. By virtue of the dissipativity of the equation (1.2), we can assume, without loss of generality, that $S(t)u_0$ belongs to the absorbing set $\mathcal{B}_0$ (see (2.1)) for all $t \geq 0$. Let $t_0 = \delta$ and let $t > t_0$; we can write, as in §2,

$$u(t) = e^{-tA}u_0 - \int_0^t e^{t-\tau A}R(u(\tau))d\tau.$$

By translation in time we can assume that u is defined on $(-t, 0)$ and replace t by 0. Hence

$$u(0) = e^{-tA}u(-t) - \int_{-t}^0 e^{\tau A}R(u(\tau))d\tau$$

(where now $u(-t) = u_0$), and after projecting on QH:

$$q(0) = e^{-tA}q(-t) - \int_{-t}^0 e^{\tau A}QR(u(\tau))d\tau. \tag{3.19}$$

The integral in the right-hand side of (3.19) is treated as before:

$$\int_{-t}^0 = \int_{-t}^{-\delta} + \int_{-\delta}^0 ,$$

$$\left|\int_{-t}^{-\delta} e^{\tau A}QR(u(\tau))d\tau\right| \leq \rho_0 \int_{-\infty}^{-\delta} e^{\tau\Lambda}d\tau \leq \frac{\rho_0}{\Lambda}e^{-\delta\Lambda} ,$$

$$\int_{-\delta}^0 e^{\tau A}f(\tau)d\tau = \int_{-\delta}^0 e^{\tau A}f_N(\tau)d\tau + \int_{-\delta}^0 e^{\tau A}(f(\tau) - f_N(\tau))d\tau ,$$

$$\int_{-\delta}^0 e^{\tau A}f_N(\tau)d\tau = J_N(u(0)) ,$$

while

$$\left|\int_{-\delta}^0 e^{\tau A}(f(\tau) - f_N(\tau))d\tau\right| \leq \frac{\rho_0\delta}{2^N}.$$

Since

$$|e^{-tA}q(-t)| \leq r_0e^{-t\Lambda}$$

(see (2.4)), we conclude that

$$\begin{aligned}&|Q(u(0) + J_N(u(0)))|\\ &\leq \frac{\rho_0e^{-\delta\Lambda}}{\Lambda} + \frac{\rho_0\delta}{2^N} + r_0e^{-t\Lambda}.\end{aligned}$$

Shifting back to forward time we obtain

$$Q(S(t)u_0 + J_N(S(t)u_0)) \leq \frac{\rho_0}{\Lambda}e^{-\delta_0\Lambda/2} + \frac{\rho_0\delta_0}{2^{N+1}} + r_0e^{-t\Lambda} \quad \forall u_0 \in \mathcal{B}, \quad \forall t \geq 0. \tag{3.20}$$

It follows that for m, N satisfying the conditions (3.16) with ϵ replaced by $\epsilon/2$, we have

$$|Q(S(t)u_0 + J_N(S(t)u_0))| \leq \frac{\epsilon}{2} + r_0 e^{-t\lambda_{m+1}} \quad \forall u_0 \in \mathcal{B}, \quad \forall t \geq 0. \tag{3.21}$$

Therefore we can now conclude with the following.

THEOREM 3.3. *The hypotheses are those of Theorem* 3.1, *and m and N are chosen as indicated above. Then there exist a time t_0 depending only on $|u_0|$ (and of course on the equation* (1.2)) *such that*

$$|Q_m(u(t) + J_N(u(t)))| < \epsilon \quad \forall t \geq t_0.$$

Remark 3.4. The results of Theorems 3.1 and 3.3 can be related to the concepts of Approximate Inertial Manifolds (AIMs) introduced in [5]. We recall that an AIM is a smooth finite-dimensional manifold which attracts all orbits in one of its neighborhoods. In that sense the set $\mathcal{M}_J$ defined by (3.17) is an AIM. However a major difference with the AIMs in [5] and with other AIMs subsequently constructed is that all these AIMs are graphs above PH, while $\mathcal{M}$ is a graph above PH only if λ_{m+1} is very large (see §7 below). For more complete information about AIMs the reader is referred to [5], [3], [18], [19], and [27].

The approximation I_N. We give now a simplified form of the approximation. Starting as in the previous case (approximation J_N) we consider the expression (3.13) of S_n that we write

$$S_n = S_n' \left(I - \left(\sum_{j=0}^{n} \frac{\delta^j}{j!} A^j \right) e^{-\delta A} \right) \tag{3.22}$$

with

$$S_n' = (-1)^n A^{-n-1}. \tag{3.23}$$

Then, for every fixed $n = 0, 1, \ldots,$

$$\begin{aligned} |S_n'^{-1} S_n - I|_{\mathrm{op}} &\leq \sup_{\lambda \geq \Lambda} \left(e^{-\delta\lambda} \sum_{j=0}^{n} \frac{(\delta\lambda)^j}{j!} \right) \\ &= e^{-\delta\Lambda} \sum_{j=0}^{n} \frac{(\delta\Lambda)^j}{j!} = \int_{\delta\Lambda}^{\infty} \frac{e^{-\alpha}\alpha^n}{n!} d\alpha \leq e^{-\delta\Lambda/2}(n+1)2^{n+1}; \end{aligned} \tag{3.24}$$

here, as well as throughout, $|\cdot|_{\mathrm{op}}$ denotes the operator norm on the appropriate space, namely H in the present case. Therefore we can expect that for large values of Λ, the function I_N of u_0 defined by

$$I_N(u_0) = \sum_{n=0}^{N} S_n' f_n, \tag{3.25}$$

will have the same approximation properties as J_N. It is clear that I_N is a compact analytic polynomial map from $\mathcal{D}(A^{N+1})$ into H. We note that

$$u_0 + I_N(u_0) = u_0 + \sum_{n=0}^{N} (-1)^n A^{-n-1} f_n.$$

On the other hand, (1.2) yields by successive differentiation

$$u_{n+1} + Au_n + f_n = 0, \qquad n \geq 0. \tag{3.26}$$

Hence

$$(-1)^n A^{-n-1} u_{n+1} + (-1)^n A^{-n} u_n + (-1)^n A^{-n-1} f_n = 0$$

and by summing for $n = 0, \ldots, N$, there remains

$$(-1)^N A^{-N-1} u_{N+1} + u_0 + I_N(u_0) = 0.$$

Thus

$$\begin{cases} u_0 + I_N(u_0) &= (-1)^{N+1} A^{-N-1} u_{N+1} \\ &= (-1)^{N+1} A^{-N-1} \frac{d^{N+1}}{dt^{N+1}} (S(t)u_0)|_{t=0}. \end{cases} \tag{3.27}$$

Now, by using the estimated (3.6), we obtain

$$\begin{aligned} |Q_m(u_0 + I_N(u_0))| &\leq \frac{1}{\Lambda^{N+1}} \left| \frac{d^{N+1}}{dt^{N+1}} S(t)u_0 \right|_{t=0} \\ &= \frac{1}{\Lambda^{N+1}} |u_{N+1}| \leq \frac{\mu_0 (N+1)!}{(\delta_0 \Lambda)^{N+1}} \quad \forall u_0 \in \mathcal{A} \end{aligned} \tag{3.28}$$

and (by using the fact that the disk of radius 2δ with center t is contained in $\Delta_1(r_1)$ for $t \geq 2\delta_0$)

$$|Q_m(S(t)u_0 + I_N(S(t)u_0))| \leq \frac{1}{\Lambda^{N+1}} \left| \frac{d^{N+1}}{dt^{N+1}} S(t)u_0 \right| \leq \frac{\mu_0 (N+1)!}{(\delta_0 \Lambda)^{N+1}} \quad \forall u_0 \in \mathcal{B} \tag{3.29}$$

provided that $t \geq 2\delta_0$.

We can now state the following.

THEOREM 3.5. *The hypotheses are those of Theorem* 3.1. *Let I_N be the function defined by* (3.25). *Then, I_N is a compact analytic polynomial map from $\mathcal{D}(A^{N+1})$ into H and for every ϵ in $(0, \mu_0/2)$,*

$$|Q_m(u_0 + I_N(u_0))| \leq \epsilon \quad \forall u_0 \epsilon \mathcal{A} \tag{3.30}$$

provided

$$\lambda_{m+1} = \Lambda \geq \frac{1}{\delta_0}(N+2), \qquad N \geq 2 \log \frac{\mu_0}{\epsilon}. \tag{3.31}$$

Proof. By virtue of of (3.29) we need only to verify that if λ_{m+1} and N satisfy (3.31) then

$$\frac{\mu_0 (N+1)!}{(\delta_0 \Lambda)^{N+1}} < \epsilon. \tag{3.32}$$

In order to check when (3.32) is valid, we could use Stirling's formula, but for our purpose it suffices to use the trivial estimate

$$(N+1)! \leq (N+2)^{N+2} e^{-N-1}.$$

Therefore, if Λ satisfies the first relation in (3.31), we obtain

$$\frac{\mu_0(N+1)!}{(\delta_0\Lambda)^{N+1}} \leq \mu_0(N+2)\left(\frac{1}{e}\right)^{N+1}.$$

Obviously the right-hand side above is less than ϵ if N satisfies the second relation in (3.31). □

Theorem 3.5 expresses the fact that, for the indicated values of m and N, $\mathcal{A}$ lies in a neighborhood of the set $\mathcal{M}_I$ of equation

$$Q_m(u_0 + I_N(u_0)) = 0, \tag{3.33}$$

more precisely in the set

$$|Q_m(u_0 + I_N(u_0))| \leq \epsilon. \tag{3.34}$$

From the estimate (3.29) we readily infer the following analogue of Theorem 3.5.

THEOREM 3.6. *The hypotheses are those of Theorem* 3.5 *and* m *and* N *are chosen to satisfy* (3.31) *with* ϵ *replaced by* $\epsilon/2$. *Then there exists a time* t_0 *depending only on* $|u_0|$ *such that*

$$|Q_m(S(t)u_0 + I_N(S(t)u_0))| \leq \epsilon \quad \forall t \geq t_0. \tag{3.35}$$

Remark 3.7. Remarks analogous to Remarks 3.2 and 3.4 are valid here too; in particular, the set $\mathcal{M}_I$ is an Approximate Inertial Manifold in the sense of [5]. However there is an important difference between the estimates (3.15) and (3.30). Namely the former is valid for all large enough λ_{n+1} and N while the latter is certainly valid only if morevoer $\delta_0\lambda_{m+1}$ is at least of the same size as N.

4. An interpolation formula. In view of the study of the third type of approximation of attractors considered in §5 we study here an interpolation problem in a context independent of (2.12) but we have in view, of course, its application to (2.12).

Let $\mathcal{H}$ be a separable (complex) Hilbert space and let A be a strictly positive self-adjoint operator in $\mathcal{H}$. We set $\lambda = \inf\{(Ah, h) : h \in \mathcal{D}(A)\ ,\ |h| = 1\} > 0$, where we denote by $(\cdot\ ,\ \cdot)$ the scalar product in $\mathcal{H}$ and $|\cdot| = (\cdot\ ,\ \cdot)^{1/2}$. Also the norm of a bounded linear operator T on $\mathcal{H}$ will be denoted by $|T|_{op}$. Let $H^\infty(\mathcal{H})$ denote the space of all bounded and analytic functions $g : \mathbb{D} = \{z \in \mathbb{C}\ : |z| < 1\} \to \mathcal{H}$. The norm in $H^\infty(\mathcal{H})$ is defined by

$$\|g\|_\infty = \sup\{|g(z)|,\ z \in \mathbb{D}\}.$$

Since

$$\int_{-1}^{0} \left|\left(\frac{1+\sigma}{1-\sigma}\right)^A\right|_{\text{op}} \frac{d\sigma}{1-\sigma^2} \leq \int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^\lambda \frac{d\sigma}{1-\sigma^2} \leq \int_{-1}^{0} (1+\sigma)^{\lambda-1} d\sigma < \infty\ ,$$

the integral

$$T(g) = \int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^A g(\sigma) \frac{d\sigma}{1-\sigma^2}, \tag{4.1}$$

where $g \in H^\infty(\mathcal{H})$, is absolutely convergent and defines an operator T on $H^\infty(\mathcal{H})$.

We will consider the following interpolation problem:

Find operators T_0 , T_1 , ... *such that*

$$\sup\left\{\left|T(g)-\sum_{n=0}^{N}\frac{1}{n!}T_n g^{(n)}(0)\right|_{\mathrm{op}} \; ; \; g\in H^\infty(\mathcal{H}) \; , \; \|g\|_\infty \le 1\right\} \tag{4.2}$$

goes to 0, *for* $N \to \infty$, *as fast as possible.*

We will give an answer to this problem, which although not optimal, will nevertheless be sufficient for our purpose.

We start by defining

$$T_n = \int_{-1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^A \frac{\sigma^n d\sigma}{1-\sigma^2} \qquad (n=0,1,2,\ldots), \tag{4.3}$$

and by noticing that for any $g \in H^\infty(\mathcal{H})$, the function $g(e^{i\theta}) = \text{strong lim}\, g(re^{i\theta})(0 \le \theta < 2\pi)$ exists almost everywhere, $|g(e^{i\theta})| \le \|g\|_\infty$ almost everywhere and

$$g(z) = \frac{1}{2\pi}\int_0^{2\pi}\frac{g(e^{i\theta})d\theta}{1-e^{-i\theta}z} \qquad (z\in\mathbb{D}) \tag{4.4}$$

(see [22, Ch. III]). With the choice (4.3) we define

$$T_N(g) = T(g) - \sum_{n=0}^{N}\frac{1}{n!}T_n g^{(n)}(0) \qquad (N=1,2,\ldots). \tag{4.5}$$

LEMMA 4.1. *For all* $g \in H^\infty(\mathcal{H})$ *and* $n = 0,1,2,\ldots$, *we have*

$$|T_N(g)| \le \|g\|_\infty \int_{-1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^\lambda \left(\frac{1}{2\pi}\int_0^{2\pi}\frac{d\theta}{|e^{i\theta}-\sigma|}\right)\frac{|\sigma|^{N+1}d\sigma}{1-\sigma^2}. \tag{4.6}$$

Proof. We note that by (4.1) and (4.4)

$$\begin{aligned} T(g) &= \int_{-1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^A\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{g(e^{i\theta})d\theta}{1-e^{-i\theta}\sigma}\right)\frac{d\sigma}{1-\sigma^2} \\ &= \sum_{n=0}^{N}\int_{-1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^A\left(\frac{1}{2\pi}\int_0^{2\pi} g(e^{i\theta})e^{-in\theta}\sigma^n d\theta\right)\frac{d\sigma}{1-\sigma^2} \\ &\quad + \int_{-1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^A\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{g(e^{i\theta})e^{-i(N+1)\theta}\sigma^{N+1}}{1-e^{-i\theta}\sigma}d\theta\right)\frac{d\sigma}{1-\sigma^2}, \end{aligned}$$

and that, by (4.3) and (4.5),

$$\begin{aligned} |T_N(g)| &= \left|\int_{-1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^A\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{g(e^{i\theta})e^{-i(N+1)\theta}\sigma^{N+1}}{1-e^{-i\theta}\sigma}d\theta\right)\frac{d\sigma}{1-\sigma^2}\right| \\ &\le \int_{-1}^{0}\left|\left(\frac{1+\sigma}{1-\sigma}\right)^A\right|_{\mathrm{op}}\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{|g(e^{i\theta})|d\theta}{|1-e^{-i\theta}\sigma|}\right)\frac{|\sigma|^{N+1}d\sigma}{1-\sigma^2} \\ &\le \int_{-1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^\lambda\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{\|g\|_\infty d\theta}{|e^{i\theta}-\sigma|}\right)\frac{|\sigma|^{N+1}d\sigma}{1-\sigma^2}. \end{aligned}$$

This establishes (4.6). □

LEMMA 4.2. *For all* $N = 0, 1, 2, \ldots$, *we have*

$$\int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\lambda} \left(\frac{1}{2\pi}\int_0^{2\pi} \frac{d\theta}{|e^{i\theta}-\sigma|}\right) \frac{|\sigma|^{N+1}d\sigma}{1-\sigma^2} \tag{4.7}$$
$$\leq \frac{8}{\pi}\int_{-1}^{0}(1+\sigma)^{\lambda-1}\left(\log\frac{2}{1+\sigma}\right)|\sigma|^{N+1}d\sigma.$$

Proof. Since

$$\int_0^{2\pi}\frac{d\theta}{|e^{i\theta}-\sigma|} \leq 4\int_{\pi/2}^{\pi}\frac{d\theta}{|e^{i\theta}-\sigma|} \qquad (-1<\sigma\leq 0),$$

the left-hand side of (4.7) is bounded from above by

$$\frac{2}{\pi}\int_{-1}^{0}(1+\sigma)^{\lambda-1}\left(\int_{\pi/2}^{\pi}\frac{d\theta}{|e^{i\theta}-\sigma|}\right)\frac{|\sigma|^{N+1}d\sigma}{(1-\sigma)^{\lambda+1}}$$
$$\leq \frac{2}{\pi}\int_{-1}^{0}(1+\sigma)^{\lambda-1}|\sigma|^{N+1}\left(\int_{\pi/2}^{\pi}\frac{d\theta}{|e^{i\theta}-\sigma|}\right)d\sigma$$

and so it remains to check that

$$\int_{\pi/2}^{\pi}\frac{d\theta}{|e^{i\theta}-\sigma|} \leq 4\,\log\frac{2}{1+\sigma}\,. \tag{4.8}$$

Indeed,

$$\int_{\pi/2}^{\pi}\frac{d\theta}{|e^{i\theta}-\sigma|} = \int_0^{\pi/2}\frac{d\theta}{|e^{i\theta}+\sigma|} = \int_0^1 \frac{dy}{\sqrt{(1-y^2)(1+\sigma^2+2\sigma y)}}$$
$$\leq \int_0^1 \frac{dy}{\sqrt{(1-y)(1+\sigma^2+2\sigma y)}} = \int_0^1 \frac{2dz}{\sqrt{(1+\sigma)^2-2\sigma z^2}}$$
$$\leq 2\sqrt{2}\int_0^1 \frac{dz}{1+\sigma+z\sqrt{2|\sigma|}} = \frac{2}{\sqrt{|\sigma|}}\log\frac{1+\sigma+\sqrt{2|\sigma|}}{1+\sigma}$$

and this last function is bounded from above by $4\log[(3/2)(1+\sigma)^{-1}] \leq 4\log[2(1+\sigma)^{-1}]$ if $-1 < \sigma \leq -1/4$ and by $8\sqrt{2}/3 \leq 4\log 8/3 \leq 4\log[2(1+\sigma)^{-1}]$ if $-1/4 \leq \sigma \leq 0$. □

We can now pass to the main result of this section. To this aim we define the Nth error of our interpolation of $T(g)$ by $\sum_{n=0}^{N}\frac{1}{n!}T_n g^{(n)}(0)$ (where T_n are given in (4.3)) by

$$\epsilon_N = \sup\{|T_N(g)|\ ,\ g\in H^\infty(\mathcal{H})\ ,\ \|g\|\leq 1\}\,. \tag{4.9}$$

Then we have the following

PROPOSITION 4.3. *For the error* ϵ_N, *defined in* (4.9), *the following estimate holds*

(4.10)
$$\epsilon_N \leq \frac{8}{\pi}\left(\frac{1}{\lambda}+\log\left(1+\frac{N+1}{\lambda}\right)+2\right)\frac{\lambda^\lambda(N+2)^{N+2}}{(\lambda+N+1)^{\lambda+N+2}} \quad \textit{for all } N=0,1,2,\ldots.$$

Proof. By virtue of Lemmas 4.1 and 4.2, we must only estimate from above the integral

$$I_N(\lambda) = \int_0^1 (1-\sigma)^{\lambda-1} \left(\log \frac{2}{1-\sigma}\right) \sigma^{N+1} d\sigma.$$

For that purpose we integrate by parts and using the Bernouilli's beta function B

$$B(x,y) = \int_0^1 s^{x-1}(1-s)^{y-1} ds, \qquad x, y > -1$$

we write

$$\begin{aligned}
I_N(\lambda) &= \frac{1}{\lambda} \int_0^1 (1-\sigma)^{\lambda} \left(\sigma^{N+1} \log \frac{2}{1-\sigma}\right)' d\sigma \\
&= \frac{N+1}{\lambda} I_{N-1}(\lambda+1) + \frac{1}{\lambda} \int_0^1 (1-\sigma)^{\lambda-1} \sigma^{N+1} d\sigma \\
&= \frac{N+1}{\lambda} I_{N-1}(\lambda+1) + \frac{1}{\lambda} B(\lambda, N+2) \\
&= \frac{1}{\lambda} B(\lambda, N+2) + \frac{N+1}{\lambda(\lambda+1)} B(\lambda+1, N+1) + \frac{(N+1)N}{\lambda(\lambda+1)} I_{N-2}(\lambda+2) \\
&\quad \text{(by reiterating)} \\
&= \frac{1}{\lambda} B(\lambda, N+2) + \frac{(N+1)}{\lambda(\lambda+1)} B(\lambda+1, N+1) \\
&\quad + \frac{(N+1)N}{\lambda(\lambda+1)(\lambda+2)} B(\lambda+2, N) \\
&\quad + \cdots + \frac{(N+1)\cdots 3}{\lambda(\lambda+1)\cdots(\lambda+N-1)} B(\lambda+N-1, 3) \\
&\quad + \frac{(N+1)\cdots 2}{\lambda(\lambda+1)\cdots(\lambda+N-1)} I_0(\lambda+N) \\
&= \frac{(N+1)!}{\lambda^2(\lambda+1)\cdots(\lambda+N+1)} + \frac{(N+1)!}{\lambda(\lambda+1)^2\cdots(\lambda+N+1)} \\
&\quad + \frac{(N+1)!}{\lambda(\lambda+1)(\lambda+2)^2\cdots(\lambda+N+1)} \\
&\quad + \cdots + \frac{(N+1)!}{\lambda(\lambda+1)\cdots(\lambda+N-1)^2(\lambda+N)(\lambda+N+1)} \\
&\quad + \frac{(N+1)!}{\lambda(\lambda+1)\cdots(\lambda+N-1)} I_0(\lambda+N).
\end{aligned}$$

Then

$$\begin{aligned}
I_0(\mu) &= \int_0^1 (1-\sigma)^{\mu-1} \sigma \log \frac{2}{1-\sigma} d\sigma \\
&= \frac{1}{\mu} \int_0^1 (1-\sigma)^{\mu} \log \frac{2}{1-\sigma} d\sigma + \frac{1}{\mu} \int_0^1 (1-\sigma)^{\mu-1} d\sigma
\end{aligned}$$

$$= \frac{1}{\mu(\mu+1)} \int_0^1 \log \frac{2}{1-\sigma} (-(1-\sigma)^{\mu+1})' d\sigma$$
$$+ \frac{1}{\mu} \int_0^1 (1-\sigma)^{\mu-1} d\sigma$$
(after integrating by parts the first integral
and integrating the second integral)
$$= \frac{\log 2}{\mu(\mu+1)} + \frac{1}{\mu(\mu+1)^2} + \frac{1}{\mu^2}.$$

Thus

$$I_N(\lambda) = \frac{(N+1)!}{\lambda(\lambda+1)\cdots(\lambda+N)}$$
$$\cdot \left\{ \frac{1}{(\lambda+N+1)} \left(\frac{1}{\lambda} + \frac{1}{\lambda+1} + \cdots + \frac{1}{\lambda+N-1} \right) \right.$$
$$\left. + \frac{\log 2}{\lambda+N+1} + \frac{1}{\lambda+N} + \frac{1}{(\lambda+N+1)^2} \right\}$$
$$= \frac{(N+1)!}{\lambda(\lambda+1)\cdots(\lambda+N+1)} \left(\frac{1}{\lambda} + \frac{1}{\lambda+1} + \cdots + \frac{1}{\lambda+N+1} + \log 2 + 1 \right).$$

The sum

$$\frac{1}{\lambda+1} + \cdots + \frac{1}{\lambda+N+1}$$

is bounded by the integral

$$\int_\lambda^{\lambda+N+1} \frac{dx}{x} = \log \left(1 + \frac{N+1}{\lambda} \right),$$

and hence

$$I_N(\lambda) \leq \frac{(N+1)!}{\lambda(\lambda+1)\cdots(\lambda+N)} \left[\frac{1}{\lambda} + \log \left(1 + \frac{N+1}{\lambda} \right) + 2 \right] \frac{1}{\lambda+N+1}. \tag{4.11}$$

The logarithm of the term in (4.11) in front of the square bracket is

$$-\sum_{j=1}^{N+1} \log \left(1 + \frac{\lambda-1}{j} \right).$$

The sum is bounded from below by the integral

$$\int_1^{N+2} \log \left(1 + \frac{\lambda-1}{x} \right) dx = \int_1^{N+2} \log(x+\lambda-1) dx - \int_1^{N+2} \log x dx$$
$$= (\lambda+N+1)\log(\lambda+N+1) - \lambda \log \lambda - (N+2)\log(N+2)$$
$$= \log \frac{(\lambda+N+1)^{\lambda+N+1}}{\lambda^\lambda (N+2)^{N+2}}.$$

Introducing the last estimate in (4.11) we obtain (4.10). □

5. Approximation of attractors by analytic sets (II). We now return to the approximation of attractors by analytic sets. The third and last approximation that we construct here utilizes the conformal mapping described in §2 of the band of analyticity Δ (see (2.6)) onto the unit disk $\mathbb{D}$ (see (2.9)).

As before we denote by $\Lambda = \lambda_{m+1}$ an eigenvalue of A and by $Q = Q_m$ the projector in H onto the space spanned by the eigenvectors corresponding to the eigenvalues larger than or equal to λ_{m+1} ; $P = P_m = I - Q_m$ is the orthogonal projector onto the space spanned by $w_1, \ldots, w_m$ (see (1.1)).

We set $p_m = P_m u$, $q_m = Q_m u$ or dropping the indices m , $p = Pu$, $q = Qu$.

By projecting (1.2) onto the spaces PH and QH we obtain the system of equations satisfied by p and q written in (3.1) and (3.2). We recall also formula (3.3)

$$q(0) = \int_{-\infty}^{0} e^{\tau A} QR(u(\tau))d\tau, \tag{5.1}$$

which is valid at any point $u_0 = u(0)$ on the attractor. With the change of variable (2.10) we find the relations similar to (2.11) and (2.12):

$$q(0) = -2\delta \int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta A} Qg(\sigma)\frac{d\sigma}{1-\sigma^2}, \tag{5.2}$$

where, as well as in the sequel, we denote by δ the quantity $2\delta_0/\pi$ denoted δ' in §2, and

$$g(\sigma) = f(\psi(\sigma)) = R(u(\psi(\sigma))). \tag{5.3}$$

Our aim now is to show that the right-hand side of (5.2) is the sum of an analytic function of $u(0)$ and a small term; furthermore, this analytic function of $u(0)$ is a polynomial map of the domain of a large enough power of A into H. For that purpose we use the results in §4 with g as in (5.3) and A and λ replaced by δQA and $\delta\lambda_{m+1} = \delta\Lambda$. Then the right-hand side of (5.2) is $-2\delta QT(g)$ and, for N arbitrary we set

$$\begin{aligned} K_N(u_0) &= 2\delta T_N(g) \\ &= 2\delta \sum_{n=0}^{N} \frac{1}{n!} T_n g^{(n)}(0), \end{aligned} \tag{5.4}$$

where

$$T_n = \int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta A} \frac{\sigma^n}{1-\sigma^2} d\sigma, \qquad n = 0, 1, 2, \ldots . \tag{5.5}$$

We note that

$$Q(u_0 + K_N(u_0)) = 2\delta Q(T(g) - T_N(g)) = 2\delta(T(Qg) - T_N(Qg))$$

and that, by applying Proposition 4.3 to $T(Qg)$ and $T_N(Qg)$, the role of the $\mathcal{H}$ played now by QH, we obtain the estimate

$$\begin{aligned} |T(Qg) - T_N(Qg)| \leq & \frac{8}{\pi}\left(\frac{1}{\delta\Lambda} + \log\left(1 + \frac{N+1}{\delta\Lambda}\right) + 2\right) \cdot \\ & \frac{(\delta\Lambda)^{\delta\Lambda}(N+2)^{N+2}}{(\delta\Lambda + N + 1)^{\delta\Lambda+N+2}} \cdot \|g\|_\infty, \end{aligned} \tag{5.6}$$

where

$$\|g\|_\infty = \sup\{|g(z)|\ ,\ z \in \mathbb{D}\} \tag{5.7}$$

is simply majorized by ρ_0 (see (3.8)).

Now we make $K_N(u_0)$ explicit:

$$T_n = \mathcal{L}_n(\delta A), \tag{5.8}$$

where for $r > 0$

$$\mathcal{L}_n(r) = \int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^r \frac{\sigma^n}{1-\sigma^2} d\sigma\ ,$$

or setting $s = (1+\sigma)/(1-\sigma)$:

$$\mathcal{L}_n(r) = \frac{(-1)^n}{2} \int_0^1 s^{r-1} \left(\frac{1-s}{1+s}\right)^n ds. \tag{5.9}$$

Using the beta function B and the hypergeometric function F,

$$F(a,b;c;z) = \sum_{n=0}^{\infty} \frac{a(a+1)\cdots(a+n-1)b(b+1)\cdots(b+n-1)}{n!c(c+1)\cdots(c+n-1)} z^n,$$

we see that (see [13, Ch. 4, §3, eq. (4)])

$$\mathcal{L}_n(r) = \frac{(-1)^n}{2} B(r, n+1) F(r, n; r+n+1; -1)$$

which is the same, thanks to Kummer's relation (see [13, Ch. 4, §10, eq. (9)]), as

$$\mathcal{L}_n(r) = \frac{(-1)^n}{2^{n+1}} B(r, n+1) F\left(n, n+1; r+n+1; \frac{1}{2}\right)$$

or expressing B in terms of the gamma function (see [13, Ch. 3, §4])

$$\begin{aligned} \mathcal{L}_n(r) &= \frac{(-1)^n}{2^{n+1}} \frac{\Gamma(r)\Gamma(n+1)}{\Gamma(r+n+1)} F\left(n, n+1; r+n+1; \frac{1}{2}\right) \\ &= \frac{(-1)^n}{2^{n+1}} \frac{n!}{r(r+1)\cdots(r+n)} F\left(n, n+1; r+n+1; \frac{1}{2}\right). \end{aligned} \tag{5.10}$$

Alternatively we can write

$$\begin{aligned} \mathcal{L}_n(r) &= (-1)^n \sum_{j=0}^{\infty} \mathcal{L}_{n,j}(r) \\ \mathcal{L}_{n,j}(r) &= \frac{(n+j)!}{r(r+1)\cdots(r+n+j)} \frac{n(n+1)\cdots(n+j-1)}{j!} \frac{1}{2^{n+j+1}}. \end{aligned}$$

Hence

$$\begin{aligned} \mathcal{L}_n(\delta AQ) = \frac{(-1)^n}{2^{n+1}} F\left(n, n+1; \delta AQ + (n+1)I; \frac{1}{2}\right) n!(\delta AQ)^{-1} \\ \cdot (\delta AQ + I)^{-1} \cdots (\delta AQ + nI)^{-1} \end{aligned} \tag{5.11}$$

or

$$\mathcal{L}_n(\delta AQ) = (-1)^n \sum_{j=0}^{\infty} \mathcal{L}_{n,j}(\delta AQ), \tag{5.12}$$

where

$$\begin{aligned}\mathcal{L}_{n,j}(\delta AQ) = \frac{(n+j)!}{2^{n+j+1}} \cdot \frac{n(n+1)\cdots(n+j-1)}{j!}(\delta AQ)^{-1} \\ \cdot (\delta AQ + I)^{-1}\cdots(\delta AQ + (n+j)I)^{-1}.\end{aligned} \tag{5.13}$$

Although $\mathcal{L}_n$ is expressed in (5.12) as the sum of a series, we can observe that this series is rapidly convergent as

$$|\mathcal{L}_{n,j}(\delta AQ)|_{\mathrm{op}} \leq \frac{(n+j)!}{2^{n+j+1}} \frac{n(n+1)\cdots(n+j-1)}{j!} \frac{1}{\delta\Lambda(\delta\Lambda+1)\cdots(\delta\Lambda+n+j)}$$

and this bound converges rapidly to 0 if n is fixed $\leq N$ and $j \to \infty$. Therefore the formulas (5.12) and (5.13) turn out to be better suited for the numerical computation of $QT_n = \mathcal{L}_n(AQ)$ than (5.5).

Due to (1.16), (2.10), and (2.12) we express $g(z)$ as

$$\begin{aligned} g(z) &= \sum_{h=0}^{\infty} \frac{(2\delta)^h}{h!}\left(z + \frac{z^3}{3} + \frac{z^5}{5} + \cdots\right)^h f_h \\ &= \sum_{n=0}^{\infty} z^n \sum_{h=0}^{n} \frac{(2\delta)}{h!} C_{nh} f_h, \end{aligned} \tag{5.14}$$

where $C_{00} = 1$ and $C_{nh} = 0$ if $n - h$ is odd or if $n < h$; otherwise if $n - h \geq 0$ and $n - h$ is even,

$$\begin{aligned} C_{nh} &= \sum_{\substack{i_1,\ldots,i_k \geq 0 \\ i_1+\cdots+i_k = \frac{n-h}{2}}} \frac{1}{(2i_1+1)} \cdots \frac{1}{(2i_k+1)}, \\ &= \sum \binom{h}{\alpha_1, \ldots, \alpha_j} \frac{1}{1^{\alpha_0}} \frac{1}{3^{\alpha_1}} \cdots \frac{1}{(2j+1)^{\alpha_j}}, \\ &= 1 + \sum \binom{h}{\alpha_1, \ldots, \alpha_j} \frac{1}{3^{\alpha_1}} \cdots \frac{1}{(2j+1)^{\alpha_j}}. \end{aligned} \tag{5.15}$$

The last two sums are extended to the $j \geq 0$, $\alpha_1, \ldots, \alpha_j \geq 0$ such that

$$\begin{cases} \alpha_0 + \cdots + \alpha_j = h \ , \\ \alpha_1 + 2\alpha_2 + \cdots + j\alpha_j = \frac{n-h}{2} \ , \end{cases} \tag{5.16}$$

in the first case, while in the second case

$$\begin{cases} 1 \leq \alpha_1 + \cdots + \alpha_j \leq h, \\ \alpha_1 + 2\alpha_2 + \cdots + j\alpha_j \leq \frac{n-h}{2} \ . \end{cases} \tag{5.17}$$

It is clear that

$$g_n = g^{(n)}(0) = \sum_{h=0}^{n} \frac{n!}{h!} (2\delta)^h C_{nh} f_h \ . \tag{5.18}$$

Finally

$$K_N(u_0) = \sum_{n=0}^{N} \sum_{h=0}^{n} \frac{(2\delta)^{h+1}}{h!} C_{nh} \mathcal{L}_n(\delta A Q) f_h \tag{5.19}$$

$$= \sum_{h=0}^{N} \frac{(2\delta)^{h+1}}{h!} \mathcal{K}_{N,h}(\delta A Q) f_h,$$

where

$$\mathcal{K}_{N,h}(\alpha) = \sum_{n=h}^{N} C_{nh} \mathcal{L}_n(\alpha). \tag{5.20}$$

Reintroducing the index m we arrive to the conclusion of this section.

THEOREM 5.1. *The hypotheses are those of* §§1 *and* 2, *in particular* (1.3)–(1.6), (1.10)–(1.13), (1.25), (2.1)–(2.4), *and* (2.6). *Let* $\mathcal{K}_N$ *be the function of* u_0 *defined in* (5.4) (*or more precisely* $K_N(u_0)$ *being given by* (5.4), (5.5), (5.13), (5.18), (5.19), (5.20), *and* (1.17)–(1.19)).

Then $\mathcal{K}_N$ *is a compact polynomial map from* $\mathcal{D}(A^{N+1})$ *into* H *and for every* $\epsilon > 0$, *we have*

$$|Q_m(u_0 + K_N(u_0))| \leq \epsilon \quad \forall u_0 \in \mathcal{A}\ , \tag{5.21}$$

whenever either

$$\frac{\delta\Lambda}{N+1} \geq \frac{1}{\log 2} \log \frac{192}{\pi^2} \frac{\delta_0 \rho_0}{\epsilon} \quad \textit{and} \quad \frac{\delta\Lambda}{N+1} \geq 1 \tag{5.22a}$$

or

$$\frac{N+1}{\delta\Lambda} \geq \left(\frac{2}{\alpha} \left(1 + \log^+ \frac{1}{\alpha} \right) 3 \log(\beta e) \right)^{1/\alpha} \quad \textit{and} \quad \frac{N+1}{\delta\Lambda} \geq 1\ , \tag{5.22b}$$

where

$$\Lambda = \lambda_{m+1}, \quad \alpha = \delta\Lambda = \frac{2\delta_0 \lambda_{m+1}}{\pi}, \quad \beta = \frac{128}{\pi^2} \frac{\rho_0 \delta_0}{\epsilon} e^{2+\alpha} \tag{5.22c}$$

and $\log^+ = \max\{0, \log\}$.

Proof. Using (5.6) and the relation prior to (5.6), we bound the H-norm of $2\delta Q_m[T_N(g) - T(g)]$ by

$$\rho_0 \frac{16\delta}{\pi} \left(\frac{1}{\delta\Lambda} + \log\left(1 + \frac{N+1}{\delta\Lambda} \right) + 2 \right) \frac{(\delta\Lambda)^{\delta\Lambda} (N+2)^{N+2}}{(\delta\Lambda + N + 1)^{\delta\Lambda + N + 2}} \tag{5.23}$$

(where $\Lambda = \lambda_{m+1}$). In order to estimate suitable values of m and N we note that the expression (5.23) is bounded by

$$\rho_0 \frac{16\delta}{\pi} \left(\frac{1}{\delta\Lambda} + \log\left(1 + \frac{N+1}{\delta\Lambda} \right) + 2 \right) \left(\frac{1}{2} \right)^{N+1} \leq \frac{96}{\pi} \rho_0 \delta \left(\frac{1}{2} \right)^{N+1} \tag{5.24}$$

if $\delta\Lambda/(N+1) \geq 1$, and by

$$\rho_0 \frac{64\delta}{\pi}\left(\frac{1}{\delta\Lambda} + \log\left(1+\frac{N+1}{\delta\Lambda}\right)+2\right)\left(\frac{\delta\Lambda}{\delta\Lambda+N+1}\right)^{\delta\Lambda} \tag{5.25}$$

if $\delta\Lambda/(N+1) \leq 1$. If the relations in (5.22a) are satisfied, then (5.21) readily follows from the fact that the right-hand side of (5.24) is bounded by ϵ. If $\xi = (N+1)/\delta\Lambda \geq 1$, then (5.25) is bounded by (with the notation of (5.22c))

$$\frac{128}{\pi^2}\rho_0\delta_0 e^{2+\alpha}\frac{\log[e^{2+\alpha}(1+\xi)]}{[e^{2+\alpha}(1+\xi)]^{\delta\Lambda}} = \epsilon\beta\frac{\log\eta}{\eta^\alpha}, \quad \text{where } \eta = e^{2+\alpha}(1+\xi).$$

The fact that (5.21) holds if $\xi = \delta\Lambda/(N+1)$ satisfies the two relations in (5.22b), now follows from the following elementary lemma.

LEMMA 5.2. *Let $\alpha > 0$, $\beta \geq e$ and $\theta(\eta) = \beta\eta^{-\alpha}\log\eta$ for $\eta \geq 1$. Then*

$$\eta \geq \eta_1 = \left(\frac{2}{\alpha}\left(1+\log^+\frac{1}{\alpha}\right)\beta\log(\beta e)\right)^{1/\alpha} \Longrightarrow \theta(\eta) \leq 1. \tag{5.26}$$

Indeed it is easy to check that the function $\eta^\alpha(1-\theta(\eta))$ is positive for $\eta = \eta_1$ and its derivative is positive for $\eta \geq \eta_1$. □

Remark 5.3. It will be subsequently shown that the set defined by the equation

$$Q_m(u_0 + K_N(u_0)) = 0, \tag{5.27}$$

is analytic in H. Theorem 5.1 expresses the fact that the attractor lies in its neighborhood

$$\{u_0 \in M\ ,\ |Q_m(u_0 + K_N(u_0))| < \epsilon\}. \tag{5.28}$$

For Theorem 5.1 to be of interest it is desirable that m and N are not too large and that the set (5.21) is not too thick. The question of the thickness of (5.28) will also be addressed below.

We emphasize the major difference between Theorems 3.1 and 3.5 on one side and Theorem 5.1 on the other side. Namely the first two theorems are relevant only if $\delta\Lambda$ is large enough, while in the latter one, this restriction is not necessary. Unlike the I_N and J_N-approximations, the K_N-approximations can be improved by just increasing N. We shall return to this point at the end of this work.

Exponential attraction. We will now estimate the time of absorbtion into the set (5.27) of the orbits that do not lay on the attractor $\mathcal{A}$.

THEOREM 5.4. *The hypotheses are those of Theorem* 5.1 *and m and N are chosen as in the proof of Theorem* 5.1 *with ϵ replaced by $\epsilon/2$.*

There exists a constant κ such that

$$|Q_m(u(t) + K_N(u(t)))| \leq \frac{\epsilon}{2} + \kappa\exp(-t\lambda_{m+1}) \quad \forall t \geq t_0, \tag{5.29}$$

where $u(t) = S(t)u_0$, the constants κ and t_0 depending on the equation and boundedly on $|u_0|$. In particular all orbits enter the set (5.21) *at a time depending logarithmically on ϵ and boundedly on $|u_0|$.*

Proof. As in the proof of Theorems 3.3 and 3.6, we can assume that u_0 belongs to the absorbing set $\mathcal{B}_0$. Let t_0 be fixed, $t_0 = \delta_0(r_1)$. We recall that $u(t) = S(t)u_0$ is analytic in t in the region $\Delta_1(r_1)$ defined by $\mathrm{Re}\zeta \geq t_0$, $|\mathrm{Im}\zeta| \leq t_0$ (see (1.12) and (2.4)). Proceeding as for (2.5), we obtain by integration of (3.2) between t_0 and $t_1 \geq 2t_0$:

$$q(t_1) = e^{-(t_1-t_0)A}q(t_0) - \int_{t_0}^{t_1} e^{\tau A}QR(u(\tau))d\tau. \tag{5.30}$$

Using a transformation similar to (2.10)

$$\sigma = \varphi(t - t_1) = \frac{e^{(t-t_1)/\delta} - 1}{e^{(t-t_1)/\delta} + 1},$$

(5.30) yields (compare to (2.12) and (3.19)):

$$q(t_1) = e^{-(t_1-t_0)A}q(t_0) - 2\delta \int_{\sigma_1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta A} g(\sigma)\frac{d\sigma}{1-\sigma^2}, \tag{5.31}$$

where $\sigma_1 = \varphi(t_0 - t_1)$ and $g(\sigma) = R(u(t_1 + \psi(\sigma)))$, ψ as in (2.11). We proceed as in §4 and in the proof of Theorem 5.1, and write

$$T^{(1)}(g) = \int_{\sigma_1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta A} g(\sigma)\frac{d\sigma}{1-\sigma^2},$$

$$T_N^{(1)}(g) = \sum_{n=0}^{N} \left(\int_{\sigma_1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta A} \frac{\sigma^n}{1-\sigma^2}d\sigma\right) \frac{g^{(n)}(0)}{n!},$$

and

$$\begin{aligned} q(t_1) + QK_N(u(t_1)) = e^{(t_0-t_1)A}q(t_0) &+ 2\delta Q(T_N(g) - T_N^{(1)}(g)) \\ &+ 2\delta Q(T_N^{(1)}(g) - T^{(1)}(g)). \end{aligned} \tag{5.32}$$

The H-norm of the term $2\delta Q(T_N(g) - T_N^{(1)}(g))$ is bounded as follows:

$$\begin{aligned} 2\delta|Q(T_N(g) - T_N^{(1)}(g))| &\leq \sum_{n=0}^{N} \left(2\delta \int_{-1}^{\sigma_1} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta\Lambda} \frac{d\sigma}{1-\sigma^2}\right) \frac{|g^{(n)}(0)|}{n!} \\ &\quad \text{(returning to the } \tau \text{ variable)} \\ &\leq \sum_{n=0}^{N} \left(\int_{-\infty}^{t_0-t_1} e^{\tau\Lambda}d\tau\right) \frac{|g^{(n)}(0)|}{n!} \\ &= \sum_{n=0}^{N} \frac{e^{(t_0-t_1)\Lambda}}{\Lambda} \frac{|g^{(n)}(0)|}{n!}. \end{aligned}$$

The function g is analytic in the image of the region $\Delta_1(r_1)$ by the conformal mapping φ in (2.9), and it is bounded there by the constant ρ_0 defined in (3.8). As is easily shown the image of $\Delta_1(r_1)$ by φ contains the circle centered at 0 of radius $|\sigma_1| = -\varphi(t_0 - t_1)$. Hence by Cauchy's formula

$$\frac{1}{n!}|g^{(n)}(0)| \leq \frac{\rho_0}{|\sigma_1|^n} \tag{5.33}$$

and we obtain

$$2\delta|Q(T_N^{(1)}(g)-T_N(g))| \le \rho_0 \frac{e^{(t_0-t_1)\Lambda}}{\Lambda}\sum_{n=0}^{N}\frac{1}{|\sigma_1|^n} \le \frac{(N+1)\rho_0}{\Lambda|\sigma_1|^N}e^{(t_0-t_1)\Lambda} \le \kappa_1 e^{-t_1\delta\Lambda}, \tag{5.34}$$

where

$$\kappa_1 = \frac{(N+1)\rho_0}{\Lambda|\varphi(-t_0)|^N}e^{t_0\delta\Lambda}. \tag{5.35}$$

We then estimate the term $2\delta Q(T_N^{(1)}(g)-T^{(1)}(g))$. For this term, the analog of Lemma 4.1 holds and yields

$$2\delta|Q(T_N^{(1)}(g)-T^{(1)}(g))| \le 2\delta \sup_{|z|\le|\sigma_1|}|g(z)|\int_{\sigma_1}^{0}\left(\frac{1+\sigma}{1-\sigma}\right)^{\lambda_1}\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{d\sigma}{|e^{i\theta}-\sigma|}\right)\frac{|\sigma|^{N+1}}{1-\sigma^2}d\sigma, \tag{5.36}$$

where $\sup_{|z|\le|\sigma_1|}|g(z)| \le \rho_0$. The integral in the right-hand side of (5.36) can be bounded by the integral from -1 to 0; then the estimates in Lemma 4.2 and Proposition 4.3 are valid without further change and so, under the assumptions of Theorem 5.1 the right-hand side of (5.36) is $\le \epsilon/2$. Finally

$$\|e^{-(t_1-t_0)A}q(t_0)\| \le \mu_0 e^{-(t_1-t_0)\Lambda}. \tag{5.37}$$

Introducing the estimates (5.34), (5.36), and (5.37) into (5.32) we obtain

$$|Q_m(S(t)u_0+K_N(S(t)u_0))| \le \frac{\epsilon}{2}+\kappa e^{-t\Lambda} \quad \forall t \ge 2t_0, \tag{5.38}$$

where

$$\kappa = \kappa_1 + \mu_0 e^{t_0\Lambda}.$$

This concludes the proof of Theorem 5.4. □

Remark 5.5. We shall show in §7 that for λ_{m+1} large enough, the set defined by (5.27) is a graph above P_mH. However the K_N-approximation can be improved just by keeping Λ fixed and by increasing N (see (5.22b)). Therefore the neighborhood (5.28) of that set will have the exponential attraction property, although Λ is not large enough to guarantee that the set is a graph.

Remark 5.6. It is interesting to observe that the sets defined by

$$u_0 + I_N(u_0) = 0 \text{ and } u_0 + K_N(u_0) = 0, \quad \text{respectively,} \tag{5.39}$$

contain all the stationary solutions of (1.2), i.e., the points u_* such that

$$Au_* + R(u_*) = 0.$$

For I_N this follows readily from (3.28) with $u_0 = u_*$. For K_N we observe that if u_0 is a stationary point, then $u_n = 0$, $f_n = 0$, $g_n = 0$, for all $n \ge 1$, and

$$Au_0 + f_0 = 0.$$

Then, by (5.4) and (5.5),

$$\begin{aligned} K_N(u_0) &= 2\delta T_0 g_0 = -2\delta T_0 A u_0 \\ &= -2\delta A \int_{-1}^{0} \left(\frac{1+\sigma}{1-\sigma}\right)^{\delta A} \frac{d\sigma}{1-\sigma^2} u_0 \\ &= -\left(\int_{-\infty}^{0} e^{tA} dt\right) A u_0 = -u_0, \end{aligned}$$

i.e., $u_0 + K_N(u_0) = 0$.

A weaker assertion is valid for J_N:

$$u_0 + J_N(u_0) = u_0 + S_0 f_0, \qquad S_0 = (I - e^{-\delta A})\, A^{-1},$$

so that

$$u_0 + J_N(u_0) = u_0 + (I - e^{-\delta A})\, I_0(u_0) = e^{-\delta A} u_0 \tag{5.40}$$

for all $N = 0, 1, 2, \ldots$.

6. Utilization of more regular spaces. In the previous sections we have shown how to approximate the attractor $\mathcal{A}$ by smooth sets in the topology of H. In this section we want to extend the methodology to more regular topologies, and we shall study the same approximation problems in the spaces $\mathcal{D}(A^k)$, $k \geq 1$. We shall establish also some properties of the functions I_N, J_N, K_N in spaces $\mathcal{D}(A^k)$. Besides their intrinsic interest these results will be needed in §7 for proving the analytic structure of the approximating sets.

First we should recall that (1.2) in the spaces $\mathcal{D}(A^k)$ is assumed to satisfy all the regularity conditions introduced in the last part of §1.

Then we have the following easy fact.

PROPOSITION 6.1. *The hypotheses on the equation* (1.2) *are those in* §1. *Then for all* $r > 0$ *and* $= 0, 1, 2, \ldots$ *there exists* $\mu_k(r) < \infty$ *such that*

$$|A^k u(\zeta)| \leq \mu_k(r) \quad \forall \zeta \in \Delta_1(r), \quad \forall |A^{1/2} u(0)| \leq r. \tag{6.1}$$

Proof. If there is no finite $\mu = \mu_k(r)$ satisfying (6.1), we must have a sequence $u_j \in \mathcal{D}(A^{1/2})$, $|A^{1/2} u_j| \leq r$, and $t_j \geq \delta_0 = \delta_0(r)$ such that the solution $u_j(t)$ with initial data $u_j(0) = u_j$ will satisfy $|A^k u_j(t_j)| \to \infty$. Since the sequence $\{u_j\}$ is relatively compact in H, we can assume without loss of generality that u_j is convergent in H to some u_0; obviously $u_0 \in \mathcal{D}(A^{1/2})$ and $|A^{1/2} u_0| \leq r$. If $\{t_j\}$ contains a convergent subsequence, say $t_{j'} \to t_0 \geq \delta_0$, then $S(t_{j'}) u_{j'} \to S(t_0) u_0$ in $\mathcal{D}(A^k)$ and therefore $|A^k S(t_{j'}) u_{j'}| \to |A^k S(t_0) u_0| < \infty$; a contradiction. So $t_j \to \infty$. But then for j large enough, we will have $v_j = S(t_j - \delta_0) u_j \in \mathcal{B}_0$ and therefore $|A^{1/2} v_j| \leq r_1$ (see (2.4)) for all j large enough. Then by replacing u_j and t_j for all j large enough, with v_j and δ_0 we are in the first case considered above which, as we already noticed, leads to a contradiction. □

COROLLARY 6.2. *Equation* (1.2) *has an absorbing ball* $\mathcal{B}_k$ *in each* $\mathcal{D}(A^k)$, $k = 1, 2, \ldots$.

COROLLARY 6.3. *The global attractor* $\mathcal{A}$ *attracts all bounded sets in* H *in the* $\mathcal{D}(A^k)$*-norm,* $k = 1, 2, \ldots$.

Proof. Corollary 6.2 readily follows by noticing that if $\mathcal{B}_k$ denotes the ball in $\mathcal{D}(A^k)$ centered at the origin and of radius $\mu_k = \mu_k(r_1)$, then by (6.1),

$$|A^k S(t)u_0| \leq \mu_k = \mu_k(r_1) \quad \forall t \geq \delta_0 = \delta_0(r_1), \quad \forall u_0 \in \mathcal{B}_1. \tag{6.2}$$

Corollary 6.3 now results easily from the interpolation relation

$$\begin{aligned}\inf_{v\in\mathcal{A}} |A^k(S(t)u_0 - v)| &\leq \inf_{v\in\mathcal{A}} |A^{k+1}(S(t)u_0 - v)|^{\frac{k}{k+1}} |S(t)u_0 - v|^{\frac{1}{k+1}} \\ &\leq (2\mu_{k+1})^{\frac{k}{k+1}} \left(\inf_{v\in\mathcal{A}} |S(t)u_0 - v|\right)^{\frac{k}{k+1}}\end{aligned}$$

There is no difficulty in extending Theorems 3.1, 3.3, 3.5, 3.6, 5.1, and 5.4 to $\mathcal{D}(A^k)$; i.e., these theorems remain valid with H replaced by $\mathcal{D}(A^k)$ and the norm in H $|\cdot|$, replaced by the norm in $\mathcal{D}(A^k)$, $|A^k \cdot|$. The expressions of I_N and J_N are the same, although the appropriate values of m and N may now be different and may depend on k. In the course of the proof we replace the bound $\mu_2 = \mu_2(r_1)$ in $\mathcal{D}(A)$ with the bound $\mu_{k+1} = \mu_{k+1}(r_1)$; also we replace the corresponding bound ρ_0 by

$$\rho_k = \lambda_1^{\gamma-1} \left[|A^{k+1-\gamma} R_0| + \sum_{j=1}^{\nu} c'_{j,k} (\mu_{k+1})^j \right]; \tag{6.3}$$

(see (3.8) and (1.26)).

Similarly the results of §§4 and 5 extend to the spaces $\mathcal{D}(A^k)$. The interpolation results of §4 are used with $\mathcal{H} = Q\mathcal{D}(A^k)$ and §5 proceeds in essentially the same way with

$$\|Qg\|_\infty = \sup\{|QA^k g(z)|, z \in \mathbb{D}\} \leq \rho_k$$

and $|T_N(Qg)|$ replaced by $|A^k T_N(Qg)|$. Of course, the expression of K_N is the same. □

Properties of I_N, J_N, K_N. Our aim is now to establish some properties of the functions I_N, J_N, K_N considered as mappings in the spaces $D(A^k)$. More precisely, see the following.

PROPOSITION 6.4. *The functions I_N, J_N, K_N are analytic polynomial maps from $D(A^k)$ into $D(A^{k+1-\gamma})$, for all $k \geq N+1$.*

Proof. First we recall that $f_n (n = 0, 1, \ldots)$ are analytic polynomial maps from $\mathcal{D}(A^{n+1+k})$ into $\mathcal{D}(A^{k+1-\gamma})$ for all $k = 0, 1, 2, \ldots$, $n = 0, 1, 2, \ldots$ (see Remark 1.3). We also recall

$$I_N(u_0) = \sum_{n=0}^{N} (-1)^n A^{-n-1} f_n \tag{6.4}$$

(see (3.6)),

$$J_N(u_0) = \sum_{n=0}^{N} S_n f_n = \sum_{n=0}^{N} (-1)^n \left[I - \left(\sum_{j=0}^{n} \frac{1}{j!} (\delta A)^j \right) e^{-\delta A} \right] A^{-n-1} f_n \tag{6.5}$$

(see (3.12) and (3.13)),

$$K_N(u_0) = \sum_{h=0}^{N} \frac{(2\delta)^{h+1}}{h!} \left[A^{n+1} K_{N,h}(\delta A) \right] A^{-n-1} f_h \tag{6.6}$$

(see (5.19) and (5.20)), where the operators in the square brackets are continuous on any of the spaces $\mathcal{D}(A^\alpha)$ (with norm $|A^\alpha \cdot|$) for all $\alpha \geq 0$. The statement for the operator in (6.5) is obvious, while for that in (6.6) it readily follows from (see (5.10))

$$A^{n+1}K_{N,h}(\delta A) = \sum_{n=h}^{N} C_{nh}A^{n+1}\mathcal{L}_n(A)$$

$$= \sum_{n=h}^{N} C_{nh}(-1)^n \frac{n!}{2^{n+1}} F\left(n, n+1; \delta A + (n+1)I; \frac{1}{2}\right)(I+\delta A^{-1})^{-1}\dots(I+n\delta A^{-1})^{-1}.$$

Since A^{-n-1} is a continuous map from $\mathcal{D}(A^\alpha)$ into $\mathcal{D}(A^{\alpha+n+1})$ for all $\mathcal{L} \geq 0$, then $n = 0, 1, 2, \dots, A^{-n-1}f_{n+1}$ is an analytic polynomial map from $\mathcal{D}(A^{n+1+k})$ into $\mathcal{D}(A^{n+2-\gamma+k})$ for all $n, k = 0, 1, 2, \dots$. The conclusion follows now from (6.4), (6.5), and (6.6). □

Remark 6.5. Since I_N, J_N, K_N are continuous polynomial maps from $\mathcal{D}(A^{N+1+k})$ into $\mathcal{D}(A^{N+2-\gamma+k})$ these functions are bounded on bounded sets in the corresponding spaces. In particular these functions are bounded in $\mathcal{D}(A^{N+2-\gamma+k})$ on the absorbing set $\mathcal{B}_{N+1+k}$ in $\mathcal{D}(A^{N+1+k})$ (and thus on the attractor $\mathcal{A}$ too), for all $k = 0, 1, 2, \dots$, namely

$$|A^{N+2-\gamma}F_N(u)| \leq \nu_{N+1} \quad \forall u \in \mathcal{B}_{N+1+k}\ , \tag{6.7}$$

$$|A^{N+2-\gamma}\frac{d}{du}F_N(u)| \leq \nu'_{N+1} \quad \forall u \in \mathcal{B}_{N+1+k}\ , \tag{6.8}$$

where F_N is either I_N, J_N or K_N. Of course on the attractor $\mathcal{A}$ we have

$$|A^{N+1}Q_m(u + F_N(u))| \leq \epsilon_{N+1,m} \quad \forall u \in \mathcal{A}, \tag{6.9}$$

where $\epsilon_{N+1,m}$ can be made as small as we need for appropriate m and N, particularly for m large enough. Clearly we can assume that the absorbing set $\mathcal{B}_{N+1+k}$ is the ball in $\mathcal{D}(A^{N+1+k})$ centered at the origin and of radius $2\mu_{N+1+k}$ (see (6.2)).

7. Analyticity. Our aim in this section is to show that the approximating sets, defined by

$$u + F_N(u) = 0, \quad u \in \mathcal{D}(A^{N+1+k}), \quad k = 0, 1, 2, \dots, \tag{7.1}$$

are analytic sets and that the neighborhoods

$$|AQ_m(u + F_N(u))| \leq \epsilon, \tag{7.2}$$

are not thick in general. Here F_N stands for either I_N, J_N or K_N.

We fix m, N and F_N (= either I_N, J_N or K_N). For the simplification of notation we write in this section

$$X = \mathcal{D}(A^{N+1+k}), \quad \|\cdot\| = |A^{N+1+k}\cdot|, \quad F = F_N,$$
$$u = y + z, \quad y = Pu, \quad z = Qu,$$
$$\mu = \mu_{N+1}, \quad \nu_{N+1} = \nu, \quad \nu'_{N+1} = \nu', \quad \epsilon_{N,m} = \epsilon,$$

where ν_{N+1}, ν'_{N+1} and $\epsilon_{N+1,m}$ are the bounds appearing in Remark 6.5 ((6.7)–(6.9)).

For example (6.7)–(6.9) imply

$$\|z + QF(y+z)\| \leq \frac{\nu}{\Lambda^{1-\gamma}}, \qquad u = y + z \in B(0, 2\mu), \tag{7.3}$$

$$\|Q\frac{\partial}{\partial z}F(y+z)\| \leq \frac{\nu'}{\Lambda^{1-\gamma}}, \qquad u = y + z \in B(0, 2\mu). \tag{7.4}$$

We note that with the present notation we have

$$\mathcal{A} \subset B(0, \mu), \tag{7.5}$$

where we use the notation $B(u, r)$ for the ball in X centered at u of radius r. We shall study the set

$$\mathcal{X} = \mathcal{X}_{N+1,\Lambda} = \{y + z \in B(0, 2\mu)\ ,\ z + QF(y+z) = 0\}. \tag{7.6}$$

Our first result is the following

LEMMA 7.1. *If $\Lambda > \Lambda_0$, where Λ_0 is specified below, there exists a function Φ*

$$\Phi : \{y \in PX, \|y\| \leq \mu\} \to \{z \in QX, \|z\| \leq \mu\},$$

such that

$$\mathcal{X} \cap \{\|y\| \leq \mu, \|z\| \leq \mu\} = \text{graph } \Phi. \tag{7.7}$$

Proof. Fix $y \in PX, \|y\| \leq \mu$ and define

$$G_y(z) = -QF(y+z), \quad z \in QX, \quad \|z\| \leq \mu.$$

Then

$$\begin{aligned} \|G_y(z_1) - G_y(z_2)\| &\leq \left(\sup\left\{\left\|Q\frac{\partial}{\partial z}F(y+z)\right\|, y + z \in B(0, 2\mu)\right\}\right)\|z_1 - z_2\| \\ &\quad \text{(by (7.4))} \\ &\leq \frac{\nu'}{\Lambda^{1-\gamma}}\|z_1 - z_2\| \quad \forall z_1, z_2 \in B(0, \mu). \end{aligned}$$

Moreover,

$$\|G_y(z)\| \leq \frac{\nu}{\Lambda^{1-\gamma}} \quad \forall z \in B(0, \mu),$$

hence if m is sufficiently large so that $\Lambda = \lambda_{m+1}$ satisfies

$$\frac{\nu}{\Lambda^{1-\gamma}} \leq \frac{\mu}{2}, \qquad \frac{\nu'}{\Lambda^{1-\gamma}} \leq \frac{1}{2},$$

i.e.,

$$\lambda_{m+1} = \Lambda \geq \Lambda_0 = \max\left\{\left(\frac{2\nu}{\mu}\right)^{1/(1-\gamma)}\ ,\ (2\nu')^{1/(1-\gamma)}\right\}, \tag{7.8}$$

then G_y is a Lipschitz mapping, with Lipschitz constant $\leq 1/2$, that applies $B(0,\mu)$ into itself. Thus G_y has a unique fixed point denoted $\Phi(y)$ in the ball $B(0,\mu)$. By the Picard successive approximation procedure, Φ is (even after the complexification of X) a uniform limit of polynomials; hence Φ is an analytic function of y. The proof of (7.7) is also easy:

$$\begin{aligned} y+z \in \mathcal{X} \cap \{y \in B(0,\mu) \cap PX, z \in B(0,\mu) \cap QX\} \\ = \{z = G_y(z) \text{ for some } y \in B(0,\mu) \cap PX\} \\ = \{z = \Phi(y), \text{ for some } y \in B(0,\mu) \cap PX\} \\ = \text{ graph } \Phi. \end{aligned}$$

Finally, by the fact that Φ is analytic, from

$$\Phi(y) + QF(y+\Phi(y)) = 0,$$

we deduce that

$$\begin{aligned} \Phi'(y) + Q\frac{\partial}{\partial z}F(y+z)|_{z=\Phi(y)}\Phi'(y) \\ = -Q\frac{\partial}{\partial y}F(y+z)|_{z=\Phi(y)}. \end{aligned}$$

Hence

$$\begin{aligned} \|\Phi'(y)\| &\leq \| - \left(I + Q\frac{\partial}{\partial z}F(y+z)\right)^{-1} Q\frac{\partial}{\partial y}F(y+z)|_{z=\Phi(y)}\| \\ &\leq \frac{1}{1-\frac{1}{2}}\|Q\frac{\partial}{\partial y}F(y+z)|_{z=\Phi(y)}\| \\ &\leq \frac{2\nu'}{\Lambda^{1-\gamma}} \leq 1. \qquad \square \end{aligned} \tag{7.9}$$

Remark 7.2. By increasing Λ (i.e., m), we can make

$$\sup_{\|y\| \leq \mu} \|\Phi'(y)\|$$

as small as desirable.

LEMMA 7.3. *Let $u = y + z \in \mathcal{A}$ and let Λ and Φ be as in Lemma* 7.1. *Then*

$$\|z - \Phi(y)\| \leq \frac{2\epsilon}{\Lambda^{1-\gamma}}. \tag{7.10}$$

Proof. Let $x + y = z \in \mathcal{A}$. Then $\|y\| \leq \mu$, $\|z\| \leq \mu$ and let G_y be the map of $\{z \in QX, \|z\| \leq \mu\}$ into itself, defined in the proof of Lemma 7.1. Then setting

$$z_0 = z, \qquad z_1 = G_y(z_0), \ldots, z_{n+1} = G_y(z_n), \ldots$$

we have $\Phi(y) = \lim z_n$ and

$$\begin{aligned} \|\Phi(y) - z\| &\leq \sum_{n=0}^{\infty} \|z_{n+1} - z_n\| \\ &\leq \left(\sum_{n=0}^{\infty} \frac{1}{2^n}\right) \|z_1 - z_0\| \\ &\leq 2\|G_y(z) - z\| \\ &= 2\|z + QF(y+z)\| \leq \frac{2\epsilon}{\Lambda^{1-\gamma}}. \qquad \square \end{aligned}$$

LEMMA 7.4. *Let m and N be fixed, m sufficiently large so that* (7.8) *is satisfied. Then for every $\Lambda' = \lambda_{m'+1} < \Lambda = \lambda_{m+1}$, the set*

$$\mathcal{X}' = \{u = y + z, y \in P_{m'}X, z \in Q_{m'}X, Q_{m'}(u + F(u)) = 0, \|y\|, \|z\| \leq \mu\} \tag{7.11}$$

is an analytic subset of $\mathcal{X}_{N,\Lambda}$.

Proof. Let us denote $\mathcal{X} = \mathcal{X}_{N,\Lambda}$ and let Φ be the map given in Lemma 7.1, that is $\mathcal{X} = \text{graph } \Phi$. If $u \in \mathcal{X}'$, then

$$Q_{m'}(u + F(u)) = 0 \tag{7.12}$$

and since $Q_m = Q_m Q_{m'}$

$$Q_m(u + F(u)) = 0, \tag{7.13}$$

i.e., $u \in \mathcal{X}$. Thus,

$$\mathcal{X}' \subset \mathcal{X} = \mathcal{X}_{N,\Lambda} \quad \text{for } \Lambda' \leq \Lambda. \tag{7.14}$$

Moreover,

$$\begin{aligned}(7.15)\ &(Q_{m'} - Q_m)u \\ &+ (Q_{m'} - Q_m)F(P_{m'}u + (Q_{m'} - Q_m)u + \Phi(P_{m'}u + (Q_{m'} - Q_m)u)) = 0\end{aligned}$$

and (7.13) is equivalent to (7.12). But the left-hand side of (7.15) is in the arguments

$$p' = P_{\Lambda'}u, \qquad q' = (Q_{\Lambda'} - Q_\Lambda)u$$

an analytic map of a ball in P_mX into $(Q_{m'} - Q_m)X$ and both spaces P_mX and $(Q_{m'} - Q_m)X$ are of finite dimension. Hence $\mathcal{X}'$ is an analytic set in X. □

We can now sum up all the discussion on this section with the following.

THEOREM 7.5. *If the hypotheses of* §§1, 2, *and* 6, *and* (7.8) *are valid, then*

(i) *$X_{N,\Lambda}$ is an analytic set (which is algebraic if H is of finite dimension);*

(ii) *$X_{N,\Lambda'} \subset X_{N,\Lambda}$ for $\Lambda' \leq \Lambda$;*

(iii) *$X_{N,\Lambda} = \text{graph } \Phi$, where $\Phi : B(0,\mu) \cap PX \to B(0,\mu) \cap QX$ is $\mathcal{C}$-analytic and $\|\Phi'\| \leq 1$;*

(iv) *$\|Q_\Lambda u - \Phi(u)\| \leq 2/(\Lambda^{1-\gamma}) \sup_{v \in \mathcal{A}} \|Q_\Lambda(v + F(v))\|$ for all $u \in \mathcal{A}$.*

Remark 7.6. In the last statement of the preceding theorem, in the case $F = K_N$ we can make

$$\epsilon = \sup_{v \in \mathcal{A}} \|Q(v + F(v))\|$$

very small just by increasing N (see Theorem 5.1). Therefore once $\Lambda \geq \Lambda_0$ we can force the set $\{u = y + z : \|z + QF(y + z)\| \leq \epsilon\}$ to be a very thin neighborhood of graph Φ (and also of the attractor $\mathcal{A}$) just by increasing N. Moreover, for this case we can infer a local approximating property even if $\Lambda < \Lambda_0$. Indeed, it can happen that in a ball $B(z_0, r)$ in QX centered at $z_0 = Qu_0$ for some $u_0 \in \mathcal{A}$, the following two conditions (7.16) and (7.17) hold. Namely,

$$\left\| \left(I + \frac{\partial}{\partial z} QF(u) \right)^{-1}_{u=u_0} \right\| \leq \varpi, \tag{7.16}$$

where the norm is the operator norm on QX and I denotes the identity operator on QX, and

$$(7.17)\quad \left\| \left(I + \frac{\partial}{\partial z} QF(y_0 + z) \right)^{-1} - \left(I + \frac{\partial}{\partial z} QF(y_0 + z_0) \right)^{-1} \right\| \leq \eta \quad \forall z \in B(z_0, r),$$

where $y_0 = Pu_0$. Then if

$$\varpi < r(1 - \varpi\eta r)\|Q(u_0 + F(u_0))\|,$$

there exists $z_1 \in B(z_0, r)$ such that $z_1 + QF(y_0 + z_1) = 0$. In other words, in this case near the point u_0 on the attractor $\mathcal{A}$ we find a point $u_1 = Pu_0 + z_1$ in our set $\mathcal{X}$.

The above statement easily follows by considering the strictly contractive map

$$z - z_0 \mapsto z - z_0 - \left(I + \frac{\partial}{\partial z} QF(u) \right)^{-1} \Big|_{u=u_0} Q(I + F(y_0 + z))$$

of $B(z_0, r)$ into itself.

8. Concluding remarks. We have shown in this paper that the global attractor of a dissipative evolution equation can be approximated, under some favorable conditions, by the zero set of some explicitly constructed polynomial maps. We have presented three types of polynomial maps, namely the polynomials $Q_m(u_0 + I_N(u_0))$, $Q_m(u_0 + J_N(u_0))$, and $Q_m(u_0 + K_N(u_0))$. For the first type our estimates improved with the increase of m and worsened with the increase of N. For the second type, our estimates improved with the increase of both m and N, while for the third type, the estimates improved with the increase of either m or N. The unlimited increase in m is possible only for partial differential equations; therefore we can expect that in the numerical application of our methods to ordinary differential equations, the third type of approximation will turn out to be more effective. Such numerical experiments, which confirm this expectation, will be presented in [4]. Indeed, in considering the classical Lorenz equation in $\mathbb{R}^3$, our numerical computations show that the I_N-approximations are not aberrant only for very small values of N, say $N = 0, 1, 2$, and rapidly deteriorate with the increase of N. The J_N-approximations first improve with the increase of N, then around $N \approx 10$ they start deteriorating. However, the K_N-approximations, slowly improve with the increase of N, at least up to values of $N \sim 1000$. It is worth mentioning that the Lorenz system with the usual parameters $\sigma = 10$, $b = 8/3$, $r = 28$ does not satisfy the assumptions we made in §§3 or 7 because $\delta\Lambda$ is in this case of the order of unity.

REFERENCES

[1] J. E. Billoti and J. P. La Salle, *Dissipative periodic processes*, Bull. Amer. Math. Soc., 77, (1971), pp. 1082–1088.

[2] P. Constantin, C. Foias, and R. Temam, *Attractors representing turbulent flows*, Mem. Amer. Math. Soc., 53 (1985), 67 + vii pages.

[3] C. Foias, M. S. Jolly, I. G. Kevrikides, G. R. Sell, and E. S. Titi, *On the computation of inertial manifolds*, Phys. Lett. A, 131 (1988), pp. 433–436.

[4] C. Foias and M. S. Jolly, *On the numerical algebraic approximation of global attractors*, submitted to Nonlinearity.

[5] C. Foias, O. Manley, and R. Temam, *Modelling of the interaction of small and large eddies in two-dimensional turbulent flows*, Math. Mod. and Num. Anal., 22 (1988), pp. 93–114.

[6] C. FOIAS, *The approximation by algebraic sets of the attractors of dissipative ordinary or partial differential equations*, Frontiers in Pure and Applied Math., Robert Dautray, ed., North-Holland, Amsterdam, 1991, pp. 95–116.

[7] C. FOIAS AND R. TEMAM, *Some analytic and geometric properties of the evolution Navier–Stokes equations*, Journal de Math. Pures et Appl., 58 (1979), pp. 339–368.

[8] ———, *Approximation Algebrique des attracteurs* I. *Le Cas de la dimension finie*, C.R. Acad. Sci. Paris, 307, Ser. I, 1988, pp. 5–8; II. Le cas de la dimension infinie, C.R. Acad. Sci. Paris, 307, Ser. I, 1988, pp. 67–70.

[9] ———, *The algebraic approximation of attractors; The finite dimensional case*, Phys. D, 32 (1988), pp. 163–182.

[10] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[11] J. HALE, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monographs, Vol. 25, American Mathematical Society, Providence, RI, 1988.

[12] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi–groups*, Amer. Math. Soc. Colloq. Publ., 31 (1948), revised ed., Vol. 31, 1957.

[13] H. HOCHSTADT, *The Functions of Mathematical Physics*, Pure Appl. Mathematics, Vol. 22, Wiley-Interscience, New York, 1971.

[14] G. IOOSS, *Secondary bifurcation of a steady solution into invariant torus for evolution problems of Navier–Stokes type*, Lecture Notes in Math., Vol. 503, Springer-Verlag, New York, 1976, pp. 354–365.

[15] J. MALLET–PARET, *Negatively invariant sets of compact maps and an extension of a theorem of Cartwright*, J. Differential Equations, 22 (1976), pp. 331–348.

[16] B. MANDELBROT, *The Fractal Geometry of Nature*, Freeman, San Francisco, 1982.

[17] R. MAÑÉ, *On the Dimension of the Compact Invariant Sets of Certain Nonlinear Maps*, Lecture Notes in Math., Vol. 898, Springer-Verlag, 1981, pp. 230–242.

[18] M. MARION, *Approximate inertial manifolds for the pattern formation Cahn-Hilliard equation*, Math. Mod. Num. Anal., 23 (1989), pp. 463–488.

[19] ———, *Approximate inertial manifolds for reaction-diffusion equations in high space dimension*, Dyn. and Diff. Equ., 1 (1989), pp. 269–298.

[20] F. J. MASSEY III, *Analyticity of solutions of nonlinear evolution equations*, J. Differential Equations, 22 (1976), pp. 416–427.

[21] D. RUELLE AND F. TAKENS, *On the nature of turbulence*, Comm. Math. Phys., 20 (1971), pp. 167–192 and pp. 343–344.

[22] S. SMALE, *Differential dynamical systems*, Bull. Amer. Math. Soc., 73 (1967), pp. 747–817.

[23] ———, *Dynamical systems and turbulence*, Lecture Notes in Math., Vol 615, Springer-Verlag, 1978, pp. 48–70.

[24] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[25] R. TEMAM, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci., Vol. 68, Springer-Verlag, New York, 1988.

[26] ———, *Attractors for the Navier–Stokes equations, localization and approximation*, J. Fac. Sci. Tokyo, Sec. IA, 36 (1989), pp. 629–647.

[27] E. S. TITI, *Une varieté approximante de l'attracteur univérsel des equation de Navier-Stokes, non linéaire, de dimension finie*, C.R. Acad. Sci. Paris, 307 (1988), pp. 383–385.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1303–1319, September 1994

004

INVESTIGATIONS OF SOLUTIONS OF NONLINEAR HYPERBOLIC EQUATIONS WITH A SMALL NONLINEARITY AND APPLICATIONS *

A. ŁADA†

Abstract. The Cauchy problem for the nonlinear hyperbolic equation $L_\varepsilon u + \varepsilon^p N_\varepsilon(t, x, D^\alpha u, |\alpha| \le q) = f_\varepsilon(t,x)$ of order $m \ge 2, 0 \le q \le m-1$, is studied; $\varepsilon \in (0,1]$ is a small parameter. The problem possesses a solution on any given finite interval of time provided ε is sufficiently small. Estimations for Sobolev and L^∞ norms of solutions are derived. Applications to the stability problem for hyperbolic equations, and to justification of the nonlinear geometric optics method for spin glass models, are given.

Key words. hyperbolic equation, nonlinear geometric optics method, stability of solutions, asymptotic solutions, spin glass model

AMS subject classifications. 35L75, 35L70, 82D40

Introduction. In the first part of the paper we investigate the following problem:

$$L_\varepsilon u + \varepsilon^P N_\varepsilon\,(t, x, D^\alpha u, |\alpha| \le q) = f_\varepsilon\,(t,x)\,; \qquad (t,x) \in [0,T] \times R^d$$

(0)

$$\partial_t^l u\,|_{t=0} = 0, \qquad l = 0, \dots, m-1,$$

where L_ε is the hyperbolic operator of order $m, \varepsilon \in (0,1]$ is a small parameter, $T > 0$ is given.

We investigate the case when coefficients of $L_\varepsilon, N_\varepsilon$ and f_ε depend singularly on ε, which is important for applications to the nonlinear geometric optics method.

DEFINITION 0. *Let $g_\varepsilon \in C^\infty(\Omega), \varepsilon \in (0,1]$. We say that the dependence of g_ε on ε is nonsingular when for each multiindex α and compact $K \subset \Omega, \sup\{|\partial^\alpha g_\varepsilon(x)| : \varepsilon \in (0,1], x \in K\}$ is finite. In the opposite case we say that g_ε depends on ε in a singular way.*

The case of nonsingular dependence on ε is an easier one, so the methods developed in this paper will work in this case as well.

We derive estimations for Sobolev and L^∞ norms of solutions of (0), which will be important for the remaining parts of the paper.

In §2 we study the stability of solutions of nonlinear hyperbolic equations with respect to small perturbations of initial data and the right-hand side of the equation.

In §3 we prove justification of the nonlinear geometric optics method applied to the spin glass model with a rapidly oscillating external magnetic field.

A problem similar to (0) was studied in [2] as an admissible problem for justifying the nonlinear geometric optics method applied to second-order nonlinear hyperbolic equations on $R^{d+1}, d = 2, 3$, with quadratic form nonlinearity compatible to a linear part. However there is interest in studying the problem in its general setting. The nonlinearity in the spin glass model is not purely a quadratic form. In contrast with

* Received by the editors January 11, 1993; accepted for publication (in revised form) July 7, 1993.

† Institute of Mathematics and Physics, Agricultural and Technikal Academy, 85-790 Bydgoszcz, al. Kaliskiego 7 Poland.

[2], in our considerations the interval of time $[0,T]$ is not necessary small, but ε should be sufficiently small.

We use the following notation.

$D_t = [0,t] \times R^d, L^k = L^k(R^d), H^k = H^k(R^d)$ denotes the Sobolev space, $\partial^\alpha = \partial_1 \cdots \partial_d, D^\alpha = \partial_t^{\alpha_0} \partial^{\alpha'}, \alpha = (\alpha_0, \alpha')$,

$$\|u\|_{H_\varepsilon^k} = \sum_{|\alpha| \le k} \left\| \varepsilon^{|\alpha|} \partial^\alpha u \right\|_{L^2},$$

$$U^{m,k}(t,\varepsilon) = \sum_{l=0}^{m-1} \sum_{|\alpha| \le k} \left\| \varepsilon^{|\alpha|} \partial^\alpha \partial_t^l u \right\|_{L^\infty([0,t], H^{m-l-1})},$$

$$X_{m,k}(t) = \bigcap_{l=0}^{m-1} C^l\left([0,t], H^{m+k-1}\right),$$

$$\tilde{X}_{m,k}(t,\varepsilon_0) = \Big\{ \{u_\varepsilon : \varepsilon \in (0,\varepsilon_0]\} : u_\varepsilon \in X_{m,k}(t), \text{for each } \varepsilon \in (0,\varepsilon_0), \quad \text{and}$$
$$\Big\{ \Big\{ U_\varepsilon^{m,k}(t,\varepsilon) : \varepsilon \in (0,\varepsilon_0] \Big\} \text{is bounded} \Big\}.$$

1. Problem (0). On $L_\varepsilon, N_\varepsilon, f_\varepsilon, m, p, q, d$, we impose the following hypotheses.

HYPOTHESIS H1. *The operator*

$$L_\varepsilon = \partial_t^m + \sum_{\substack{l+|\alpha|=m \\ l \le m-1}} a_{l,\alpha}(t,x) \partial^\alpha \partial_t^l + \sum_{\substack{l+|\alpha|<m-l \\ l \le m-1}} b_{l,\alpha}(t,x,\varepsilon) \partial^\alpha \partial_t^l$$

is regularly hyperbolic on D_T [8] in the direction of $t, m \ge 2$, the coefficients are smooth, and $D^\beta a_{l,\alpha}, \{\varepsilon^{|\beta|} D^\beta b_{l,\alpha}(\cdot,\varepsilon) : \varepsilon \in (0,1]\}$ are bounded on D_T for each $0 \le l \le m-1, |\alpha| \le m$ and β.

HYPOTHESIS H2. *$0 \le q \le m-1, 1 \le d \le \min\{2m, 4(m-q)\}$, when $L_\varepsilon, N_\varepsilon, f_\varepsilon$ depend singularly on ε then $p > d/2$, and $p = 1$ in the opposite case.*

HYPOTHESIS H3. *$N_\varepsilon \in C^\infty(D_T \times R_\eta^{q(d+1)}), N_\varepsilon(t,x,0) = 0$, for each t, x, ε, and moreover*

$$\text{(H3a)} \qquad \left| \frac{\partial}{\partial \eta_\alpha} N_\varepsilon(t,x,\eta) \right| \le M\left(\sum_{|\beta| \le q} |\eta_\beta| \right), \quad \text{for } |\alpha| \le q,$$

$(t,x,\eta,\varepsilon) \in D_T \times R^{q(d+1)} \times (0,1]$, where $M : R_+ \to R_+$ is continuous and nondecreasing.

(H3b) For any $\{v_\varepsilon : \varepsilon \in (0,1]\} \in \tilde{X}_{m,m}(T,1)$,

$$\|N(\cdot, D^\alpha v_\varepsilon, |\alpha| \le q)\|_{L^\infty([0,t], H_\varepsilon^m)} \le M_0 \left(\sum_{|\alpha| \le m-1} \|D^\alpha v_\varepsilon\|_{L^\infty(D_t)} \right)$$
$$\times \left[V_\varepsilon^{m,m}(t,x) + \sum_{\substack{\alpha \le m \\ \beta \le q}} \left\{ \sum_{\{\gamma+\delta=\alpha, |\gamma,\delta| \ge 1\}} \|\varepsilon^{|\gamma|} \partial^\gamma D^\beta v_\varepsilon\|_{L^\infty([0,t],L^4)} \right.\right.$$
$$\left.\left. \times \|\varepsilon^{|\delta|} \partial^\delta D^\beta v_\varepsilon\|_{L^\infty([0,t],L^4)} \right\} \right],$$

where $M_0 : R_+ \to R_+$ is continuous and nondecreasing.

(H3c) $\text{supp} N_\varepsilon(t,\cdot,\eta) \subset B_\rho \equiv \{|x| \le \rho\}$, for $(t,\eta,\varepsilon) \in [0,T] \times R^{q(d+1)} \times (0,1]$.

HYPOTHESIS H4. $\{||f_\varepsilon||_{L^\infty([0,T],H^m_\varepsilon)} : \varepsilon \in (0,1]\}$ *is bounded, and*

(H4a) supp $f_\varepsilon(t,\cdot) \subset B_\rho, (t,\varepsilon) \in [0,T] \times (0,1]$.

The fundamental result of this paper is the following theorem.

THEOREM 1. *Under Hypotheses* H1–H4, *there exists* $\varepsilon_0 \in (0,1]$, *that is, for* $\varepsilon \in (0,\varepsilon_0]$, *the problem* (0) *possesses a unique solution* $u_\varepsilon \in X_{m,m}(T)$, *and moreover*

$$\{u_\varepsilon : \varepsilon \in (0,\varepsilon_0]\} \in \tilde{X}_{m,m}(T,\varepsilon_0), \tag{1.1}$$

$$\lim_{\varepsilon \to 0} \varepsilon^p \sum_{|\alpha| \le m-1} \|D^\alpha u_\varepsilon\|_{L^\infty(D_T)} = 0. \tag{1.2}$$

Proof. The solution u_ε will be looked for as a limit of the sequence $u^n_\varepsilon, n \ge 0$, where $u^0_\varepsilon = 0$, and

$$L_\varepsilon u^{n+1}_\varepsilon + \varepsilon^p N^n_\varepsilon = f_\varepsilon \quad \text{on } D_t \quad \text{and} \quad \partial^l_t u^{n+1}_\varepsilon|_{t=0} = 0, \tag{1.3}$$

$0 \le l \le m-1$; for $n \ge 0$.

We have denoted $N^n_\varepsilon = N_\varepsilon(t,x,D^\alpha u^n_\varepsilon, |\alpha| \le q)$.

In particular $L_\varepsilon u^1_\varepsilon = f_\varepsilon$; hence from [4, Thm. 23.2.2] there exists a unique solution $u^1_\varepsilon \in X_{m,m}(T), \varepsilon \in (0,1]$, solving (1.3) for $n = 0$. Moreover, from [8, Thm. 6.10], supp $u^1_\varepsilon(t,\cdot) \subset B_{\rho+bt}, (t,\varepsilon) \in [0,T] \times (0,1]$ where $b = \sup\{|\lambda_l(t,x,\xi)| : 1 \le l \le m, |\xi| = 1, (t,x) \in D_T\}, \lambda_l$ are characteristic roots for the principal symbol of L_ε.

Now, $L_\varepsilon u^2_\varepsilon = f_\varepsilon - \varepsilon^p N^1_\varepsilon$. After using (H3b) together with estimations (A1), (A2) (see Appendix), we claim that $f_\varepsilon - \varepsilon^p N^1_\varepsilon \in L^\infty([0,T],H^m), \varepsilon \in (0,1]$. Hence arguing again in the same way as for u^1_ε we obtain existence of $u^2_\varepsilon \in X_{m,m}(t)$, such that supp $u^2_\varepsilon(t,\cdot) \subset B_{\rho+bt}, (t,\varepsilon) \in [0,T] \times (0,1]$. The process can be continued step by step, so for each $n \ge 0$ we obtain the existence of the unique solution $u^{n+1}_\varepsilon \in X_{m,m}(T)$ for (1.3), and moreover

$$\text{supp } u^{n+1}_\varepsilon(t,\cdot) \subset B_{\rho+bT}, (t,\varepsilon) \in [0,T] \times (0,1]. \tag{1.4}$$

To proceed with our considerations we need the energy estimate.

Energy estimate. Let $v \in X_{m,k}(T), \partial^l_t v|_{t=0} = 0, 0 \le l \le m-1$, and $L_\varepsilon v \in L^1([0,T],H^k); \varepsilon \in (0,1]$. Then

$$V^{m,k}(t,\varepsilon) \le c_0 \int_0^t \|L_\varepsilon v(r,\cdot)\|_{H^k_\varepsilon}\, dr, (t,\varepsilon) \in [0,T] \times (0,1], \tag{1.5}$$

where $c_0 > 0$ is a universal constant.

This estimation follows from [4, Lem. 23.2.1]. Returning to the proof, we apply (1.5), with $k = m$, to u^{n+1}_ε. Because of (1.3) this gives

$$\left(U^{n+1}_\varepsilon\right)^{m,m}(t,\varepsilon) \le c_0 \left\{ tK(t) + \varepsilon^p \int_0^t \|N^n_\varepsilon(r,\cdot)\|_{H^m_\varepsilon}\, dr \right\}, \tag{1.6}$$

for $(t,\varepsilon) \in [0,T] \times (0,1]$, where, for brevity, we have denoted

$$K(t) = \sup\left\{ \|f_\varepsilon\|_{L^\infty([0,t],H^m_\varepsilon)} : \varepsilon \in (0,1] \right\}.$$

Of course $(U_\varepsilon^1)(t,\varepsilon) \le c_0 t K(t)$, because $N_\varepsilon^0 = 0$.

Our goal now is to show the existence of $\varepsilon_1 \in (0,1]$ such that

$$(U_\varepsilon^n)^{m,m}(t,\varepsilon) \le 2c_0 t K(t), \qquad (t,\varepsilon) \in [0,T] \times (0,\varepsilon_1], \tag{1.7}$$

$n \ge 1$. To show this, let us note first that from (H3b) together with (A1), (A2),

$$\begin{aligned} \varepsilon^p \int_0^t \|N_\varepsilon^n(r,\cdot)\|_{H_\varepsilon^m} &\le c\varepsilon^{p-(d/2)} \\ &\times \int_0^t \Big\{ M_0\left((U_\varepsilon^n)^{m,m}(r,\varepsilon)\right) \Big[(U_\varepsilon^n)^{m,m}(r,\varepsilon) \\ &\quad + \left((U_\varepsilon^n)^{m,m}(r,\varepsilon)\right)^2 \Big] dr \Big\}, \end{aligned} \tag{1.8}$$

for $(t,x) \in [0,T] \times (0,1], n \ge 1$.

Arguing by induction, let us suppose that (1.7) holds for $n = n_0$ (for $n = 1$ this was already found). Because of (1.8) this yields

$$\varepsilon^p \int_0^t \|N^{n_0}(r,\cdot)\|_{H_\varepsilon^m}\, dr \le 2c_0 t K(t)\, c\varepsilon^{p-(d/2)} A(t), \tag{1.9}$$

where $A(t) = M_0(2c_0 t K(t))(1 + 2c_0 t K(t))$.

Because of (1.6) this gives

$$\left(U_\varepsilon^{n_0+1}\right)^{m,m}(t,\varepsilon) \le c_0 t K(t) \left(1 + 2c\varepsilon^{p-(d/2)} A(t)\right), \qquad (t,\varepsilon) \in [0,T] \times (0,1].$$

So when we choose $\varepsilon_1 \in (0,1]$ such that $1 + 2c\varepsilon_1^{p-(d/2)} A(T) \le 2$ then (1.7) will hold.

Now it remains to investigate the convergence of u^n when $n \to \infty$.

Because $\Delta_\varepsilon^n \equiv u_\varepsilon^{n+1} - u_\varepsilon^n$ satisfies $L_\varepsilon \Delta_\varepsilon^n + \varepsilon^p (N_\varepsilon^n - N_\varepsilon^{n-1}) = 0$ at D_T, and $\partial_t^l \Delta_\varepsilon^n|_{t=0} = 0, 0 \le l \le m-1$, we obtain from (1.5) (with $k = 0$) that

$$(\Delta_\varepsilon^n)^{m,0}(t,\varepsilon) \le c_0 \varepsilon^p \int_0^t \left\| N_\varepsilon^n(r,\cdot) - N_\varepsilon^{n-1}(r,\cdot) \right\|_{L^2} dr. \tag{1.10}$$

Combining the expression

$$\begin{aligned} \left(N_\varepsilon^n - N_\varepsilon^{n-1}\right)(t,x) = \sum_{|\alpha| \le q} \Big\{ \int_0^1 \frac{\partial}{\partial \eta_\alpha} N_\varepsilon\left(t,x, rD^\beta u^n + (1-r) D^\beta u^{n-1}, |\beta| \le q\right) \\ \times \left(D^\alpha u^n - D^\alpha u_\varepsilon^{n-1}\right) \Big\} \end{aligned}$$

with (H3a), (A1), and (1.7) we show that the right-hand side of (1.10) is bounded by $cc_0 \varepsilon^p M(ctK(t)) t (\Delta_\varepsilon^n)^{m,0}(t,\varepsilon), (t,\varepsilon) \in [0,T] \times (0,\varepsilon_1]$.

Therefore, whenever $\varepsilon_0 \in (0,\varepsilon_1]$ is such that $cc_0 \varepsilon_0^p M(cTK(T))T < 1$, then this yields that

$$\{u_\varepsilon^n : n \ge 1\} \tag{1.11}$$

is a Cauchy sequence in $X_{m,0}(T)$ for each $\varepsilon \in (0,\varepsilon_0]$.

From now on $\varepsilon \in (0, \varepsilon_0]$ everywhere in the proof.

We denote by u_ε the limit of u_ε^n in $X_{m,0}(T)$. Because of (1.4) we infer that also $D^\alpha u_\varepsilon^n$ converges almost everywhere on D_T to $D^\alpha u_\varepsilon$ when $|\alpha| \leq q$. Therefore, N_ε^n converges almost everywhere on D_T to $N_\varepsilon = N_\varepsilon(t, x, D^\alpha u_\varepsilon, |\alpha| \leq q)$.

From (1.9) we have

$$\int_0^T \|N_\varepsilon^n(r, \cdot)\|_{H_\varepsilon^m} \, dr \leq 2c_0 \varepsilon^{-p} TK(T). \tag{1.12}$$

Because of (1.11) this gives $N_\varepsilon^n \to N_\varepsilon$ in $L^2(D_T), n \to \infty$, (see [6, Lem. 1.3]). We see from (1.11) and (1.7) that for $0 \leq l \leq m-1, t \in [0,T], \partial_t^l u_\varepsilon^n(t, \cdot) \to \partial_t^l u_\varepsilon(t, \cdot)$ in $H^{2(m-1)-l}$ when $n \to \infty$.

We write (1.3) in the form $Pu_\varepsilon^{n+1} = f_\varepsilon - \varepsilon^p N_\varepsilon^n - SL_\varepsilon u_\varepsilon^{n+1}$, where P is the principal part of $L_\varepsilon, SL_\varepsilon = L_\varepsilon - P$. From the above, the right-hand side tends in $L^2(D_T)$ to the limit $f_\varepsilon - \varepsilon^p N_\varepsilon - SL_\varepsilon u_\varepsilon$, when $n \to \infty$. This gives that Pu_ε^n converges in $L^2(D_T)$ to Pu_ε, when $n \to \infty$, because it does in the distribution sense.

By (1.12), $N_\varepsilon \in L^1([0,T], H^m)$. By [4, Thm. 23.2.2] this yields that u_ε, as a solution of $L_\varepsilon u_\varepsilon = f_\varepsilon - \varepsilon^p N_\varepsilon$, belongs to $X_{m,m}(T)$. The assertion (1.1) follows from (1.7). Since $p > d/2$, from (1.1) and (A1) we have (1.2). This finishes the proof.

By suitable modifications of the scheme of proof we can obtain the following results.

PROPOSITION 2. *Under Hypotheses* H1–H4 *there exists* $T_0 \in (0, T]$, *that is, for each* $\varepsilon \in (0, 1]$, *problem* (0) *possesses a unique solution* $u_\varepsilon \in X_{m,m}(T_0)$, *and moreover*

$$\{u_\varepsilon : \varepsilon \in (0,1]\} \in \tilde{X}_{m,m}(T_0, 1), \qquad \lim_{\varepsilon \to 0} \varepsilon^p \sum_{|\alpha| \leq m-1} \|D^\alpha u_\varepsilon\|_{L^\infty(D_{T_0})} = 0.$$

PROPOSITION 3. *We admit Hypotheses* H1–H4 *with the following modifications:*

(i) *in Hypothesis* H1 *the dependence of all* $b_{l,\alpha}$ *on* ε *is nonsingular*;

(ii) *in Hypothesis* H2, $p = 1$;

(iii) *in Hypothesis* H3 *instead of* (H3b) *we impose the inequality for* $v \in X_{m,m}(T)$ *where on the left-hand side we have* H^m *instead of* H_ε^m *and on the right-hand side instead of* $V_\varepsilon^{m,m}(t, \varepsilon)$ *we have*

$$\sum_{l=0}^{m-1} \|\partial_t^l v\|_{L^\infty([0,T], H^{2m-l-1})},$$

and $\varepsilon = 1$;

(iv) *in Hypothesis* H4, $\{||f_\varepsilon||_{L^\infty([0,T], H^m} : \varepsilon \in (0,1)\}$ *is bounded.*

Then there exists $\varepsilon_0 \in (0, 1]$, *such that for each* $\varepsilon \in (0, \varepsilon_0]$, *problem* (0) *possesses a unique solution* $u_\varepsilon \in X_{m,m}(T)$, *and*

$$\left\{ \sum_{l=0}^{m-1} \|\partial_t^l u_\varepsilon\|_{L^\infty([0,T], H^{2m-l-1})} : \varepsilon \in (0, \varepsilon_0] \right\} \tag{1.13}$$

is bounded,

$$\lim_{\varepsilon \to 0} \varepsilon \sum_{|\alpha| \leq m-1} \|D^\alpha u_\varepsilon\|_{L^\infty(D_T)} = 0. \tag{1.14}$$

Remark 4. Each solution u_ε obtained in Theorem 1, Propositions 2 and 3 satisfies supp $u_\varepsilon(t,\cdot) \subset B_{\rho+bT}, t \in [0,T]$, for those ε under consideration.

Comments. Proposition 2 is an analogy of [2, Prop. 3.1]. We can also formulate the time-local version of Proposition 3, similar to Proposition 2. Whenever we impose Hypotheses H1–H4 without (H3c), (H4a), then Theorem 1 and Propositions 2 and 3, reformulated in local space, remain valid; see [2]. Condition (H3b) is not a most general condition, for the scheme of proof of Theorem 1 can be proved. Such a type of inequality as in (H3b) is observed in the spin glass model. However, on the right-hand side of (H3b) we can consider L^p norms, $p \neq 4, p > 2$, as well.

2. The stability problem. Let $u^0 \in C^\infty(D_T)$ be a solution of

$$\begin{aligned} &Lu + F\left(D^\alpha u, |\alpha| \leq q\right) = f\left(t,x\right);\ (t,x) \in D_T, \\ &\partial_t^l u\,|_{t=0} = u_l, \qquad 0 \leq l \leq m-1, \end{aligned} \tag{2.1}$$

and $D^\alpha u^0$ is bounded on D_T for each α.

Consider the perturbed problem

$$\begin{aligned} &Lu + F\left(D^\alpha u, |\alpha| \leq q\right) = f\left(t,x\right) + \varepsilon^p g_\varepsilon\left(t,x\right);\ (t,x) \in D_T, \\ &\partial_t^l u\,|_{t=0} = u_l + \varepsilon^p v_l; \qquad 0 \leq l \leq m-1. \end{aligned} \tag{2.2}$$

Assumptions about $L, F, g_\varepsilon, v_l, 0 \leq l \leq m-1, p, q, d, m$, will be given later.

We seek the solution of (2.2) in the form

$$u_\varepsilon = u^0 + \varepsilon^p\left(v + w_\varepsilon\right), \tag{2.3}$$

where $v = \sum_{l=0}^{m-1} t^l v_l$.

We find that w_ε should satisfy

$$\begin{aligned} &L_\varepsilon w + \varepsilon^p N_\varepsilon\left(t,x,D^\alpha w, |\alpha| \leq q\right) = f_\varepsilon\left(t,x\right);\ (t,x) \in D_T, \\ &\partial_t^l w\,|_{t=0} = 0, \qquad 0 \leq l \leq m-1, \end{aligned} \tag{2.4}$$

where

$$L_\varepsilon = L + \sum_{|\alpha| \leq q} \left\{ \frac{\partial}{\partial \eta_\alpha} F\left(D^\beta\left(u^0 + \varepsilon^p v\right), |\beta| < q\right) D^\alpha \right\},$$

$$f_\varepsilon = g_\varepsilon - Lv - \sum_{|\alpha| \leq q} \left\{ \int_0^1 \frac{\partial}{\partial \eta_\alpha} F\left(D^\beta\left(u^0 + r\varepsilon^p v\right), |\beta| < q\right) dr D^\alpha v \right\},$$

$$N_\varepsilon = \sum_{|\alpha|,|\beta| \leq q} \left\{ \int_0^1 \frac{\partial^2}{\partial \eta_\alpha \partial \eta_\beta} F\left(D^\gamma\left(u^0 + \varepsilon^p v + r\varepsilon^p w\right), |\gamma| < q\right)(1-r)\, dr D^\alpha w D^\beta w \right\}.$$

We assume that (i) L satisfies Hypothesis H1, but coefficients do not depend on ε;

(ii) $F \in C^\infty(R^{q(d+1)})$ is such that the condition (H3b) for N_ε holds. For example, if the dependence of F on $D^\alpha u, 1 \leq |\alpha| \leq q$, is quadratic, then the latter takes place; see §3.3. Moreover we assume $0 \leq q \leq m-1$;

(iii) $v_l \in C_0^\infty(R^d)$, supp $v_l \subset B_\rho, 0 \le l \le m-1$;

(iv) if g_ε depends singularly on ε we suppose that this satisfies Hypothesis H4. In the opposite case this satisfies condition (iv) in Proposition 3 and (H4a);

(v) p, d satisfy Hypothesis H2.

Whenever we apply Theorem 1 or Proposition 3 to the problem (2.4) we immediately infer the following result.

COROLLARY 5. *Under the above assumptions, there exists* $\varepsilon_0 \in (0,1]$, *that is, for* $\varepsilon \in (0, \varepsilon_0]$, *problem* (2.2) *possesses a unique solution* $u_\varepsilon \in X_{m,m}(T)$, *and*

$$\lim_{\varepsilon \to 0} \sum_{l=0}^{m-1} \left\| \partial_t^l (u_\varepsilon - u^0) \right\|_{L^\infty([0,T], H^{m-l-1})} = 0, \tag{2.5}$$

$$\lim_{\varepsilon \to 0} \sum_{|\alpha| \le m-1} \left\| D^\alpha (u_\varepsilon - u^0) \right\|_{L^\infty(D_T)} = 0. \tag{2.6}$$

Example 6. We consider the problem having its origin in the spin glass model; see [1], [7] and §3 of this paper. We have

$$\begin{aligned} &\Box u - \frac{2u}{1+u^2} \left(|\partial_t u|^2 - |\nabla_x u|^2 \right) = \delta g(t,x); \qquad (t,x) \in D_T \subset R \times R^3 \\ &\partial_t^l u = u_l + \delta v_l; \qquad l = 0,1, \quad \text{for } t = 0, \end{aligned} \tag{2.7}$$

where $\Box = \partial_t^2 - (\partial_1^2 + \partial_2^2 + \partial_3^2)$, $g \in L^\infty([0,T], H^2)$, supp $g(t,\cdot) \subset B_\rho$ for $t \in [0,T]$, $\delta \in [0,1]$. For u_l, v_l, we impose the same assumptions as in Corollary 5.

When $\delta = 0$, the solution u^0 of (2.7) has the following form [7]:

$$u^0(t,x) = \tan\left(\arctan u_0(x) + \int_0^t \tilde{u}(r,x)\, dr \right), \tag{2.8}$$

where $\tilde{u}$ is the solution of

$$\Box \tilde{u} = 0, \quad \tilde{u}_{t=0} = \frac{u_1}{(1+u_0^2)}, \quad \partial_t \tilde{u}_{t=0} = \sum_{l=0}^{3} \partial_l \left(\frac{\partial_l u_0}{1+u_0^2} \right).$$

When $u_0, u_1 = 0$ then $u^0 = 0$. It is possible to choose assumptions on u_0, u_1 [7] so that u^0 given by (2.8) is nontrivial and exists globally on $R \times R^3$. Therefore we can make the assumption that u_0, u_1 admit the existence of u^0 on D_T. From Corollary 5 we have that there exists $\delta_0 \in (0,1]$ such that for each $\delta \in (0, \delta_0]$, the problem (2.7) possesses a unique solution $u_\delta \in X_{2,2}(T)$, and

$$\lim_{\delta \to 0} \sum_{l=0}^{1} \left\| \partial_t^l (u_\delta - u^0) \right\|_{L^\infty([0,T], H^{1-l})} = 0, \tag{2.9}$$

$$\lim_{\delta \to 0} \sum_{|\alpha| \le 1} \left\| D^\alpha (u_\delta - u^0) \right\|_{L^\infty(D_T)} = 0. \tag{2.10}$$

3. Model of the spin glass.

3.1. Description of the model. We consider the model of Andreev and Marčenko [1]. Let the orientation of spins of a spin glass material in an equilibrium state be described by the field $M_0 : R^3 \to S^2$; S^2 is the two-dimensional sphere.

We assume the following situation: At the moment $t = 0$ the material is in an equilibrium state, and then an external magnetic field is switched on. The influence of the magnetic field on the orientation of the fields is described by the law

$$M = M_0 + \frac{2v}{1+|v|^2}\left(v \times M_0 + v \times (v \times M_0)\right),$$

where $M \equiv M(t,x) \in S^2$ describes the orientation of spins for $t > 0$, $\times$ denotes the vector product in R^3, and $v : R \times R^3 \to R^3$ satisfies the system

(3.1)

$$\Box v - \left(1+|v|^2\right)^{-1}\left(\partial_t |v|^2 \partial_t v - \sum_{k=1}^{3} \partial_k |v|^2 \partial_k v\right) + \partial_t H$$
$$+ \partial_t (H \times v) + \left(1+|v|^2\right)^{-1}\left(2\langle \partial_t v, H + H \times v\rangle - \partial_t |v|^2 (H + H \times v)\right) = 0;$$

here H denotes the external magnetic field, $\langle \cdot, \cdot \rangle$ denotes the euclidean scalar product.

The appropriate initial conditions for v will be

$$\partial_t^l v\,|_{t=0} = 0, \quad l = 0, 1. \tag{3.2}$$

In this paper we reduce to the case $H = (H_1, 0, 0), H_1(t, \cdot) = -\int_0^t f_\varepsilon(r, \cdot)\, dr$; f_ε is a rapidly oscillating function and will be described later, $\varepsilon \in (0, 1]$. The uniqueness theorem [5, p. 48] allows us to seek the solution v of (3.1), (3.2) in the form $v = (u, 0, 0)$, where u satisfies

$$\Box u - a(u)\left(|\partial_t u|^2 - |\nabla_x u|^2\right) - f_\varepsilon(t,x) = 0,$$
$$\partial_t^l u\,|_{t=0} = 0, \quad l = 0, 1. \tag{3.3}$$

Here we have denoted $a(u) = 2u(1+u^2)^{-1}$.

Below we consider the problem (3.3) on $D_T \subset R \times R^3; T > 0$ is given.

The function f_ε will be considered in the form

$$f_\varepsilon(t,x) = \sum_{k \in K} f_k\left(t, x, \varepsilon^{-1}\phi_k(t,x), \varepsilon\right),$$

where K is a finite set of indices, for each $k \in K$, $f_k(t,x,\theta,\varepsilon)$ is periodic with respect to θ with period 2π.

HYPOTHESIS H5. (i) $f_k \in C^\infty(D_T \times [0, 2\pi] \times [0, 1])$, supp $f_k(t, \cdot, \theta, \varepsilon) \subset B_\rho$, $(t, \theta, \varepsilon) \in [0, T] \times [0, 2\pi] \times [0, 1], k \in K$.

(ii) *Whenever*

$$f_k(t,x,\theta,\varepsilon) = f_k^0(t,x,\theta) + \varepsilon f_k^1(t,x,\theta) + \varepsilon^2 f_k^2(t,x,\theta,\varepsilon),$$

then we impose $f_k^l(0,x,\theta) = 0 (x,\theta) \in R^3 \times [0, 2\pi], l = 0, 1, k \in K$.

(iii) *Let*

$$hf_k^l = (2\pi)^{-1}\int_0^{2\pi} f_k^l\, d\theta, \quad k \in K, \quad hf^l = \sum_{k \in K} hf_k^l, \qquad l = 0, 1, 2$$

then we impose $hf_k^2 = 0, k \in K$, *and* hf^0 *is so small that the problem* (2.7) *in Example* 6, *with the right-hand side* hf^0 *instead of* δg, *has a solution on* D_T *satisfying* $\partial_t^l u|_{t=0} = 0, l = 0, 1$.

HYPOTHESIS H6. (i) $\phi_k \in C^\infty(D_T)$ *and* $|\partial_t \phi_k|^2 - |\nabla_x \phi_k|^2 = 0$ *on* $D_T, k \in K$.

(ii) $d\phi_k \neq 0$ *on* $D_T, k \in K$.

(iii) *For each* $k \neq 1, |B(d\phi_k, d\phi_l)| \geq c > 0$ *on* D_T, *where we denote* $B(df, dg) = \partial_t f \partial_t g - \sum_{l=1}^3 \partial_l f \partial_l g$.

This hypothesis immediately yields the following results.

Remark 7. We have

$$\partial_t \phi_k \neq 0, \ \nabla_x \phi_k \neq 0 \quad \text{on } D_T, \ k \in K, \tag{3.4}$$

$$\begin{aligned} &\text{for each } k \neq 1, \ d(c_1\phi_k + c_2\phi_l) \neq 0, \text{ and } c_1\phi_k + c_2\phi_l \\ &\quad \text{does not satisfy the eikonal equation on } D_T. \\ &\quad \text{Here } c_1, c_2 \neq 0 \text{ are constants.} \end{aligned} \tag{3.5}$$

3.2. Construction of the asymptotic solution.

The asymptotic solution $\tilde{u}_\varepsilon$ of (3.3) will be constructed by the nonlinear geometric optics method [2]. So $\tilde{u}_\varepsilon$ will be looked for in the form

$$\begin{aligned} \tilde{u}_\varepsilon(t,x) = u^0(t,x) &+ \varepsilon \left\{ v^0(t,x) + \sum_{k \in K} u_k^1(t,x,\varepsilon^{-1}\phi_k) \right\} \\ &+ \varepsilon^2 \left\{ \sum_{k \in K} v_k^1(t,x,\varepsilon^{-1}\phi_k) + \sum_{(k,l) \in K^2} u_{(k,l)}^2(t,x,\varepsilon^{-1}\phi_k, \varepsilon^{-1}\phi_l) \right\} \\ &+ \varepsilon^3 \left\{ \sum_{(k,l) \in K^2} v_{(k,l)}^2(t,x,\varepsilon^{-1}\phi_k, \varepsilon^{-1}\phi_l) + \sum_{\bar{k} \in K^3} u_{\bar{k}}^3(t,x,\varepsilon^{-1}\phi_{\bar{k}}) \right\}; \end{aligned} \tag{3.6}$$

we have denoted $\bar{k} = (k_1, k_2, k_3)$, and $\phi_{\bar{k}} = (\phi_{k_1}, \phi_{k_2}, \phi_{k_3})$.

PROPOSITION 8. *There exist* $u^0, v^0 \in C^\infty(D_T), u_k^1, v_k^1 \in C^\infty(D_T \times [0, 2\pi]), k \in K, u_{\bar{k}}^2, v_{\bar{k}}^2 \in C^\infty(D_T \times [0, 2\pi]^2), \bar{k} \in K^2, u_{\bar{k}}^3 \in C^\infty(D_T \times [0, 2\pi]^3), \bar{k} \in K^3$, *such that*

$$\Box \tilde{u}_\varepsilon - a(\tilde{u}_\varepsilon)\left(|\partial_t \tilde{u}_\varepsilon|^2 - |\nabla_x \tilde{u}_\varepsilon|^2\right) - f_\varepsilon = \varepsilon^2 R_\varepsilon, \ \partial_t \tilde{u}_\varepsilon|_{t=0} = 0,$$

where $\{\varepsilon^{|\alpha|} D^\alpha R_\varepsilon : \varepsilon \in (0,1]\}$ *is bounded in* $L^\infty(D_T)$ *for each multiindex* α. *Moreover*

(i) *all the functions quoted above (without* R_ε*) and their first derivatives with respect to* t *vanish at* $t = 0$.

(ii) *There exists a ball* $B, B_\rho \subset B$, *such that supports with respect to* x *of all the functions quoted above, together with* R_ε, *are contained in* B.

(iii) $u_k^1, v_k^1, k \in K$, *are periodic with respect to* θ *with period* 2π *and their mean values with respect to* θ *over* $[0, 2\pi]$ *equal* 0. *The same property with respect to* $\theta_1, \theta_2 \in [0, 2\pi]$ *holds for* $u_{\bar{k}}^2, v_{\bar{k}}^2, \bar{k} \in K^2$, *and also for* $u_{\bar{k}}^3, \bar{k} \in K^3$, *with respect to* $\theta_1, \theta_2, \theta_3 \in [0, 2\pi]$.

Proof. We insert into (3.3) the function $\tilde{u}_\varepsilon$ in the form (3.6) and collect the terms of ε^0 and ε^1. This leads to equations describing the functions quoted in Proposition 8.

Step 1. *Determination of* $u^0, u_k^1, k \in K, u_{(k,l)}^2, k \neq l$.

The expression, which stands at ε^0, is a sum of left-hand sides of the following equations:

$$\Box u^0 - a(u^0)\left(|\partial_t u^0|^2 - |\nabla_x u^0|^2\right) = hf^0(t,x), \tag{3.7}$$

$$\sum_{k\in K}\{2X_k\partial_\theta u_k^1 - 2a(u^0)B(du^0, d\phi_k)\partial_\theta u_k^1 + hf_k^0 - f_k^0\} = 0, \tag{3.8}$$

$$\sum_{(k,l)\in K^2}\left\{2B(d\phi_k, d\phi_l)\left[\partial_{\theta_1}\partial_{\theta_2}u^2_{(k,l)} - a(u^0)\partial_{\theta_1}u_k^1\partial_{\theta_2}u_l^1\right]\right\} = 0 \tag{3.9}$$

where we have denoted $X_k = \partial_t\phi_k\partial_t - \sum_{l=1}^3 \partial_l\phi_k\partial_l$, $k\in K$.

Equation (3.7). Under the assumption on hf^0 we can infer from the considerations proved for Example 6 that there exists a solution of (3.7) having the initial datum zero and having the desired properties. The smoothness follows from the smoothness of hf^0, and the assertion concerning support follows from compactness of the support of hf^0 and Remark 4.

Equation (3.8). We seek the solution of (3.8) satisfying $u_k^1(0,x,\theta) = 0, k\in K$. First, let us consider the following admissible problems:

$$\partial_t w_k - (\partial_t\phi_k)^{-1}\left\{\sum_{l=1}^3 \partial_l\phi_k\partial_l w_k + a(u^0)B(du^0, d\phi_k)w_k + f_k^0(t,x,\theta) - hf_k^0(t,x)\right\} = 0,$$

$w_k(0,x,\theta) = 0$, for $k\in K$.

There exists a good theory [4], [8] for such linear problems. Therefore there exist smooth solutions w_k such that supp $w_k(t,\cdot)\subset B, t\in[0,T], k\in K$. Moreover it is not difficult to observe that w_k are periodic with respect to θ with period 2π. After integration of the equation for w_k with respect to θ over $[0,2\pi]$ we claim that $\int_0^{2\pi} w_k\, d\theta$ solves a homogeneous problem, so equals 0. Therefore,

$$u_k^1(t,x,\theta) = \int_0^\theta w_k(t,x,s)\, ds - (2\pi)^{-1}\int_0^{2\pi}\int_0^s w_k(t,x,r)\, dr\, ds$$

satisfies

$$2X_k\partial_\theta u_k^1 - 2a(u^0)B(du^0, d\phi_k)\partial_\theta u_k^1 + hf_k^0 - f_k^0 = 0,$$

and has the desired properties.

Equation (3.9). For $k\neq l$ we put $u^2_{(k,l)}(t,x,\theta_1,\theta_2) = a(u^0)u_k^1(t,x,\theta_1)u_l^1(t,x,\theta_2)$. The terms with $k=l$ are not important in (3.9) because $B(d\phi_k, d\phi_k) = 0$.

Determination of $v^0, v_k^1, u^2_{(k,k)}, k\in K, v^2_{\bar{k}}, \bar{k}\in K^2, u^3_{\bar{k}}, \bar{k}\in K^3$.

The expression standing at ε^1 is a sum of the left-hand sides of the following equations:

$$\Box v^0 - 2a(u^0)B(du^0, dv^0) - a'(u^0)B(du^0, du^0)v^0 - hf^1 = 0, \tag{3.10}$$

$$\begin{aligned}\sum_{k\in K}\{&2[X_k\partial_\theta v_k^1 - a(u^0)B(du^0, d\phi_k)\partial_\theta v_k^1]\\ &- 2a(u^0)(B(du^0, du_k^1) + B(du^0, d\phi_k)\partial_\theta u_k^1) - a'(u^0)(B(du^0, du^0)u_k^1\\ &+ 2B(du^0, d\phi_k)v^0\partial_\theta u_k^1) + hf_k^1 - f_k^1\} = 0,\end{aligned} \tag{3.11}$$

$$\sum_{k\in K}\left\{2X_k\left(\partial_{\theta_1}+\partial_{\theta_2}\right)u^2_{(k,k)}+G_{(k,k)}\right\}=0, \tag{3.12}$$

$$\sum_{k\neq l}\left\{2B\left(d\phi_k,d\phi_l\right)v^2_{(k,l)}+2X_k\left(\partial_{\theta_1}+\partial_{\theta_2}\right)u^2_{(k,l)}+G_{(k,l)}\right\}=0, \tag{3.13}$$

$$\begin{aligned}\sum_{\bar k\in K^2} 2\Big\{&B\left(d\phi_{k_1},d\phi_{k_2}\right)\left[\partial_{\theta_1}\partial_{\theta_2}u_k-\tfrac{1}{3}a\left(u^0\right)\left(\partial_{\theta_1}u^1_{k_1}\partial_{\theta_2}u^2_{(k_1,k_2)}\right.\right.\\&+\partial_{\theta_2}u^1_{k_2}\partial_{\theta_1}u^2_{(k_1,k_3)}\Big)-\tfrac{1}{6}a'\left(u^0\right)u^1_{k_3}\partial_{\theta_1}u^1_{k_1}\partial_{\theta_2}u^1_{k_2}\Big]\\&+B\left(d\phi_{k_2},d\phi_{k_3}\right)\left[\partial_{\theta_2}\partial_{\theta_3}u^3_{\bar k}-\tfrac{1}{3}a\left(u^0\right)\left(\partial_{\theta_2}u^1_{k_2}\partial_{\theta_3}u^2_{(k_1,k_2)}\right.\right.\\&+\partial_{\theta_3}u^1_{k_3}\partial_{\theta_2}u^2_{(k_1,k_2)}\Big)-\tfrac{1}{6}a'\left(u^0\right)u^1_{k_1}\partial_{\theta_2}u^1_{k_2}\partial_{\theta_3}u^1_{k_3}\Big]\\&+B\left(d\phi_{k_1},d\phi_{k_3}\right)\left[\partial_{\theta_1}\partial_{\theta_3}u^3_{\bar k}-\tfrac{1}{3}a\left(u^0\right)\left(\partial_{\theta_1}u^1_{k_1}\partial_{\theta_3}u^2_{(k_2,k_3)}\right.\right.\\&+\partial_{\theta_3}u^1_{k_3}\partial_{\theta_1}u^2_{(k_1,k_2)}\Big)-\tfrac{1}{6}a'\left(u^0\right)u^1_{k_2}\partial_{\theta_1}u^1_{k_1}\partial_{\theta_3}u^1_{k_3}\Big]\Big\}=0,\end{aligned} \tag{3.14}$$

where we have denoted

$$\begin{aligned}G_{(k,l)}=&\left(\Box\phi_k\partial_{\theta_1}+\Box\phi_l\partial_{\theta_2}\right)u^2_{(k,l)}-a\left(u^0\right)\left\{\left[B\left(du^0,d\phi_k\right)\partial_{\theta_1}\right.\right.\\&+B\left(du^0,d\phi_l\right)\partial_{\theta_2}\Big]u^2_{(k,l)}+\tfrac{1}{2}\left(\partial_{\theta_1}u^1_k\partial_{\theta_2}v^1_l+\partial_{\theta_2}u^1_l\partial_{\theta_1}v^1_k\right)\\&+\tfrac{1}{2}\left[B\left(d\phi_k,du^1_l\right)\partial_{\theta_1}u^1_k+B\left(d\phi_l,du^1_k\right)\partial_{\theta_2}u^1_l\right]\Big\}-a'\left(u^0\right)\left\{B\left(d\phi_k,d\phi_l\right)v^0\right.\\&\left.\times\partial_{\theta_1}u^1_k\partial_{\theta_2}u^1_l+\tfrac{1}{2}\left[B\left(du^0,d\phi_k\right)u^1_l\partial_{\theta_1}u^1_k+B\left(du^0,d\phi_l\right)u^1_k\partial_{\theta_2}u^1_l\right]\right\}.\end{aligned}$$

Equation (3.10). We consider the solutions subject to initial conditions $\partial^l_t v^0|_{t=0}=0, l=0,1$. The problem is linear, so it can be solved by known methods.

Equation (3.11). We consider the solutions subject to initial conditions $v^1_k(0,x,\theta)=0,(x,\theta)\in R^3\times[0,2\pi],k\in K$. The problem is solved by the same argument as used for (3.8).

Equation (3.12). The structure of the equation and of $G_{(k,k)},k\in K$, suggests that we seek $u^2_{(k,k)}(t,x,\theta,\theta)$ as a solution of

$$\begin{aligned}&X_ku^2_{(k,k)}+\tfrac{1}{2}\left(\Box\phi_k-a\left(u^0\right)B\left(du^0,d\phi_k\right)\right)u^2_{(k,k)}\\&\quad-\tfrac{1}{4}\left\{a\left(u^0\right)\left(u^1_kv^1_k+B\left(d\phi_k,du^1_k\right)u^1_k\right)-a'\left(u^0\right)B\left(du^0,d\phi_k\right)\left(u^1_k\right)^2\right\}=0,\\&\qquad u^2_{(k,k)}|_{t=0}=0,\quad \text{for } k\in K.\end{aligned}$$

This problem can be solved by the same argument as for (3.8).

Equation (3.13). We put $v^2_{(k,l)}=-(2B(d\phi_k,d\phi_l))^{-1}\{G_{(k,l)}+2X_k(\partial_{\theta_1}+\partial_{\theta_2})u^2_{(k,l)}\}$ when $k\neq l$, and $v^2_{(k,k)}=0$.

Equation (3.14). For $k_1=k_2=k_3$ we put $u^3_{\bar k}=0$, for $k_1\neq k_2\neq k_3, u^3_{\bar k}=(\frac{1}{3}a^2(u^0)+\frac{1}{6}a'(u^0))\prod^3_{l=1}u^1_{k_l}$. In the latter formula we have taken into account the form of $u^2_{\bar k}$, for $k_1\neq k_2$.

Now we fix our attention on $\bar k=(k,l,l),k\neq l$. The remaining cases for $\bar k$ can be investigated by analogy. So, we must find $u^3_{(k,l,l)}$ satisfying

$$\begin{aligned}\partial_{\theta_1}\left(\partial_{\theta_2}+\partial_{\theta_3}\right)u^3_{(k,l,l)}|_{\theta_2=\theta_3}=&\Big\{\tfrac{1}{3}a\left(u^0\right)\left[\partial_{\theta_1}u^1_k\left(\partial_{\theta_2}+\partial_{\theta_3}\right)u^2_{(l,l)}\right.\\&+\partial_{\theta_1}u^2_{(k,l)}\left(\partial_{\theta_2}+\partial_{\theta_3}\right)u^1_l\Big]+\tfrac{1}{6}a'\left(u^0\right)u^1_l\partial_{\theta_1}u^1_k\left(\partial_{\theta_2}+\partial_{\theta_3}\right)u^1_l\Big\}|_{\theta_2=\theta_3}.\end{aligned}$$

Taking into account the identity

$$\partial_{\theta_1} u^2_{(l,l)} \mid_{\theta_1=\theta_2} = \partial_{\theta_3} u^2_{(l,l)} \mid_{\theta_1=\theta_2},$$

and the form of $u^2_{(k,l)}, k \neq l$, we ensure the functions

$$\begin{aligned} u^3_{(k,l,l)}(\cdot,\theta_1,\theta_2,\theta_3) = {} & \tfrac{1}{3} a(u^0) u^1_k(\cdot,\theta_1) u^2_{(l,l)}(\cdot,\theta_2,\theta_3) \\ & + \left(\tfrac{1}{3} a^2(u^0) + \tfrac{1}{12} a'(u^0)\right) u^1_k(\cdot,\theta_1) \\ & \times \left[(u^1_l(\cdot,\theta))^2 - \frac{1}{2\pi}\int_0^{2\pi} (u^1_l(\cdot,\theta))^2 \, d\theta\right] \end{aligned}$$

enjoy the conditions of the Proposition.

The properties of R_ε can be followed from its exact formula, but we omit this because of the tedious notation. More information about R_ε will be given in §3.4.

3.3. Existence of the solution and validity of the asymptotic development.

PROPOSITION 9. *There exists* $\varepsilon_0 \in (0,1]$, *such that for each* $\varepsilon \in (0,\varepsilon_0]$ *the problem* (3.3) *has a unique solution* $u_\varepsilon \in X_{2,2}(T) \cap C^\infty(D_T)$, *and*

$$\lim_{\varepsilon\to 0} \sum_{l=0}^{1} \left\| \partial_t^l (u_\varepsilon - \tilde{u}_\varepsilon) \right\|_{L^\infty([0,T],H^{1-l})} = 0, \tag{3.15}$$

$$\lim_{\varepsilon\to 0} \sum_{|\alpha|\leq 1} \left\| D^\alpha (u_\varepsilon - \tilde{u}_\varepsilon) \right\|_{L^\infty(D_T)} = 0, \tag{3.16}$$

$$\left\{ \sum_{l=0}^{1} \sum_{|\alpha|\leq 2} \left\| \varepsilon^{|\alpha|} \partial^\alpha \partial_t^l r_\varepsilon \right\|_{L^\infty([0,T],H^{1-l})} : \varepsilon \in (0,\varepsilon_0] \right\} \tag{3.17}$$

is bounded; *here* $r_\varepsilon = \varepsilon^{-2}(u_\varepsilon - \tilde{u}_\varepsilon)$.

Proof. For r_ε we derive

$$\Box r_\varepsilon + 2a(\tilde{u}_\varepsilon) B(d\tilde{u}_\varepsilon, dr_\varepsilon) + h_\varepsilon r_\varepsilon + \varepsilon^2 N_\varepsilon(t,x,D^\alpha r_\varepsilon, |\alpha| \leq 1) = R_\varepsilon \tag{3.18}$$

at D_T, and $\partial_t^l r_\varepsilon|_{t=0} = 0, l = 0,1$.

We have denoted $h_\varepsilon = a'(\tilde{u}_\varepsilon) B(d\tilde{u}_\varepsilon, d\tilde{u}_\varepsilon)$,

$$\begin{aligned} N_\varepsilon = {} & a(\tilde{u}_\varepsilon) B(dr_\varepsilon, dr_\varepsilon) + 2a'(\tilde{u}_\varepsilon) r_\varepsilon B(d\tilde{u}_\varepsilon, dr_\varepsilon) + \varepsilon^2 a'(\tilde{u}_\varepsilon) r_\varepsilon B(dr_\varepsilon, dr_\varepsilon) \\ & + \tfrac{1}{2} F(t,x,r_\varepsilon,\varepsilon) \left[B(d\tilde{u}_\varepsilon, d\tilde{u}_\varepsilon) + 2\varepsilon^2 B(d\tilde{u}_\varepsilon, dr_\varepsilon) + \varepsilon^4 B(dr_\varepsilon, dr_\varepsilon)\right], \end{aligned}$$

where

$$F(\cdot, r_\varepsilon, \varepsilon) = (2\varepsilon^2)^{-1} \int_{\tilde{u}_\varepsilon}^{\tilde{u}_\varepsilon + \varepsilon^2 r_\varepsilon} a''(s) (\tilde{u}_\varepsilon + \varepsilon^2 r_\varepsilon - s) \, ds.$$

The sum of the first three terms in N_ε we denote by $N1_\varepsilon$ and by $N2_\varepsilon$ the remaining terms. Our idea of proving the proposition is to apply Theorem 1 to the problem (3.18). Therefore it remains to verify Hypotheses H1–H4. It is obvious that the operator L_ε defined by $L_\varepsilon v \equiv \Box v + 2a(\tilde{u}_\varepsilon) B(d\tilde{u}_\varepsilon, dv) + h_\varepsilon v$ satisfies the conditions of Hypothesis H1. Because $m = 2, p = 2, q = 1, d = 3$, Hypothesis H2 is satisfied. From the exact

formula for R_ε it can be concluded that this satisfies H4. It only remains to verify the condition (H3b) for N_ε. Therefore, let $\{v_\varepsilon : \varepsilon \in (0,1]\} \in \tilde{X}_{2,2}(T,1)$. We can derive that

$$\begin{aligned}\|N1_\varepsilon\|_{L^\infty([0,T],H^2_\varepsilon)} \le V_\varepsilon^{2,2}(t,\varepsilon) \sum_{|\alpha|\le 1} \|D^\alpha v_\varepsilon\|_{L^\infty(D_t)} \\ + \varepsilon^2 \|v_\varepsilon\|^2_{L^\infty([0,t],L^4)} + \varepsilon^2 V_\varepsilon^{2,2}(t,\varepsilon) \left(\sum_{|\alpha|\le 1} \|D^\alpha v_\varepsilon\|_{L^\infty(D_t)} \right)^2 \\ + \varepsilon^4 \|v_\varepsilon\|_{L^\infty([0,t],L^4)} \sum_{|\alpha|\le 1} \|D^\alpha v_\varepsilon\|_{L^\infty(D_t)}, \quad (t,\varepsilon) \in [0,T] \times (0,1].\end{aligned}$$

In $N2_\varepsilon$ we examine only F, because the remaining terms are the same as in $N1_\varepsilon$. We begin from the following observations:

(i) $|F| \le 5\varepsilon^2 v_\varepsilon^2$, because $|a''| \le 10$.

(ii) For $|\alpha| = 1$,

$$\partial^\alpha F = (2\varepsilon^2)^{-1} \int_{\tilde{u}_\varepsilon}^{\tilde{u}_\varepsilon + \varepsilon^2 v_\varepsilon} a''(s)\, ds \partial^\alpha (\tilde{u}_\varepsilon + \varepsilon^2 v_\varepsilon) - \tfrac{1}{2} a''(\tilde{u}_\varepsilon)\, v_\varepsilon \partial^\alpha \tilde{u}_\varepsilon,$$

whence $|\partial^\alpha F| \le 5|v_\varepsilon|(2|\partial^\alpha \tilde{u}_\varepsilon| + \varepsilon^2 |\partial^\alpha v_\varepsilon|)$.

After differentiation of the formula for $\partial^\alpha F$ given above we derive

(iii) for $|\alpha| = 2$,

$$\begin{aligned}|\partial^\alpha F| \le 5\,|v_\varepsilon|\,|\partial^\alpha (\tilde{u}_\varepsilon + \varepsilon^2 v_\varepsilon)| + c \Bigg\{ |v_\varepsilon| + \sum_{\{(\beta,\gamma):\, \beta+\gamma=\alpha, |\beta,\gamma|=1\}} (|\partial^\beta \tilde{u}_\varepsilon| + \varepsilon^2 |\partial^\beta v_\varepsilon|) \\ \times [|\partial^\gamma \tilde{u}_\varepsilon| + |v_\varepsilon| (|\partial^\beta \tilde{u}_\varepsilon| + |\partial^\gamma v_\varepsilon|)] \Bigg\}.\end{aligned}$$

These allow us to derive

$$\begin{aligned}\|N2_\varepsilon\|_{L^\infty([0,t],H^2_\varepsilon)} \le c \Bigg\{ \left(1 + \sum_{|\alpha|\le 1} \|D^\alpha v_\varepsilon\|_{L^\infty(D_t)}\right) \\ \times \left(V_\varepsilon^{2,2}(t,\varepsilon) + \sum_{|\alpha|\le 2} \left\|\varepsilon^{|\alpha|} d^\alpha v_\varepsilon\right\|_{L^\infty([0,t],L^4)} \right) \\ + V_\varepsilon^{2,2}(t,\varepsilon) \left(\sum_{|\alpha|\le 1} \|D^\alpha v_\varepsilon\|_{L^\infty(D_t)} \right)^2 \Bigg\},\end{aligned}$$

which finishes the verification of (H3b).

Hence, we are in a position to apply Theorem 1 to the problem (3.18). The smoothness of the solution r_ε follows from the theory of propagation of the regularity.

Let us remark from (3.17) that $\{\|r_\varepsilon\|_{H^1(D_T)} : \varepsilon \in (0, \varepsilon_0]\}$ is bounded. Hence, there exists in $H^1(D_T)$ the weak limit r_0 of r_ε, when $\varepsilon \to 0$.

3.4. Investigation of r_0. After detailed examination we can distinguish in R_ε the following parts:

$$R_\varepsilon = a(u^0)\left\{ B(dv_0, dv_0) + \sum_{(k,l)\in K^2} B(du_k^1, du_l^1) \right.$$
$$+ 2\sum_{l\in K}\left[\sum_{k\in K} B(d\phi_l, dv_k^1) + \sum_{\bar{k}\in K^2} B(d\phi_l, du_{\bar{k}}^2)\right]\partial_\theta u_l^1$$
$$+ \sum_{l\in K}\left(\sum_{k\in K} B(du_l^1, d\phi_k)\,\partial_\theta v_k^1 + \sum_{\bar{k}\in K}[B(du_l^1, d\phi_{k_1})\,\partial_{\theta_1} u_{\bar{k}}^2\right.$$
$$\left.\left. + B(du_l^1, d\phi_{k_2})\,\partial_{\theta_2} u_{\bar{k}}^2]\right)\right\} + \sum_{l=1}^{3}\sum_{\bar{k}\in K^l} R_{\bar{k}}^1(t, x, \varepsilon^{-1}\phi_{\bar{k}}) + \varepsilon R1_\varepsilon + \sum_{k\in K} f_k^2,$$

where the functions $R_k^1(t,x,\theta), R_{\bar{k}}^2(t,x,\theta_{\bar{k}}), R_{\bar{k}}^3(t,x,\theta_{\bar{k}})$ have, respectively, the properties of $u_k^1, u_{\bar{k}}^2, u_{\bar{k}}^3$ described in assertions (ii) and (iii) of Proposition 8.

We divide the expression contained in brackets $\{\cdots\}$ into the following parts:

$$\{\cdots\} = R^0(t,x) + \sum_{l=1}^{3}\sum_{\bar{k}\in K^l} \tilde{R}_{\bar{k}}^l(t, x, \varepsilon^{-1}\phi_{\bar{k}}), \tag{3.19}$$

where $\tilde{R}_k^1, \tilde{R}_{\bar{k}}^2, \tilde{R}_{\bar{k}}^3$ have the same properties as $R_k^1, R_{\bar{k}}^2, R_{\bar{k}}^3$, respectively.

HYPOTHESIS H7. *For any integers $n_1, n_2, n_3 \neq 0$, and any $k_1 \neq k_2 \neq k_3$, $d(\sum_{l=1}^3 n_l\phi_{k_l}) \neq 0$ on D_T.*

PROPOSITION 10. *Under Hypotheses* H5–H7, *there are*

$$\lim_{\varepsilon\to 0}\int R_{\bar{k}}^1(t, x, \varepsilon^{-1}\phi_{\bar{k}})\,\varphi(t,x)\,dt\,dx = 0,$$
$$\lim_{\varepsilon\to 0}\int \tilde{R}_{\bar{k}}^1(t, x, \varepsilon^{-1}\phi_{\bar{k}})\,\varphi(t,x)\,dt\,dx = 0, \quad \text{for } \bar{k}\in K^1, \qquad l = 1, 2, 3,$$

and for any $\varphi \in C_0^\infty(D_T)$.

Proof. It will be enough to make the proof for $R_{\bar{k}}^3, \bar{k} \in K^3$. Therefore, let us consider the Fourier expansion

$$R_{\bar{k}}^3(t, x, \theta_{\bar{k}}) = \sum_{\bar{n}\in Z^3} a_{\bar{n}}(t,x)\exp\left(i\sum_{l=1}^{3} n_l\theta_l\right).$$

The properties of $R_{\bar{k}}^3$ lead to the following conclusions:

(a) $a_{(0,0,0)} = 0$;

(b) if $\bar{k} = (l,l,l), l \in K$, then $a_{\bar{n}} = 0$ for each $\bar{n}$ such that $\sum_{l=1}^3 n_l = 0$;

(c) if $\bar{k} = (k,l,l), k \neq l, k, l \in K$, then $a_{\bar{n}} = 0$ for each $\bar{n}$ such that $n_p = 0, n_q + n_r = 0$, where $p \neq q \neq r, p, q, r \in \{1,2,3\}$.

Let us consider $I_{\bar{n}}(\varepsilon) = \int a_{\bar{n}}(t,x)\varphi(t,x)\exp(i\varepsilon^{-1}\sum_{l=1}^3 n_l\phi_{k_l})\,d$ where $a_{\bar{n}} \neq 0$. Because of $d\phi_k \neq 0, k \in K$, of Remark 7, (3.5), and Hypothesis H7, in any case, there

will be $d(\sum_{l=1}^{3} n_l \phi_{k_l}) \neq 0$ on D_T. Hence, from the localization principle [4, Thm. 7.7.1], $\lim_{\varepsilon \to 0} I_{\bar{n}}(\varepsilon) = 0$. This gives the result.

PROPOSITION 11. *Let Hypotheses* H5–H7 *hold. Then* r_0 *is a distributional solution of the problem*

$$\Box r_0 - a(u^0) B(du^0, dr^0) - a'(u^0) B(du^0, du^0) r_0 = a(u^0) R^0, \tag{3.20}$$

at $D_T, r_0 = 0$ *for* $t < 0$.

Outline of proof. From (3.18) and the boundedness of r_ε in $H^1(D_T)$ we infer the following fact.

Fact 1. $\{||\Box r_\varepsilon||_{L^2(D_T)} : \varepsilon \in (0, \varepsilon_0]\}$ is bounded.

Now, it will be useful to rewrite (3.18) in the form

$$\begin{aligned} \Box r_\varepsilon - a(u^0) B(du^0, dr_\varepsilon) - a'(u^0) B(du^0, du^0) r_\varepsilon &= R_\varepsilon + \varepsilon R2_\varepsilon \\ &+ a(u^0) \sum_{k \in K} \partial_\theta u_k^1 (t, x, \varepsilon^{-1}\phi_k) B(d\phi_k, dr_\varepsilon) \\ &+ a'(u^0) r_\varepsilon \sum_{k \neq l} \partial_\theta u_k^1 (t, x, \varepsilon^{-1}\phi_k) \partial_\theta u_l^1 (t, x, \varepsilon^{-1}\phi_l) B(d\phi_k, d\phi_l). \end{aligned} \tag{3.21}$$

From Proposition 10 we ensure the following facts.

Fact 2. We have that

$$\lim_{\varepsilon \to 0} \int (R_\varepsilon + \varepsilon R2_\varepsilon) \varphi \, dt \, dx = \int a(u^0) R^0 \varphi \, dt \, dx, \qquad \varphi \in C_0^\infty(D_T).$$

Fact 3. We have that

$$\begin{aligned} \lim_{\varepsilon \to 0} \int a'(u^0) r_\varepsilon \varphi \sum_{k \neq l} \{ &\partial_\theta u_k^1 (\cdot, \varepsilon^{-1}\phi_k) \partial_\theta u_l^1 (\cdot, \varepsilon^{-1}\phi_l) \\ &\times B(d\phi_k, d\phi_l) \} \, dt \, dx = 0, \qquad \varphi \in C_0^\infty(D_T). \end{aligned}$$

Proof. The use of Fourier series expansions for $\partial_\theta u_j^1, j = k, l$, with respect to θ, changes the problem into

$$\lim_{\varepsilon \to 0} \int \exp(i\varepsilon^{-1}(m\phi_k + n\phi_l)) B(d\phi_k, d\phi_l) r_\varepsilon a'(u^0) \varphi \, dt \, dx = 0, \qquad m, n \neq 0.$$

Because of the strong convergence of r_ε to r_0 in $L^2(D_T)$, when $\varepsilon \to 0$, it will be enough to check that

$$\lim_{\varepsilon \to 0} \int \exp(i\varepsilon^{-1}(m\phi_k + n\phi_l)) \chi_n \, dt \, dx = 0, \qquad n \geq 1,$$

where $\chi_n \in C_0^\infty(D_T)$ and χ_n converges in $L^2(D_T)$ to $B(d\phi_k, d\phi_l) r_0 a'(u^0) \varphi$ when $n \to \infty$.

But the latter follows from the localization principle because $d(m\phi_k + n\phi_l) \neq 0$ (see (3.5)).

Fact 4. We have that

$$\lim_{\varepsilon \to 0} \int a(u^0) \varphi \sum_{k \in K} \partial_\theta u_k^1 (\cdot, \varepsilon^{-1}\phi_k) B(d\phi_k, dr_\varepsilon) \, dt \, dx = 0, \quad \text{for any } \varphi \in C_0^\infty(D_T).$$

The proof of this fact is given in [2] (see Lemma 3.2). We only inform the reader that the proof is based on one version of the compensated compactness theorem [3] and that Fact 1 is needed for this. Now, we multiply (3.21) by $\varphi \in C_0^\infty(D_T)$, integrate over D_T, and pass to the limit $\varepsilon \to 0$. Because of Facts 2–4 this yields the assertion of Proposition 11.

Appendix.

PROPOSITION A1. *Let $1 \le d < 2k, k$ integer, then*

$$\|v\|_{L^\infty(R^d)} \le c\varepsilon^{-d/2} \sum_{|\alpha|\le k} \left\|\varepsilon^{|\alpha|}\partial^\alpha v\right\|_{L^2(R^d)};$$

c is a universal constant.

Proof. We have that

$$|v(x)| = \left| \int e^{i\langle x,\xi\rangle} \hat{v}(\xi) \frac{\left(1+|\varepsilon\xi|^2\right)^{k/2}}{\left(1+|\varepsilon\xi|^2\right)^{k/2}} d\xi \right|$$
$$\le \left(\int \left(1+|\varepsilon\xi|^2\right)^{-k} d\xi \right)^{1/2} \sum_{|\alpha|\le k} \left\|\varepsilon^{|\alpha|}\partial^\alpha v\right\|_{L^2}.$$

When $d < 2k$, the first multiplier is finite and equals $c\varepsilon^{-d/2}$.

COROLLARY A2. *Let $d < 2m, \{v_\varepsilon : \varepsilon \in (0,\varepsilon_0]\} \in \tilde{X}_{m,m}(T,\varepsilon_0), m \ge 2$. Then*

$$\text{(A1)} \qquad \sum_{|\beta|\le m-1} \|D^\beta v_\varepsilon\|_{L^\infty(D_t)} \le c\varepsilon^{-d/2} \sum_{l=0}^{m-1} \sum_{|\alpha|\le m} \left\|\varepsilon^{|\alpha|}\partial^\alpha \partial_t^l v_\varepsilon\right\|_{L^\infty([0,t],H^{m-l-1})},$$

for each $(t,\varepsilon) \in [0,T] \times (0,\varepsilon_0]$.

PROPOSITION A3. *Let $1 \le d < 4k, k$ integer, $v \in H^k(R^d)$. Then*

$$\|v\|_{L^4(R^d)} \le c\varepsilon^{-d/4} \sum_{|\alpha|\le k} \left\|\varepsilon^{|\alpha|}\partial^\alpha v\right\|_{L^2(R^d)},$$

c is a universal constant.

Proof. We have that

$$\|v\|_{L^4} \le \|\hat{v}\|_{L^{4/3}} = \left(\int |\hat{v}|^{4/3} \frac{\left(1+|\varepsilon\xi|^2\right)^{2k/3}}{\left(1+|\varepsilon\xi|^2\right)^{2k/3}} d\xi \right)^{3/4}$$
$$\le \left(\int \left(1+|\varepsilon\xi|^2\right)^{-2k} \right)^{1/4} \left\| |\hat{v}|^{4/3} \left(1+|\varepsilon\xi|^2\right)^{2k/3} \right\|_{L^{3/2}}^{3/4}.$$

When $d < 4k$, the first multiplier is finite and equals $c\varepsilon^{-d/4}$.

COROLLARY A4. *Let $d < 4(m-q), 0 \le q \le m-1, m \ge 2$. Then for any $\{v_\varepsilon : \varepsilon \in (0,\varepsilon_0]\} \in \tilde{X}_{m,m}(T,\varepsilon_0)$,*

$$\text{(A2)} \qquad \sum_{\substack{|\alpha|\le q \\ |\beta|\le m-1}} \left\|\varepsilon^{|\beta|} d^\beta D^\alpha v_\varepsilon\right\|_{L^\infty([0,t],L^4)} \le c\varepsilon^{-d/4} \sum_{l=0}^{m-1} \sum_{|\alpha|\le m} \left\|\varepsilon^{|\alpha|}\partial^\alpha \partial_t^l v_\varepsilon\right\|_{L^\infty([0,t],H^{m-l-1})},$$

for each $(t, x) \in [0, T] \times (0, \varepsilon_0]$.

REFERENCES

[1] A. F. ANDREEV AND V. I. MARČENKO, *Symmetry and macroscopic dynamics of magnets*, Uspiekhi Fiz. Nauk, 130 (1980), pp. 39–63. (In Russian.)

[2] J. M. DELORT, *Oscillations semi-lineaires multiphases compatibles en dimension* 2 *ou* 3 *d'espace*, Comm. Partial Differential Equations, 16 (1991), pp. 845–872.

[3] P. GÉRARD, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1761–1794.

[4] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators*, Vol. 1 and Vol. 3, Springer-Verlag, Berlin, 1983 and 1985.

[5] F. JOHN, *Blow-up for quasi-linear wave equations in three space dimensions*, Comm. Pure Appl. Math. 34 (1982), pp. 29–51.

[6] J. L. LIONS, *Quelquees methodes de resolution des problemes aux limites non lineaires*, Gauthier-Villars, Paris, 1969.

[7] A. ŁADA, *Some solutions for the spin glass system and for systems originating from it*, Math. Methods Appl. Sci. 13 (1990), pp. 31–41.

[8] S. MIZOHATA, *The Theory of Partial Differential Equations*, Cambridge University Press, Cambridge, 1973.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1320–1356, September 1994

005

NONLINEAR PERTURBATION OF BOUNDARY VALUES FOR REACTION-DIFFUSION SYSTEMS: INERTIAL MANIFOLDS AND THEIR APPLICATIONS*

YOSHIHISA MORITA†, HIROKAZU NINOMIYA‡, AND EIJI YANAGIDA§

Abstract. The asymptotic behavior of solutions to a reaction-diffusion system with nonlinear boundary conditions is discussed. It is assumed that the boundary values are controlled by a positive parameter ϵ and that the boundary conditions reduce to the homogeneous Neumann boundary ones if ϵ tends to 0. Under appropriate conditions it is shown that for each small ϵ there exists an inertial manifold $\mathcal{M}_\epsilon$, that is, a finite-dimensional Lipschitz (or C^1) manifold which is invariant and attracts every solution exponentially. Moreover, it is proved that as $\epsilon \to 0$ the manifold $\mathcal{M}_\epsilon$ converges to $\mathcal{M}_0$, which is the one for the homogeneous Neumann boundary conditions. The dynamics on the manifold are investigated through the reduced ordinary differential equation on it, called the inertial form, for specific cases; for instance, a specific example shows that the boundary values induce a relaxation-oscillating periodic motion in the manifold while every solution converges to a steady state in the case $\epsilon = 0$.

Key words. nonlinear boundary condition, inertial manifold, inertial form, cone property, relaxation oscillation

AMS subject classifications. 35K57, 35K60, 35B40, 35B10

1. Introduction.

Inertial manifold theory for evolution equations has been an extensive subject since the pioneering work by [9] was published. Here the inertial manifold is defined by a finite-dimensional Lipschitz (or C^1) manifold which is positively invariant and attracts every solution exponentially ([5], [6], and [7]). When such a manifold exists, the asymptotic behavior of solutions can be described by a finite-dimensional system (ordinary differential equation (ODE) system) on the manifold. Therefore, this theory plays a crucial role in the understanding of some dynamical properties of the equation. In particular, for reaction-diffusion equations in a bounded domain, the inertial manifold theory was developed by the work [18] under the Neumann, Dirichlet, or periodic boundary conditions. In a general situation, however, it is difficult to verify whether the evolution equations possess their inertial manifolds or not [17]. Moreover, there are not many examples for which the dynamics on the manifold are exhibited concretely (cf. [19] and [20]).

In this paper we will discuss the existence of the inertial manifold and its applications for a system of reaction-diffusion equations in a bounded domain supplemented with a nonlinear boundary condition. Let $\Omega \subset \mathbb{R}^n$ $(1 \le n \le 3)$ be a bounded domain with a smooth boundary $\partial\Omega$, and consider the equation

$$\frac{\partial u}{\partial t} = D\Delta u + F(u), \qquad (t, x) \in \mathbb{R}_+ \times \Omega, \tag{1.1}$$

* Received by the editors May 28, 1992; accepted for publication (in revsied form) April 20, 1993.

† Department of Applied Mathematics and Informatics, Ryukoku University, Seta Ohtsu 520-21, Japan.

‡ Department of Applied Physics, Tokyo Institute of Technology, Oh-Okayama, Meguro-ku, Tokyo 152, Japan.

§ Department of Information Science, Tokyo Institute of Technology, Oh-Okayama Meguro-ku Tokyo 152, Japan.

subject to the nonlinear boundary condition

$$\frac{\partial u}{\partial \nu} = \epsilon G(u(t,x)), \qquad (t,x) \in \mathbb{R}_+ \times \partial\Omega. \tag{1.2}_\epsilon$$

Here $\mathbb{R}_+ = (0,\infty), D = \text{diag}(d_1, \dots, d_m) \quad (d_i > 0),$

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}, \quad \Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_n^2}, \quad (x_1, x_2, \dots, x_n) \in \mathbb{R}^n,$$

and ϵ is a nonnegative parameter. The functions F and G are assumed to be so smooth that (1.1) with $(1.2)_\epsilon$ may have a unique local solution provided that initial data are chosen in an appropriate function space (for the existence theory, see [10]). If the equation is dissipative, that is, there exists a bounded set $\mathcal{B}$ (absorbing set) such that every solution enters $\mathcal{B}$ at a finite time $t = t_0(\mathcal{B})$ and remains for $t \geq t_0(\mathcal{B})$, then we can assert the existence of the maximal compact invariant set attracting every solution, called a "global attractor" (for the condition of the existence of the global attractor, see [12] and [24]).

Our main purpose in the study of the system (1.1) with $(1.2)_\epsilon$ is to show that, for each small ϵ, there exists a C^1-inertial manifold $\mathcal{M}_\epsilon$ containing the global attractor under a certain condition. We also show that the manifold $\mathcal{M}_\epsilon$ converges to $\mathcal{M}_0$ as $\epsilon \to 0$, where $\mathcal{M}_0$ is the manifold for the Neumann zero boundary condition. Remark that although our interest is mainly in the dissipative equations, our result is also applicable to studying some dynamical property of nondissipative equations (see §4).

Before stating our main results, we introduce some notations. We denote by $H^k(\Omega, \mathbb{R}^m)$, $k = 1, 2, \dots$, the subspace of $L^2(\Omega, \mathbb{R}^m)$, whose derivatives of order up to k belong to $L^2(\Omega, \mathbb{R}^m)$. We will often abbreviate $L^2(\Omega, \mathbb{R}^m)$ and $H^k(\Omega, \mathbb{R}^m)$ as L^2 and H^k, respectively. Let $\{\lambda_j\}_{j=1,2,..}$ and $\{\phi_j\}_{j=1,2,..}$ be eigenvalues and the corresponding normalized eigenvectors of $A = -D\Delta + I$ with the domain

$$D(A) = \left\{ u \in H^2; \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega \right\},$$

that is,

$$A\phi_j = \lambda_j \phi_j,$$

and

$$\begin{aligned} &1 = \lambda_1 = \lambda_2 = \cdots = \lambda_m < \lambda_{m+1} < \lambda_{m+2} < \cdots, \\ &(\phi_j, \phi_k) = \int_\Omega \phi_k(x) \cdot \phi_j(x) dx = \delta_{jk}, \end{aligned} \tag{1.3}$$

where $u \cdot v$ $(u, v \in \mathbb{R}^m)$ denotes the inner product of $\mathbb{R}^m$ and δ_{jk} is the Kronecker delta. Let us define the projection operators in L^2 as follows:

$$P : u \longmapsto \sum_{j=1}^{N} (u, \phi_j)\phi_j, \qquad Q = I - P. \tag{1.4}$$

We smoothly modify the functions F and G outside the region

$$\{u \in \mathbb{R}^m; |u| < R\}$$

so that

$$F(u) = -u, \quad G(u) = 0 \quad \text{for } u, \ |u| > 2R, \tag{C.1}$$

where $R > 0$ is taken so large that $\{u \in \mathbb{R}^m; |u| < R\}$ contains all the bounded solutions in our interest. One may think that the condition (C.1) seems to be too restrictive. However, as far as a bounded solution is concerned, this modification does not cause any change by taking R sufficiently large; see Corollary B. Under the condition (C.1) the equation has a unique global solution $u(t,x;u_0) \in C([0,\infty);L^2) \cap C((0,\infty);H^2)$ for any initial data $u_0 \in H^2$. Moreover, it is shown in §2 (Proposition 2.2) that there is an absorbing set $\mathcal{B}_1$ in H^2. We define a solution map $S_\epsilon(t) : H^2 \longrightarrow H^2$ $(t \geq 0)$ by $S_\epsilon(t)u_0 = u(t,\cdot\,;u_0)$. Note that the mappings

$$u(x) \longrightarrow F(u(x)), \qquad u(x) \longrightarrow G(u(x)) \tag{1.5}$$

define smooth functions of H^2 into itself and $H^{\frac{3}{2}}(\partial\Omega;\mathbb{R}^m)$ into itself, respectively.

Now we are ready to state our main result.

THEOREM A. *In addition to the condition* (C.1), *assume that there exists a positive integer N such that*

$$\lambda_{N+1} - \lambda_N > C_F, \tag{C.2}$$

where C_F is some positive constant which is determined by the function F (see §3.1 *for the definition of C_F). Let $\mathcal{B}_1$ be an H^2-bounded absorbing set. Then there is a positive ϵ_0 such that for each $\epsilon \in [0,\epsilon_0]$ there exists a C^1-inertial manifold $\mathcal{M}_\epsilon \subset \mathcal{B}_1$ satisfying the following:*

(i) *$\mathcal{M}_\epsilon$ is positively invariant, that is, $S_\epsilon(t)\mathcal{M}_\epsilon \subset \mathcal{M}_\epsilon$ $(t \geq 0)$ and there exists a C^1 function $\Phi_\epsilon : PH^2 \longmapsto (I-P)H^2$ such that $\mathcal{M}_\epsilon = graph(\Phi_\epsilon) \cap \mathcal{B}_1$.*

(ii) *For any solution $u(t,\cdot)$ in $\mathcal{B}_1$ of* (1.1) *with* $(1.2)_\epsilon$ *there is a $p_0 \in PH^2$ such that*

$$\|u(t,\cdot) - S_\epsilon(t)(p_0 + \Phi_\epsilon(p_0))\|_{H^2} \leq Ce^{-\nu t}\|u(0,\cdot) - p_0 - \Phi_\epsilon(p_0)\|_{H^2} \qquad (t \geq 0),$$

where ν and C are some positive constants independent of the initial value $u(0,\cdot)$.

(iii) *As $\epsilon \to 0$,*

$$\Phi_\epsilon(p) \underset{H^2}{\longrightarrow} \Phi_0(p), \quad \frac{\partial\Phi_\epsilon(p)}{\partial p}\eta \underset{H^2}{\longrightarrow} \frac{\partial\Phi_0(p)}{\partial p}\eta \quad (\eta \in H^2)$$

uniformly in $p \in PH^2$.

(iv) *The inertial form (the reduced* ODE*) on the manifold is written as*

$$p_t + (A-I)p = PF(p+\Phi_\epsilon(p)) + \epsilon\sum_{j=1}^{N}\left(\int_{\partial\Omega} G(p+\Phi_\epsilon(p))\cdot\phi_j dS\right)\phi_j. \tag{1.6}$$

Because of the nonlinear boundary condition we encounter some technical difficulties for the proof of this theorem. First, the equation (1.1) with $(1.2)_\epsilon$ does not admit the variation-of-constants formula or the abstract ODE in any state space. Therefore, it seems to be difficult to apply the inertial manifold theorem [9], based on the Lyapunov–Perron method, to our case. To overcome such a difficulty, we will use the graph transformation method due to Hadamard. Such an approach has been taken by Mallet-Paret and Sell [18] for solving some delicate and important problems regarding the existence of an inertial manifold for a scalar reaction-diffusion equation with the conventional boundary condition (see also Bates and Jones [3]). The graph transformation method is based on some geometric property of the semiflow, called the cone property, and it does not require the variation-of-constants formula for proving the existence of the manifold (cf. [22]). Second, we have to discuss the existence problem

of the inertial manifold not in the L^2-framework but in the H^2-framework in order to investigate the dynamics on the manifold through the inertial form. Indeed, as seen in (iv) of Theorem A, the inertial form has the integrated terms over the boundary. This implies that if one constructs the manifold in the L^2-framework as [18] does, even the Lipschitz continuity of the inertial form cannot be guaranteed. However, such a difficulty can be overcome provided that one constructs the manifold in a more regular space. In this paper, as mentioned above, we will discuss it in the H^2-framework because it seems to be appropriate for (1.1) and $(1.2)_\epsilon$.

We will give some corollaries of Theorem A which are more convenient for applications. The next corollary holds even if the condition (C.1) is not satisfied.

COROLLARY B. *Let $u_\epsilon(t)$ be any solution of* (1.1) *with* $(1.2)_\epsilon$ *satisfying* $\|u_\epsilon\|_{L^\infty} \leq R$ $(-\infty < t < \infty,\ 0 \leq \epsilon \leq \epsilon_0)$. *Then this solution is in $\mathcal{M}_\epsilon$, constructed in Theorem A.*

If we can take $N = m$ in Theorem A, we have

$$Pu = \frac{1}{|\Omega|} \begin{pmatrix} \int_\Omega u_1(x)dx \\ \vdots \\ \int_\Omega u_m(x)dx \end{pmatrix}, \tag{1.7}$$

and we can show that $\Phi_\epsilon(p)$ converges to a zero function. In other words, the manifold converges to the m-dimensional subspace which consists of constant functions in x-variables. More precisely, the next result holds.

COROLLARY C. *Let σ_2 be the second eigenvalue of $-\Delta$ with the homogeneous Neumann boundary condition. Assume the same conditions in Theorem A under N=m, namely,* (C.1) *and*

$$\lambda_{m+1} = \min\{d_1, \dots, d_m\}\sigma_2 > C_F. \tag{C.2$'$}$$

Then the function Φ_ϵ can be expanded by ϵ as

$$\Phi_\epsilon(p) = \epsilon\Phi_1 + \epsilon\tilde{\Phi}_\epsilon(\epsilon, p),$$

where $\tilde{\Phi}_\epsilon(\epsilon, p)$ and its first derivative in p are bounded in p. Moreover, the inertial form is written as

$$p_t = F(p) + \epsilon\frac{|\partial\Omega|}{|\Omega|}DG(p) + \epsilon R(D, \epsilon, p), \tag{1.8}$$

$$|R(D, \epsilon, p)| = O(\epsilon d_*), \quad \left|\frac{\partial R(D, \epsilon, p)}{\partial p}\right| = O(\epsilon^\delta d_*) \quad (\textit{uniformly in } p),$$

where δ is some positive constant and $d_* = \min(d_1, \dots, d_m) > 0$.

We see from this corollary that with $\epsilon = 0$ the ODE on the manifold is given by

$$p_t = F(p). \tag{1.9}$$

This implies that in the case of the homogeneous Neumann boundary condition, the asymptotic behavior of the solutions is completely determined by the ODE of (1.9). This observation is consistent with the work by Conway, Hoff, and Smoller [8] and Hale [11]. One may think that the dynamical structure of the ODE of (1.8) with $\epsilon > 0$ is the same as that of (1.9). We can, however, present specific examples for which their dynamics are drastically changed through the presence of the boundary condition $(1.2)_\epsilon$ (see §4).

Finally, we note that our method is applicable to the case where the diffusion constants depend on the parameter ϵ. For example, the following corollary holds.

COROLLARY D. *Assume that the diffusion constants are given by*

$$d_i = \frac{\tilde{d}_i}{\epsilon}, \qquad i = 1, 2, \dots, m.$$

If (C.1) *holds and* ϵ *is sufficiently small, then the same conclusion as in Theorem A holds with* $N = m$, *and the inertial form is given by*

$$p_t = F(p) + \alpha \tilde{D} G(p) + O(\epsilon), \tag{1.10}$$

where

$$\tilde{D} = \operatorname{diag}(\tilde{d}_1, \tilde{d}_2, \dots, \tilde{d}_m), \qquad \alpha = \frac{|\partial\Omega|}{|\Omega|}. \tag{1.11}$$

In relation to Corollary D we remark on the study by Hale and Rocha [13]. They considered the case where the boundary condition is linearly perturbed and obtained some results similar to Corollary D. We can derive some of their results from a direct application of Corollary D. Their approach is based on studying the behavior of eigenvalues of $-D\Delta$ with their linear boundary condition as $\epsilon \to 0$. So we cannot apply their technique to the case of nonlinear boundary conditions. Moreover, in order to prove Corollary D (not to speak of Theorem A) we have to evaluate the boundary values with much more attention, which will be seen in the successive sections.

2. Preliminary.

2.1. Notation and auxiliary inequalities. We introduce function spaces and their norms which will appear often in this paper.

$L^p(\Omega; \mathbb{R}^m)$ = the set of pth power integrable functions from Ω into $\mathbb{R}^m$,

$$\|u\|_{L^p} = \left\{ \int_\Omega \left(\sum_{j=1}^m u_j^2 \right)^{\frac{p}{2}} dx \right\}^{\frac{1}{p}}, \qquad u \in L^p(\Omega; \mathbb{R}^m),$$

$W^{k,p}(\Omega; \mathbb{R}^m)$ = the set of functions whose derivatives up to k in the distribution sense belong to $L^p(\Omega; \mathbb{R}^m)$,

$$\|u\|_{W^{k,p}} = \left(\sum_{j_1+\cdots+j_n \le k} \left\| \frac{\partial^{j_1+\cdots+j_n} u}{\partial x_1^{j_1} \dots \partial x_n^{j_n}} \right\|_{L^p}^p \right)^{\frac{1}{p}}.$$

We simply write

$$L^p = L^p(\Omega; \mathbb{R}^m), \qquad W^{k,p} = W^{k,p}(\Omega; \mathbb{R}^m),$$

and in the case $p = 2$,

$$H^k = H^k(\Omega; \mathbb{R}^m) = W^{k,2}(\Omega; \mathbb{R}^m).$$

In addition to the above notation we will abbreviate the L^2-norm, i.e.,

$$\|u\| = \|u\|_{L^2} \qquad (u \in L^2).$$

Similarly, we define the function spaces $L^p(\partial\Omega; \mathbb{R}^m)$, $W^{k,p}(\partial\Omega; \mathbb{R}^m)$, and $H^k(\partial\Omega; \mathbb{R}^m)$ and write

$$L^p(\partial\Omega) = L^p(\partial\Omega; \mathbb{R}^m), \quad W^{k,p}(\partial\Omega) = W^{k,p}(\partial\Omega; \mathbb{R}^m), \quad H^k(\partial\Omega) = H^k(\partial\Omega; \mathbb{R}^m).$$

Their norms are denoted by

$$\|\cdot\|_{L^p(\partial\Omega)}, \quad \|\cdot\|_{W^{k,p}(\partial\Omega)}, \quad \|\cdot\|_{H^k(\partial\Omega)},$$

respectively. Define the operator $\tilde{A}$ by

$$\tilde{A} = -D\Delta + I,$$

with the domain

$$D(\tilde{A}) = H^2.$$

Let the operator A be as in the Introduction. Then

$$\tilde{A}u = Au, \qquad u \in D(A).$$

Note that

$$PAu = APu, \qquad P\tilde{A}u = \tilde{A}Pu - \sum_{j=1}^{N} \Big(\int_{\partial\Omega} \Big(D\frac{\partial u}{\partial \nu} \Big) \cdot \phi_j dS \Big) \phi_j.$$

We define the fractional power A^s $(s \geq 0)$ of A as usual, that is,

$$A^s u = \sum_{j=1}^{\infty} \lambda_j^s (u, \phi_j)\phi_j, \qquad u \in D(A^s).$$

Denote

$$(u,v)_1 = (D^{\frac{1}{2}}\nabla u, D^{\frac{1}{2}}\nabla v) + (u,v), \qquad u, v \in H^1,$$
$$\|u\|_1 = (u,u)_1^{\frac{1}{2}}.$$

This norm $\|\cdot\|_1$ gives equivalent topology to that of H^1-space. By the assumption $1 \leq n \leq 3$ we have

$$D(A^s) = H^{2s}, \qquad s \leq \frac{3}{4},$$

in particular, $D(A^{\frac{1}{2}}) = H^1$ and

$$\|u\|_1 = \|A^{\frac{1}{2}}u\|, \qquad u \in D(A^{\frac{1}{2}}).$$

We place some inequalities coming from Sobolev embeddings, trace theory, and [1]:

$$\begin{cases} \|u\|_{W^{k,p}} \leq C\|u\|_{H^l} & \Big(k - \dfrac{n}{p} \leq l - \dfrac{n}{2}\Big), \\ \|u\|_{H^k} \leq C\|u\|_{H^l}^{\theta}\|u\|^{1-\theta} & (k \leq l\theta, 0 \leq \theta \leq 1), \\ \|u\|_{L^\infty} \leq C\|u\|_{H^2}, & \\ \|u\|_{H^k(\partial\Omega)} \leq C\|u\|_{H^{k+\frac{1}{2}}}, & \\ \|u\|_{H^l(\Omega)} \leq C\left(\|\Delta u\|_{H^{l-2}(\Omega)} + \|u\|_{H^{l-2}(\Omega)} + \left\| \dfrac{\partial u}{\partial \nu} \right\|_{H^{l-1}(\partial\Omega)} \right) & (l \geq 0). \end{cases}$$

From the above inequalities we have

$$\begin{aligned}\|\tilde{A}u\|^2 &= \int_\Omega |-D\Delta u+u|^2 dx \\ &= \|D\Delta u\|^2 + \|u\|^2 - 2\int_\Omega u\cdot D\Delta u dx \\ &\geq \|D\Delta u\|^2 + \|D^{\frac{1}{2}}\nabla u\|^2 + \|u\|^2 - 2\left\|D\frac{\partial}{\partial\nu}u\right\|_{L^2(\partial\Omega)}\|u\|_{L^2(\partial\Omega)} \\ &\geq C\|u\|_{H^2}^2 - C'\|u\|^2 - C'\left\|\frac{\partial u}{\partial\nu}\right\|_{H^1(\partial\Omega)}^2.\end{aligned}$$

The next lemma immediately follows from the above inequalities and the definitions of A and $\tilde{A}$.

LEMMA 2.1. *For $u, v \in H^2$,*

(i) $C^{-1}\|u\|_{H^2} \leq \|\tilde{A}u\| + C\|u\| + C\left\|\frac{\partial u}{\partial\nu}\right\|_{H^1(\partial\Omega)} \leq C'\|u\|_{H^2} + C\left\|\frac{\partial u}{\partial\nu}\right\|_{H^1(\partial\Omega)}$,

(ii) $(u,\tilde{A}v) + \int_{\partial\Omega} u\cdot D\frac{\partial v}{\partial\nu}dS = (u,v)_1 = (\tilde{A}u,v) + \int_{\partial\Omega} v\cdot D\frac{\partial u}{\partial\nu}dS$,

(iii) $\|A^s Pu\| \leq \lambda_N^s\|Pu\|$ *for* $s \leq 1$,

(iv) $\|Qu\|_1 = \|A^{\frac{1}{2}}Qu\| \geq \lambda_N^{\frac{1}{2}}\|Qu\|$,

(v) $\|u\|_{W^{k,p}} \leq C\left(\|\tilde{A}u\|_1^\theta + \|u\|^\theta + \left\|\frac{\partial u}{\partial\nu}\right\|_{H^1(\partial\Omega)}^\theta\right)\|u\|^{1-\theta}$ $\left(k - \frac{n}{p} + \frac{n}{2} \leq 3\theta \leq 3\right)$,

(vi) $\|u\|_{L^\infty} \leq C\|\tilde{A}u\| + C\|u\| + C\left\|\frac{\partial u}{\partial\nu}\right\|_{H^1(\partial\Omega)}$,

(vii) $\|u\|_{L^2(\partial\Omega)} \leq C\|u\|_1$.

Especially if $\|\frac{\partial u}{\partial\nu}\|_{H^1(\partial\Omega)} \leq C\epsilon\|u\|_{H^1(\partial\Omega)}$, then the boundary data appearing in the inequalities (i), (v), *and* (vii) *can be removed.*

For simplicity of notation we will, hereafter, denote most of the constants in the computation by C_N if they depend on λ_N; otherwise we will denote them by C. We also abbreviate the inner product $u\cdot v$ of $\mathbb{R}^m$ as uv. To avoid troublesome expressions, we write

$$|F| = \sup_{u\in\mathbb{R}^m}|F(u)|, \qquad |F|_k = \sum_{\substack{\beta=(\beta_1,\dots,\beta_m)\\ \beta_1+\cdots+\beta_m\leq k}} \sup_{u\in\mathbb{R}^m}\left|\frac{\partial^\beta F}{\partial {u_1}^{\beta_1}\dots\partial_{u_m}^{\beta_m}}\right|$$

for any function $F(\cdot)$ from $\mathbb{R}^m$ into $\mathbb{R}^m$.

2.2. Existence of an absorbing set. We set $\tilde{F}(u) = F(u) + u$. By the condition (C.1)

(C.1′) $$\tilde{F}(u) = 0, \quad G(u) = 0 \quad \text{for } |u| > 2R.$$

Then the equation (1.1) is written as

(1.1′) $$\frac{\partial u}{\partial t} - D\Delta u + u = \tilde{F}(u).$$

Taking the inner product between u and (1.1′) and integrating by parts yield

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u\|^2 &= -(\tilde{A}u,u)+(\tilde{F}(u),u)\\ &\le -\|u\|_1^2+\epsilon|DG|\ |\partial\Omega|^{\frac{1}{2}}\|u\|_{L^2(\partial\Omega)}+|\tilde{F}|\ |\Omega|^{\frac{1}{2}}\|u\|\\ &\le -\frac{1}{2}\|u\|_1{}^2+\frac{1}{2}R_0^2,\end{aligned}$$

where $R_0=C(|\tilde{F}|+|\epsilon G|)$. Hence

$$\mathcal{B}_0=\{u\in H^2;\ \|u\|\le R_0\}$$

is an absorbing set.

Next we discuss the existence of an H^2-bounded absorbing set. By (C.1′), we may consider the solution with $\|u\|_{L^\infty}\le R$. In a manner similar to the L^2 case above,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\tilde{A}u\|^2 &= -(\tilde{A}^2u,\tilde{A}u)+(\tilde{A}\tilde{F}(u),\tilde{A}u)\\ &= -\|\tilde{A}u\|_1^2+\int_{\partial\Omega}\tilde{A}uD\frac{\partial}{\partial\nu}\tilde{A}u\,dS+(D\Delta\tilde{F}(u),\tilde{A}u)+(\tilde{F}(u),\tilde{A}u)\\ &\le -\|\tilde{A}u\|_1^2+\int_{\partial\Omega}\tilde{A}uD\frac{\partial}{\partial\nu}\Big(-u_t+\tilde{F}(u)\Big)dS\\ &\quad+\|D\tilde{F}'(u)\Delta u\|\ \|\tilde{A}u\|+|D\tilde{F}''(u)|\|\nabla u\|_{L^4}^2\|\tilde{A}u\|+\|\tilde{F}(u)\|\ \|\tilde{A}u\|\\ &\le -\|\tilde{A}u\|_1^2+\int_{\partial\Omega}\tilde{A}uD\frac{\partial}{\partial\nu}\Big(-u_t+\tilde{F}(u)\Big)dS\\ &\quad+C|\tilde{F}'|(\|\tilde{A}u\|+\|u\|)\|\tilde{A}u\|+C|\tilde{F}''|\|u\|_{W^{1,4}}^2\|\tilde{A}u\|+C|\tilde{F}|\|\tilde{A}u\|.\end{aligned}\tag{2.1}$$

Prepare the inequalities which follow from Lemma 2.1 (v):

$$\|u\|_{W^{1,4}}\le C(\|\tilde{A}u\|_1^{\frac{7}{12}}+\|u\|^{\frac{7}{12}}+\|\epsilon G\|_{H^1(\partial\Omega)}^{\frac{7}{12}})\|u\|^{\frac{5}{12}}\le CR^{\frac{5}{12}}(\|\tilde{A}u\|_1^{\frac{7}{12}}+R^{\frac{7}{12}}+\epsilon^{\frac{7}{12}}),$$

$$\|\tilde{A}u\|\le C(\|\tilde{A}u\|_1^{\frac{2}{3}}+\|u\|^{\frac{2}{3}}+\|\epsilon G\|_{H^1(\partial\Omega)}^{\frac{2}{3}})\|u\|^{\frac{1}{3}}\le CR^{\frac{1}{3}}(\|\tilde{A}u\|_1^{\frac{2}{3}}+R^{\frac{2}{3}}+\epsilon^{\frac{2}{3}}).$$

By these inequalities, we have

$$\|u\|_{W^{1,4}}^2\|\tilde{A}u\|\le CR^{\frac{7}{6}}(\|\tilde{A}u\|_1^{\frac{11}{6}}+R^{\frac{11}{6}}+\epsilon^{\frac{2}{3}}).$$

The term integrated on the boundary of (2.1) is

$$\frac{\partial}{\partial\nu}(-u_t+\tilde{F}(u))=-\epsilon G'(u)u_t+\tilde{F}'(u)\frac{\partial u}{\partial\nu},$$

by which

$$\begin{aligned}&\left|\int_{\partial\Omega}\tilde{A}uD\frac{\partial}{\partial\nu}(-u_t+\tilde{F}(u))dS\right|\\ &\qquad\le\epsilon\|u\|_{H^2(\partial\Omega)}(|DG'|\ \|-\tilde{A}u+\tilde{F}(u)\|_{L^2(\partial\Omega)}+|D\tilde{F}'||G|\ |\partial\Omega|)\\ &\qquad\le\epsilon C(1+\|\tilde{A}u\|_1^2).\end{aligned}$$

Substituting the above inequalities into (2.1), we get

$$\frac{1}{2}\frac{d}{dt}\|\tilde{A}u\|^2\le-\frac{1}{2}\|\tilde{A}u\|_1^2+\frac{1}{2}R_1^2,$$

where

$$R_1 = C(|\tilde{F}|_2 + |\tilde{F}|_2^{12} + \epsilon^{\frac{1}{3}}).$$

Hence we have the following proposition.

PROPOSITION 2.2. *There exists a positive constant ϵ_0 such that*

$$\mathcal{B}_1 = \{u \in H^2;\ \|u\| \le R_0,\ \|\tilde{A}u\| \le R_1\}$$

is an absorbing set for each $\epsilon \le \epsilon_0$, where R_1 is a constant that depends only on ϵ_0.

Considering this proposition, we make a modification of the equation outside $\mathcal{B}_1$. Let R' be a number satisfying

$$\mathcal{B}_1 \subset \{u = p + q \in H^2;\ p = Pu,\ q = Qu,\ \|Ap\|^2 + \|q\|_1^2 \le R'^2\}.$$

From now on we consider the following modified equation instead of (1.1) and $(1.2)_\epsilon$:

$$\text{(2.2)} \qquad \begin{cases} p_t + Ap = f_1(p+q), \\ q_t + \tilde{A}q = f_2(p+q), \\ \dfrac{\partial q}{\partial \nu} = \epsilon g(p+q), \end{cases}$$

where

$$\begin{cases} f_1(p+q) = \chi(\|Ap\|^2 + \|q\|_1^2)\left\{P\tilde{F}(p+q) + \displaystyle\sum_{j=1}^{N} \epsilon \int_{\partial\Omega} DG(p+q)\phi_j dS \phi_j\right\}, \\ f_2(p+q) = \chi(\|Ap\|^2 + \|q\|_1^2)\tilde{F}(p+q) - f_1(p+q), \\ g(p+q) = \chi(\|Ap\|^2 + \|q\|_1^2)G(p+q), \end{cases}$$

and χ is a smooth function satisfying

$$\chi(r) = \begin{cases} 1 & (0 \le r \le R'^2), \\ 0 & (r \ge 4R'^2). \end{cases}$$

We simply say that $u(t) = p(t) + q(t)$ is a solution to (2.2) for $p(t), q(t)$ satisfying (2.2). This modification also guarantees the existence of an absorbing set for (2.2). Furthermore, we see that there is an H^4-bounded absorbing set.

PROPOSITION 2.3. *There exists an absorbing set $\mathcal{B}_2$ ($\supset \mathcal{B}_1$) for* (2.2),

$$\mathcal{B}_2 = \{u = p + q \in H^2;\ \|u\| \le R_2, \|\tilde{A}u\| \le 2R_1\}.$$

Moreover, there exists a number $R_3 > 0$ such that

$$\mathcal{B} = \{u = p + q \in H^4;\ \|u\| \le R_2, \|\tilde{A}u\| \le 2R_1, \|\tilde{A}^2 q\| \le R_3\}$$

is an H^4-absorbing set for (2.2), *where R_3 is chosen independently of N.*

Proof. The estimates of $\|u\|, \|\tilde{A}u\|$ are obtained similarly. Now we prove the

existence of a bound of $\|\tilde{A}^2 q\|$. By taking the scalar product, we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\tilde{A}^2 q\|^2 =&(\tilde{A}^2 q, \tilde{A}^2 q_t)\\ =&(\tilde{A}^2 q, \tilde{A} q_t)_1 + \int_{\partial\Omega} \tilde{A}^2 q D\frac{\partial}{\partial \nu}\tilde{A} q_t dS\\ =&-\|\tilde{A}^2 q\|_1 + \int_{\partial\Omega} \tilde{A}^2 q D\frac{\partial}{\partial \nu}(-q_{tt} + f_2(p+q)_t) dS\\ &+ (\tilde{A}^2 q, \tilde{A} f_2(p+q))_1.\end{aligned}$$

We use the following lemma.

LEMMA 2.4. *For a solution $p+q$ of (2.2) with $\|\tilde{A} q\| \le 2R_1$,*

(i) $\|\tilde{A} f_2(p+q)\|_1 \le \{C(1+|\tilde{F}|_3)(1+R_1)^3 + \epsilon C\}(1+\lambda_N^{\frac{1}{2}} + \|\tilde{A} q\|_1)$,

(ii) $\|D\frac{\partial}{\partial \nu}(-q_{tt} + f_2(p+q)_t)\|_{L^2(\partial\Omega)} \le \epsilon C_N(1+\|\tilde{A}^2 q\|_1)$.

The proof of this lemma is given in the Appendix.

Set

$$C_1 = C(1+|\tilde{F}|_1)(1+R_1)^3 + \epsilon C.$$

By virtue of Lemma 2.4 we have the following inequalities:

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\tilde{A}^2 q\|^2 \le& -\|\tilde{A}^2 q\|_1^2 + \epsilon C_N(\|\tilde{A}^2 q\|_1 + R_1)(1+\|\tilde{A}^2 q\|_1)\\ &+ C_1\|\tilde{A}^2 q\|_1(1+\lambda_N^{\frac{1}{2}} + R_1^{\frac{1}{2}}\|\tilde{A}^2 q\|_1^{\frac{1}{2}})\\ \le& -\frac{1}{2}\|\tilde{A}^2 q\|_1^2 + C_1^4 R_1^2 + \epsilon C_N,\end{aligned}$$

where we used Lemma 2.1 (v) and Schwarz's inequality to obtain the second inequality. Hence the set

$$\{p+q; \|\tilde{A} q\| \le 2R, \|\tilde{A}^2 q\| \le R_3\} \qquad \left(R_3^2 \ge \frac{2C_1^4 R_1^2 + 2\epsilon C_N}{\lambda_{N+1}}\right)$$

is a positively invariant absorbing set. □

3. Proof of the main theorem. In this section we will give the proof of the main theorem. We first introduce the *cone property* in §3.1. Then using the cone property, we show the existence of an inertial manifold with the Lipschitz continuity in §3.2. C^1-smoothness of the manifold will be discussed in §3.3. In §3.4 we prove Theorem A (iii). The proofs of Corollaries C and D are also given in §3.4.

3.1. Cone property. Let $u_i(t)$ $(i=1,2)$ be any two solutions of (2.2):

$$u_1(t) = p_1(t) + q_1(t), \qquad u_2(t) = p_2(t) + q_2(t).$$

Define the sets

$$\begin{cases} \mathcal{C} = \{p+q \in H^2;\ \|\tilde{A} q\| \ge \|Ap\|\},\\ \tilde{\mathcal{B}} = \{p+q \in H^4;\ \|\tilde{A}^2 q\| \le 2R_3\}. \end{cases} \tag{3.1}$$

The next property is called the cone property.

PROPOSITION 3.1. *Assume that two solutions* $u_1(t), u_2(t)$ *of* (2.2) *remain in* $\tilde{\mathcal{B}}$ *for* $t_1 \le t \le t_2$. *Set* $p(t) = p_1(t) - p_2(t), q(t) = q_1(t) - q_2(t)$. *Then the following property holds:*

(i) *if* $p(t_2) + q(t_2) \in \mathcal{C}$, *then* $p(t) + q(t) \in \mathcal{C}$ *for* $t_1 \le t \le t_2$ *and*

$$\|\tilde{A}q(t)\| \le \|\tilde{A}q(s)\| e^{-\gamma_2(t-s)} \quad \text{for } t_1 \le s \le t \le t_2,$$

where γ_2 *is a constant independent of* $u_1(t)$, $u_2(t)$, t_1, *and* t_2.

(ii) *if* $p(t_1) + q(t_1) \notin \mathcal{C}$, *then*

$$p(t) + q(t) \notin \mathcal{C} \quad \text{for } t_1 \le t \le t_2.$$

Proof. By (2.2), $p(t)$ and $q(t)$ satisfy

$$\begin{cases} p_t + Ap = f_1(p_1 + q_1) - f_1(p_2 + q_2), \\ q_t + \tilde{A}q = f_2(p_1 + q_1) - f_2(p_2 + q_2), \\ \dfrac{\partial q}{\partial \nu} = \epsilon g(p_1 + q_1) - \epsilon g(p_2 + q_2). \end{cases} \tag{3.2}$$

Operate $\tilde{A}$ on the above equations. Taking an inner product between them and Ap, $\tilde{A}q$, respectively, we get

$$\frac{1}{2}\frac{d}{dt}\|Ap\|^2 \ge -\|A^{\frac{3}{2}}p\|^2 - \|A(f_1(p_1 + q_1) - f_1(p_2 + q_2))\| \, \|Ap\|, \tag{3.3}$$

$$\frac{1}{2}\frac{d}{dt}\|\tilde{A}q\|^2 \le -(\tilde{A}^2 q, \tilde{A}q) + \|\tilde{A}(f_2(p_1 + q_1) - f_2(p_2 + q_2))\| \, \|\tilde{A}q\|. \tag{3.4}$$

Using Lemma 2.1 (iii), we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\tilde{A}q\|^2 &\le -\|\tilde{A}q\|_1^2 + \int_{\partial\Omega} \tilde{A}qD\frac{\partial}{\partial\nu}\tilde{A}q dS + \|\tilde{A}(f_2(p_1 + q_1) - f_2(p_2 + q_2))\| \, \|\tilde{A}q\| \\ &\le -\|\tilde{A}q\|_1^2 + \int_{\partial\Omega} \tilde{A}qD\frac{\partial}{\partial\nu}(-q_t + f_2(p_1 + q_1) - f_2(p_2 + q_2))dS \\ &\quad + \|\tilde{A}(f_2(p_1 + q_1) - f_2(p_2 + q_2))\| \, \|\tilde{A}q\|. \end{aligned} \tag{3.5}$$

To estimate the right-hand side, we have to prepare the next lemma.

LEMMA 3.2. *Suppose that*

$$u = p + q \in \operatorname{conv}\left\{u \in H^4;\ \frac{\partial u}{\partial \nu} = \epsilon g(u) \text{ on } \partial\Omega, \|\tilde{A}^2 q\| \le 2R_3\right\},$$

$$\rho + \sigma, \tilde{\rho} + \tilde{\sigma} \in \left\{\rho + \sigma \in H^4;\ \frac{\partial \sigma}{\partial \nu} = \epsilon\frac{\partial g}{\partial u}(v)(\rho + \sigma) \text{ on } \partial\Omega \text{ for some } v \in H^2\right\},$$

and define

$$h(u) = \frac{\partial g}{\partial u}(u)(-\tilde{A}u + f_1(u) + f_2(u)).$$

Then

(i) $\|A\frac{\partial}{\partial u}f_1(u)(\rho + \sigma)\| \le (K_1 + \epsilon C_N)(\|A\rho\| + \|\tilde{A}\sigma\|)$,

(ii) $\|\tilde{A}\frac{\partial}{\partial u}f_2(u)(\rho + \sigma)\| \le (K_2 + \epsilon C_N)(\|A\rho\| + \|\tilde{A}\sigma\|)$,

(iii) $\|\frac{\partial}{\partial \nu}\frac{\partial}{\partial u}f_2(u)(\rho + \sigma)\|_{L^2(\partial\Omega)} \le \epsilon C_N(\|A\rho\| + \|\tilde{A}\sigma\|)$,

(iv) $\|\frac{\partial}{\partial u}h(u)(\rho + \sigma)\|_{L^2(\partial\Omega)} \le C_N(\|A\rho\| + \|\tilde{A}\sigma\|_1)$,

where

$$K_1 = K_2 = C|\tilde{F}|_3(1+R_1)^4.$$

We give the proof in the Appendix.

We arrange (3.3) and (3.5) by using Lemma 3.2(i)–(iv) and get

$$\text{(3.6)}\quad \begin{cases} \dfrac{1}{2}\dfrac{d}{dt}\|Ap\|^2 \geq -\lambda_N\|Ap\|^2 - (K_1+\epsilon C_N)(\|Ap\|+\|\tilde{A}q\|)\|Ap\|, \\ \dfrac{1}{2}\dfrac{d}{dt}\|\tilde{A}q\|^2 \leq -\|\tilde{A}q\|_1^2 + \{(K_2+\epsilon C_N)\|\tilde{A}q\| + \epsilon C_N\|\tilde{A}q\|_1\}(\|Ap\|+\|\tilde{A}q\|) \\ \qquad\qquad + \epsilon C_N(\|Ap\|+\|\tilde{A}q\|_1)\|\tilde{A}q\|_1; \end{cases}$$

note that

$$\frac{\partial}{\partial\nu}q_t = \epsilon(h(u_1)-h(u_2)).$$

It follows from (3.6) that for sufficiently small ϵ and $(p,q), p+q \in \mathcal{C}$,

$$\text{(3.7)}\quad \begin{cases} \dfrac{1}{2}\dfrac{d}{dt}\|Ap\|^2 \geq -\gamma_1\|\tilde{A}q\|^2, \\ \dfrac{1}{2}\dfrac{d}{dt}\|\tilde{A}q\|^2 \leq -\gamma_2\|\tilde{A}q\|^2, \end{cases}$$

where we put

$$\gamma_1 = \lambda_N + 2(K_1+\epsilon C_N), \qquad \gamma_2 = \lambda_{N+1} - 2(K_2 + \epsilon C_N(1+\lambda_{N+1}^{\frac{1}{2}})^2).$$

If we take the constant C_F in (C.2) satisfying

$$C_F > 2K_1 + 2K_2,$$

then the difference $\gamma_2 - \gamma_1$ is strictly positive for sufficiently small ϵ and

$$\text{(3.8)}\quad \frac{1}{2}\frac{d}{dt}(\|\tilde{A}q\|^2 - \|Ap\|^2) \leq -(\gamma_2-\gamma_1)\|\tilde{A}q\|^2 \leq 0,$$

that is, $\mathcal{C}$ is negatively invariant under the semiflow (see Fig. 1). By considering this fact and (3.7), we can verify that the statement of the proposition is true. □

3.2. Construction of an inertial manifold. We construct an inertial manifold by the graph transformation method due to Hadamard, which has been used by Mallet-Paret and Sell [18] for the case of the homogeneous boundary conditions (see also [22]). First we introduce the following set:

$$\mathcal{M}_{t,\epsilon} = S(t)PH^2.$$

Since $\mathcal{B}$ is invariant under the semiflow and

$$f_1(p) = 0,\ f_2(p) = 0 \quad \text{if} \quad \|Ap\| \geq 2R,$$

we have

$$\mathcal{M}_{t,\epsilon} \subset \mathcal{B} \cup \{p+q \in H^2;\ \|Ap\| \geq 2R, q = 0\} \subset \tilde{\mathcal{B}}.$$

This implies that $u_1(t), u_2(t) \in \mathcal{M}_{t,\epsilon}$ satisfy the assumption of Proposition 3.1, hence, that the difference of the above solutions satisfies the cone property. We investigate

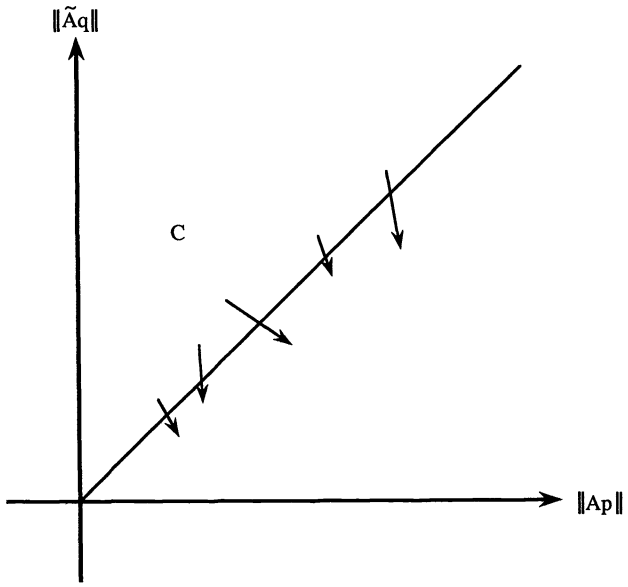


FIG. 1. *The negative invariance of C.*

a relation between $\|\tilde{A}\ \cdot\|$ and the H^2-norm. By using the mean value theorem, for $u = u_1(t) - u_2(t)$ we have

$$\begin{aligned}
\|\tilde{A}u\|^2 &= \int_\Omega |-D\Delta u + u|^2 dx \\
&\geq \|D\Delta u\|^2 + \|D^{\frac{1}{2}}\nabla u\|^2 + \|u\|^2 - 2\left\|D\frac{\partial}{\partial \nu}u\right\|_{L^2(\partial\Omega)}\|u\|_{L^2(\partial\Omega)} \\
&\geq \|D\Delta u\|^2 + \|D^{\frac{1}{2}}\nabla u\|^2 + \|u\|^2 - 2\|\epsilon D(g(u_1) - g(u_2))\|_{L^2(\partial\Omega)}\|u\|_{L^2(\partial\Omega)} \\
&\geq \|D\Delta u\|^2 + \|D^{\frac{1}{2}}\nabla u\|^2 + \|u\|^2 - \epsilon C\|u\|^2_{L^2(\partial\Omega)} \\
&\geq \|D\Delta u\|^2 + \|D^{\frac{1}{2}}\nabla u\|^2 + \|u\|^2 - \epsilon C\|u\|^2_{H^1} \\
&\geq \|D\Delta u\|^2 + \frac{1}{2}\|D^{\frac{1}{2}}\nabla u\|^2 + \frac{1}{2}\|u\|^2 \\
&\geq C\|u\|^2_{H^2}.
\end{aligned}$$

By this we can apply the theory of the inertial manifold by Mallet-Paret and Sell to our case under the space H^2. Hence we have the following proposition.

PROPOSITION 3.3. *There exists a continuous mapping $\Phi_{t,\epsilon}$ from PH^2 into QH^2 satisfying*

(i) $\mathcal{M}_{t,\epsilon} = \text{graph } \Phi_{t,\epsilon}$,

(ii) $\{\Phi_{t,\epsilon}\}_{t\geq 0}$ *is a Cauchy sequence in H^2, and the limit of $\Phi_{t,\epsilon}$, denoted by Φ_ϵ, is Lipschitz,*

(iii) $\mathcal{M}_\epsilon = \text{graph } \Phi_\epsilon$ *is an invariant manifold,*

(iv) *for any solutions $u(t)$ of (2.2), there exists $p_0 \in PH^2$ such that*

$$\|u(t) - v(t)\|_{H^2} \leq Ce^{-\gamma_2 t}\|u(0) - v(0)\|_{H^2} \qquad (t \geq 0),$$

where $v(t) = p(t;p_0) + q(t;p_0)$ *is a solution of*

$$\begin{cases} p_t + Ap = f_1(p + \Phi_\epsilon(p)), \\ q_t + \tilde{A}q = f_2(p + \Phi_\epsilon(p)), \\ \dfrac{\partial q}{\partial \nu} = \epsilon g(p + \Phi_\epsilon(p)), \end{cases} \tag{3.9}$$

with the initial data $v(0) = p_0$.

Remark. By this proposition, $\mathcal{M}_\epsilon$ is a Lipschitz inertial manifold in our sense. We note that $q(t;p_0)$ is expressed as

$$q(t;p_0) = \Phi_\epsilon(p(t;p_0)).$$

Then

$$q(0;p_0) = \Phi(p_0),$$

if we put $t = 0$. This implies that the inertial manifold is characterized by a set of initial data for which the solution can be extended for all negative time and bounded in $t \in (-\infty, 0]$.

3.3. C^1-smoothness of an inertial manifold. We consider the variation of (2.2) around the solution $(p(t;p_0), q(t;p_0))$ satisfying $p(0;p_0) = p_0$ and $q(0;p_0) = \Phi(p_0)$:

$$\begin{cases} \rho_t + A\rho = \dfrac{\partial}{\partial u} f_1(p(t;p_0) + q(t;p_0))(\rho + \sigma), \\ \sigma_t + \tilde{A}\sigma = \dfrac{\partial}{\partial u} f_2(p(t;p_0) + q(t;p_0))(\rho + \sigma), \\ \dfrac{\partial \sigma}{\partial \nu} = \epsilon \dfrac{\partial}{\partial u} g(p(t;p_0) + q(t;p_0))(\rho + \sigma). \end{cases} \tag{3.10}$$

The solution of the equation (3.10) with the initial conditions

$$\rho(0) = \xi, \quad \sigma(s) = 0 \quad (s \le t \le 0)$$

is denoted by

$$\rho^s(t;p_0,\xi) + \sigma^s(t;p_0,\xi).$$

We will prove that there exist limits of ρ^s, σ^s as $s \to -\infty$ and that the limit

$$\lim_{s \to -\infty} \sigma^s(0;p_0,\xi)$$

coincides with

$$\frac{\partial}{\partial p}\Phi_\epsilon(p_0)\xi.$$

First we present the cone property for the solution of (3.10).

LEMMA 3.4. *Let* $\mathcal{C}$ *and* $\tilde{\mathcal{B}}$ *be the sets defined by* (3.1). *If* $p(t,p_0) + q(t,p_0) \in \tilde{\mathcal{B}}$ *for* $t_1 \le t \le t_2$, *the following holds:*

(i) *if* $\rho(t_2) + \sigma(t_2) \in \mathcal{C}$, *then* $\rho(t) + \sigma(t) \in \mathcal{C}$ *for* $t_1 \le t \le t_2$ *and*

$$(\|A\rho(t)\| \le)\|\tilde{A}\sigma(t)\| \le \|\tilde{A}\sigma(s)\|e^{-\gamma_2(t-s)} \quad \text{for } t_1 \le s \le t \le t_2,$$

(ii) *if* $\rho(t_1) + \sigma(t_1) \notin \mathcal{C}$, *then* $\sigma(t) + \sigma(t) \notin \mathcal{C}$ *for* $t_1 \le t \le t_2$ *and*

$$(\|\tilde{A}\sigma(t)\| \le)\|A\rho(t)\| \le \|A\rho(s)\|e^{-\gamma_1(t-s)} \quad \text{for } t_1 \le t \le s \le t_2.$$

This lemma can be proved in the same manner as that of Proposition 3.1 except for the last inequality. The last one is shown by using a similar inequality to (3.6). We omit the detail to avoid repeating the same computation.

Applying this lemma to the solution $\rho^s + \sigma^s$ yields

$$\|\tilde{A}\sigma^s(t;p_0,\xi)\| \le \|A\rho^s(t;p_0,\xi)\| \le \|A\xi\|e^{-\gamma_1 t} \qquad (s \le t \le 0). \tag{3.11}$$

Using Lemma 3.4 and (3.11), we can show that ρ^s, σ^s make Cauchy sequences as $s \to -\infty$. Indeed, for $s \le \tau \le t \le 0$, put

$$\tilde{\rho} = \rho^s(t;p_0,\xi) - \rho^\tau(t;p_0,\xi), \qquad \tilde{\sigma} = \sigma^s(t;p_0,\xi) - \sigma^\tau(t;p_0,\xi),$$

which also satisfy (3.10). From Lemma 3.4 (i) and the inequalities (3.11), we have

$$\begin{aligned}\|A\tilde{\rho}(t;p_0,\xi)\| \le \|\tilde{A}\tilde{\sigma}(t;p_0,\xi)\| &\le \|\tilde{A}\tilde{\sigma}(\tau;\xi)\|e^{-\gamma_2(t-\tau)} \\ &\le \|\tilde{A}\sigma^s(\tau;\xi)\|e^{-\gamma_2(t-\tau)} \\ &\le \|A\xi\|e^{-\gamma_2 t+(\gamma_2-\gamma_1)\tau} \to 0 \quad \text{as } s,\tau \to -\infty.\end{aligned}$$

Hence there exist limits of $\rho^s(t;p_0,\xi), \sigma^s(t;p_0,\xi)$ as $s \to -\infty$. We denote the limits by $\rho(t;p_0,\xi), \sigma(t;p_0,\xi)$, respectively.

PROPOSITION 3.5. *H^2-valued functions $p(t;p_0)$ and $q(t;p_0)$ are continuously differentiable with respect to $p_0 \in PH^2$ and*

$$\frac{\partial}{\partial p}p(t;p_0)\xi = \rho(t;p_0,\xi), \qquad \frac{\partial}{\partial p}q(t;p_0)\xi = \sigma(t;p_0,\xi).$$

In particular, when $t = 0$,

$$\frac{\partial}{\partial p}\Phi_\epsilon(p_0)\xi = \sigma(0;p_0,\xi).$$

Proof. Put

$$\begin{aligned}\tilde{p}(t) &= p(t,p_0+\xi) - p(t,p_0) - \rho(t;p_0,\xi), \\ \tilde{q}(t) &= q(t,p_0+\xi) - q(t,p_0) - \sigma(t;p_0,\xi).\end{aligned}$$

We simply write

$$p^\xi = p(t;p_0+\xi), \quad q^\xi = q(t;p_0+\xi), \quad p = p(t;p_0), \quad q = q(t;p_0), \quad u = p+q,$$

as long as there is no confusion. We prove that

$$\begin{cases} \|A\tilde{p}(t)\| \le C_N\|A\xi\|^{1+\delta}e^{-2\gamma t}, \\ \|A\tilde{q}(t)\| \le C_N\|A\xi\|^{1+\delta}e^{-2\gamma t}, \end{cases}$$

where δ is some constant specified later. If the above inequalities are shown, this proposition immediately follows from the definitions of $\tilde{p}$ and $\tilde{q}$. We divide the proof into three parts. In the first step, we introduce auxiliary functions $X(t), Y(t)$ and derive the differential inequalities for them. Next we investigate the behaviors of $X(t), Y(t)$ as $t \to -\infty$ and finally the desired estimates of $\|A\tilde{p}\|, \|\tilde{A}\tilde{q}\|$ will be shown through such behaviors of $X(t), Y(t)$.

Step 1. The above $\tilde{p}, \tilde{q}$ satisfy

$$\begin{cases} \tilde{p}_t + A\tilde{p} = w_1, \\ \tilde{q}_t + \tilde{A}\tilde{q} = w_2, \\ \dfrac{\partial\tilde{q}}{\partial\nu} = w_3, \end{cases} \tag{3.12}$$

where

$$\begin{cases} w_1 = f_1(p^\xi + q^\xi) - f_1(p+q) - \dfrac{\partial}{\partial u} f_1(p+q)(\rho+\sigma), \\ w_2 = f_2(p^\xi + q^\xi) - f_2(p+q) - \dfrac{\partial}{\partial u} f_2(p+q)(\rho+\sigma), \\ w_3 = \epsilon \left(g(p^\xi + q^\xi) - g(p+q) - \dfrac{\partial}{\partial u} g(p+q)(\rho+\sigma) \right). \end{cases} \tag{3.13}$$

Take the inner product between (3.12) and $A\tilde{p}, \tilde{A}\tilde{q}$. Then we see

$$\begin{cases} \dfrac{1}{2}\dfrac{d}{dt}\|A\tilde{p}\|^2 \geq -\|A^{\frac{3}{2}}\tilde{p}\|^2 - \|Aw_1\| \, \|A\tilde{p}\|, \\ \dfrac{1}{2}\dfrac{d}{dt}\|\tilde{A}\tilde{q}\|^2 \leq -(\tilde{A}^2\tilde{q}, \tilde{A}\tilde{q}) + \|\tilde{A}w_2\| \, \|\tilde{A}\tilde{q}\|. \end{cases} \tag{3.14}$$

First estimate the term $(\tilde{A}^2\tilde{q}, \tilde{A}\tilde{q})$ as follows:

$$\begin{aligned} (\tilde{A}^2\tilde{q}, \tilde{A}\tilde{q}) &= \|\tilde{A}\tilde{q}\|_1^2 - \int_{\partial\Omega} \tilde{A}\tilde{q} D\frac{\partial}{\partial\nu}\tilde{A}\tilde{q} dS \\ &\geq \|\tilde{A}\tilde{q}\|_1^2 - C\|\tilde{A}\tilde{q}\|_{L^2(\partial\Omega)} \left\| \frac{\partial}{\partial\nu}(-\tilde{q}_t + w_2) \right\|_{L^2(\partial\Omega)} \\ &\geq \|\tilde{A}\tilde{q}\|_1^2 - C\|\tilde{A}\tilde{q}\|_1 \left(\|w_{3_t}\|_{L^2(\partial\Omega)} + \left\| \frac{\partial}{\partial\nu} w_2 \right\|_{L^2(\partial\Omega)} \right). \end{aligned} \tag{3.15}$$

For the desired estimates of $\|A\tilde{p}\|, \|\tilde{A}\tilde{q}\|$, we prepare the next lemma.

LEMMA 3.6. *There exist positive functions $a(t;\xi), b(t;\xi)$ and a constant γ ($\gamma_1 < \gamma < \frac{\gamma_1+\gamma_2}{2}$) such that*

$$\begin{cases} \|Aw_1\| \leq (K_1 + \epsilon C_N)(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|) + a(t;\xi), \\ \|\tilde{A}w_2\| \leq (K_2 + \epsilon C_N)(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|) + a(t;\xi), \\ \|w_{3_t}\|_{L^2(\partial\Omega)} \leq \epsilon C_N(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|_1 + b(t;\xi)), \\ \left\| \dfrac{\partial}{\partial\nu} w_2 \right\|_{L^2(\partial\Omega)} \leq \epsilon C_N(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\| + a(t;\xi)) \end{cases} \tag{3.16}$$

and

$$\begin{cases} a(t;\xi) \leq C_N \|A\xi\|^{2(1+\delta)} e^{-(1+\delta)\gamma_1 t}, \\ \displaystyle\int_{-\infty}^{t} b(s;\xi)^2 e^{2\gamma s} ds \leq C_N \|A\xi\|^{2(1+\delta)} e^{2(\gamma-(1+\delta)\gamma_1)t} \qquad (t \leq 0), \end{cases} \tag{3.17}$$

where $\delta \leq \frac{\gamma-\gamma_1}{2\gamma}$.

We will give the proof of this lemma in the last part of this subsection. Applying Lemma 3.6 to (3.14) and (3.15), we obtain

$$\begin{cases} \dfrac{1}{2}\dfrac{d}{dt}\|A\tilde{p}\|^2 \geq -\lambda_N\|A\tilde{p}\|^2 - (K_1 + \epsilon C_N)(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|)\|A\tilde{p}\| - a(t;\xi)\|A\tilde{p}\|, \\ \dfrac{1}{2}\dfrac{d}{dt}\|\tilde{A}\tilde{q}\|^2 \leq -\|\tilde{A}\tilde{q}\|_1^2 + (K_2 + \epsilon C_N)(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|)\|\tilde{A}\tilde{q}\| \\ \qquad\qquad + 2\epsilon C_N(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|)\|\tilde{A}\tilde{q}\|_1 + (a(t;\xi) + b(t;\xi))(\|\tilde{A}\tilde{q}\| + 2\epsilon C\|\tilde{A}\tilde{q}\|_1). \end{cases} \tag{3.18}$$

Because of the presence of the term that includes $b(t;\xi)$, it is difficult to handle (3.18) in this form. We introduce the following functions and rewrite (3.18) with the new functions:

$$\begin{cases} X(t)^2 = \displaystyle\int_{-\infty}^{t} e^{2\gamma s}\|A\tilde{p}(s)\|^2 ds, \\ Y(t)^2 = \displaystyle\int_{-\infty}^{t} e^{2\gamma s}\|\tilde{A}\tilde{q}(s)\|^2 ds, \\ Y_1(t)^2 = \displaystyle\int_{-\infty}^{t} e^{2\gamma s}\|\tilde{A}\tilde{q}(s)\|_1^2 ds. \end{cases}$$

We first prove the boundedness of $X(t), Y(t)$ for $t \le 0$:

$$\begin{aligned} Y(t) \le & 2\int_{-\infty}^{t} e^{2\gamma t}(\|\tilde{A}(q^{\xi}(s) - q(s))\|^2 + \|\tilde{A}\sigma(s)\|^2) ds \\ \le & 4\int_{-\infty}^{t} e^{2\gamma s}\|A\xi\|^2 e^{-2\gamma_1 s} ds \\ = & \frac{2}{\gamma - \gamma_1} e^{2(\gamma-\gamma_1)t} \le \frac{2}{\gamma - \gamma_1}; \end{aligned}$$

the boundedness of $X(t)$ is also verified in a similar manner. Multiplying $e^{2\gamma t}$ by (3.18) and integrating over $(-\infty, t]$ yields

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}X^2 \ge & -(\lambda_N - \gamma)X^2 - (K_1 + \epsilon C_N)(X+Y)X \\ & - \left(\int_{-\infty}^{t} a(s;\xi)^2 e^{2\gamma s} ds\right)^{\frac{1}{2}} X, \\ \frac{1}{2}\frac{d}{dt}Y^2 \le & -Y_1^2 + \gamma Y^2 + (K_2 + \epsilon C_N)(X+Y)Y + \epsilon C_N (X+Y)Y_1 \\ & + \left\{\left(\int_{-\infty}^{t} a(s;\xi)^2 e^{2\gamma s} ds\right)^{\frac{1}{2}} + \left(\int_{-\infty}^{t} b(s;\xi)^2 e^{2\gamma s} ds\right)^{\frac{1}{2}}\right\}(Y + 2\epsilon C_N Y_1). \end{aligned}$$

Using (3.17), we have

$$\text{(3.19)} \qquad \begin{cases} \dfrac{1}{2}\dfrac{d}{dt}X^2 \ge -(\lambda_N - \gamma)X^2 - (K_1 + \epsilon C_N)(X+Y)X - C_N\|A\xi\|^{1+\delta}X, \\ \dfrac{1}{2}\dfrac{d}{dt}Y^2 \le -Y_1^2 + \gamma Y^2 + (K_2 + \epsilon C_N)(X+Y)Y \\ \qquad\qquad + 2\epsilon C_N (X+Y)Y_1 + C_N\|A\xi\|^{1+\delta}(Y + \epsilon C_N Y_1). \end{cases}$$

Define the set

$$\text{(3.20)} \qquad \mathcal{S}(K) = \{(X,Y) \in \mathbb{R}^2; Y \ge X \ge 0, Y \ge K\},$$

where

$$K = \frac{2C_N\|A\xi\|^{1+\delta}}{\gamma - \gamma_1}.$$

We prove that $\mathcal{S}(K)$ is a negatively invariant region of the XY space and that every bounded solution in t, $-\infty < t \le 0$, is squeezed from $\mathcal{S}(K)$. We call it the *squeezing*

property. If $(X(t),Y(t)) \in \mathcal{S}(K)$, by the second inequality of (3.19) we get

$$\frac{1}{2}\frac{d}{dt}Y^2 \le -Y_1^2+\gamma Y^2+(K_2+\epsilon C_N)(X+Y)Y+2\epsilon C_N(X+Y)Y_1+\frac{\gamma-\gamma_1}{2}K(Y+2\epsilon C_N Y_1).$$

Since $Y_1(t) \ge \lambda_{N+1}^{\frac{1}{2}} Y(t) \ge \lambda_{N+1}^{\frac{1}{2}} X(t)$, we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}Y^2 \le &-(\lambda_{N+1}-\gamma)Y^2+(K_2+\epsilon C_N)(X+Y)Y \\ &+2\epsilon C_N\lambda_{N+1}^{\frac{1}{2}}(X+Y)Y+\frac{\gamma-\gamma_1}{2}K(1+2\epsilon C\lambda_{N+1}^{\frac{1}{2}})Y.\end{aligned} \tag{3.21}$$

We prove that

$$(X,Y) \notin \mathcal{S}(K) \quad \text{for all } t \le 0.$$

Indeed, if $(X,Y) \in \mathcal{S}(K)$, then one has

$$\begin{cases}\frac{1}{2}\frac{d}{dt}(Y^2-X^2) \le -\frac{\gamma_2-\gamma_1}{2}Y^2+\epsilon C_N' Y^2+\gamma(Y^2-X^2), \\ \frac{1}{2}\frac{d}{dt}Y^2 \le -(\gamma_2-\gamma)Y^2+\frac{\gamma-\gamma_1}{2}Y^2+\epsilon C_N' Y^2.\end{cases} \tag{3.22}$$

Considering the direction of the flow on the boundary $\partial\mathcal{S}(K)$, one can say that $\mathcal{S}(K)$ is negatively invariant under the flow (see Fig. 2). If there exists a t such that $(X(t) < Y(t)) \in \mathcal{S}(K)$, then the second inequality of (3.22) implies that

$$Y(t)^2 \le Y(s)^2 e^{-\frac{\gamma_2-\gamma}{2}(t-s)} \quad \text{for } s \le t \le t_0 \tag{3.23}$$

(recall the definition of γ, γ_1, and γ_2, see §3.1). Because of the boundedness of $Y(s)$, letting $s \to -\infty$ yields $Y(t) = 0$ for all $t \le 0$. This is a contradiction. Thus we get

$$(X(t),Y(t)) \notin \mathcal{S}(K) \quad \text{for all } t \le 0.$$

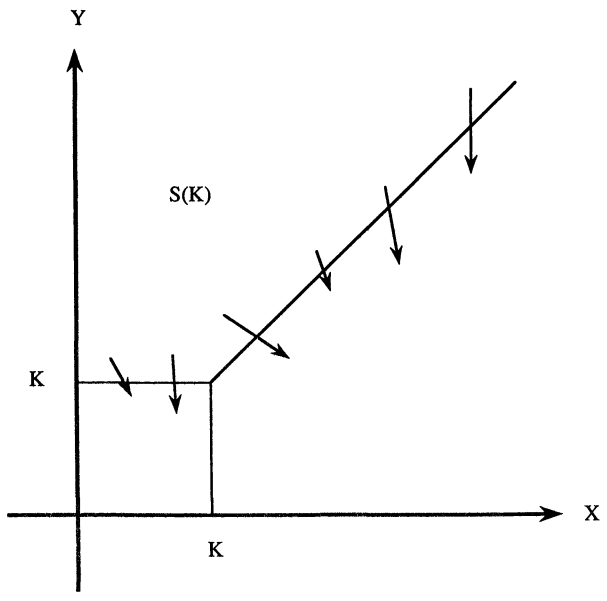


FIG. 2. *The negative invariance of* $\mathcal{S}(K)$.

Step 2. We prove that

$$0 \le X(t) \le K, \quad 0 \le Y(t) \le K \quad \text{for all } t \le 0. \tag{3.24}$$

If $X(0) > K$, then we have

$$\frac{1}{2}\frac{d}{dt}X^2\Big|_{t=0} \ge \frac{\gamma-\gamma_1}{2}X^2\Big|_{t=0} \ge \frac{\gamma-\gamma_1}{2}K^2 > 0. \tag{3.25}$$

On the other hand,

$$\frac{d}{dt}X(t)^2 = e^{2\gamma t}\|A\tilde{p}(t)\|^2,$$

which implies

$$\frac{d}{dt}X(t)^2\Big|_{t=0} = \|A\tilde{p}(0)\|^2 = 0.$$

This contradicts (3.25). Hence we have $X(0) \le K$. From this and

$$\frac{d}{dt}X(t)^2 \ge 0,$$

we see that the former inequalities of (3.24) hold. The latter inequalities follow from the former ones and $(X(t), Y(t)) \notin \mathcal{S}(K)$ $(t \le 0)$.

Step 3. By (3.19) and (3.24), we have

$$\begin{aligned}\frac{1}{2}e^{2\gamma t}\|\tilde{Aq}\|^2 \le& -Y_1^2 + \gamma Y^2 + (K_2 + \epsilon C_N)(X+Y)(Y+\epsilon C_N Y_1)\\ &+ 2\epsilon C_N(X+Y)Y_1 + \frac{\gamma-\gamma_1}{2}K(Y+\epsilon C_N Y_1)\\ \le& C_N K^2.\end{aligned}$$

From this,

$$na\tilde{q}(t)^2 \le C_N K^2 e^{-2\gamma t} \le \frac{2C_N^2}{\gamma-\gamma_1}\|A\xi\|^{2(1+\delta)}e^{-2\gamma t}. \tag{3.26}$$

We turn to the estimate of $\|A\tilde{p}\|$. Recall that

$$\frac{1}{2}\frac{d}{dt}\|A\tilde{p}\|^2 \ge -\lambda_N\|A\tilde{p}\|^2 - (K_1+\epsilon C_N)(\|A\tilde{p}\| + \|\tilde{Aq}\|)\|A\tilde{p}\| - a(t;\xi)\|A\tilde{p}\|.$$

Multiplying $e^{2\gamma t}$ and integrating over $[t, 0]$, we obtain

$$-\frac{1}{2}e^{2\gamma t}\|A\tilde{p}(t)\|^2 \ge -\gamma_1 X(0)^2 - (K_1+\epsilon C_N)(X(0)+Y(0))X(0) - \frac{\gamma-\gamma_1}{2}KX(0).$$

Use (3.24) again. Then we have

$$\|A\tilde{p}(t)\|^2 \le C_N\|A\xi\|^{2(1+\delta)}e^{-2\gamma t}.$$

This inequality and (3.26) are the desired ones. □

Next we give the proof of Lemma 3.6.

Proof of Lemma 3.6. By the mean value theorem, (3.13) can be written as

$$\begin{cases} w_1 = \displaystyle\int_0^1 \frac{\partial f_1}{\partial u}(u^\theta)(\tilde{p}+\rho+\tilde{q}+\sigma)d\theta - \frac{\partial f_1}{\partial u}(u)(\rho+\sigma), \\ w_2 = \displaystyle\int_0^1 \frac{\partial f_2}{\partial u}(u^\theta)(\tilde{p}+\rho+\tilde{q}+\sigma)d\theta - \frac{\partial f_2}{\partial u}(u)(\rho+\sigma), \\ w_3 = \epsilon\displaystyle\int_0^1 \frac{\partial g}{\partial u}(u^\theta)(\tilde{p}+\rho+\tilde{q}+\sigma)d\theta - \epsilon\frac{\partial g}{\partial u}(u)(\rho+\sigma), \end{cases}$$

where

$$u^\xi = p^\xi + q^\xi, \qquad u^\theta = \theta u^\xi + (1-\theta)u.$$

We define

$$\begin{cases} \Gamma_1(\xi) = \displaystyle\int_0^1 \left(\frac{\partial f_1}{\partial u}(u^\theta) - \frac{\partial f_1}{\partial u}(u)\right)(\rho+\sigma)d\theta, \\ \Gamma_2(\xi) = \displaystyle\int_0^1 \left(\frac{\partial f_2}{\partial u}(u^\theta) - \frac{\partial f_2}{\partial u}(u)\right)(\rho+\sigma)d\theta, \\ \Gamma_3(\xi) = \epsilon\displaystyle\int_0^1 \left(\frac{\partial g}{\partial u}(u^\theta) - \frac{\partial g}{\partial u}(u)\right)(\rho+\sigma)d\theta. \end{cases} \tag{3.27}$$

Lemma 3.2 tells us that

$$\begin{cases} \|Aw_1\| \le (K_1+\epsilon C_N)(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|) + \|A\Gamma_1\|, \\ \|\tilde{A}w_2\| \le (K_2+\epsilon C_N)(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|) + \|\tilde{A}\Gamma_2\|, \\ \|w_{3_t}\|_{L^(\partial\Omega)} \le \epsilon C_N(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|) + \|\Gamma_{3_t}\|_{L^2(\partial\Omega)}, \\ \left\|\dfrac{\partial}{\partial\nu}w_2\right\|_{L^2(\partial\Omega)} \le \epsilon C_N(\|A\tilde{p}\| + \|\tilde{A}\tilde{q}\|) + \left\|\dfrac{\partial}{\partial\nu}\Gamma_1\right\|_{L^2(\partial\Omega)}. \end{cases}$$

To obtain our desired estimates of (3.16), we need a little technical consideration. First prepare two different types of inequalities for Γ_i $(i=1,2,3)$. Applying Lemma 3.2 to (3.27) yields

$$\begin{cases} \|A\Gamma_1\| \le 2(K_1+\epsilon C_N)(\|A\rho\| + \|\tilde{A}\sigma\|) \le 4(K_1+\epsilon C_N)\|A\xi\|e^{-\gamma_1 t}, \\ \|\tilde{A}\Gamma_2\| \le 2(K_2+\epsilon C_N)(\|A\rho\| + \|\tilde{A}\sigma\|) \le 4(K_2+\epsilon C_N)\|A\xi\|e^{-\gamma_1 t}, \\ \|\Gamma_{3_t}\|_{L^2(\partial\Omega)} \le 2\epsilon C_N(\|A\rho\| + \|\tilde{A}\sigma\|_1) \le 2\epsilon C_N(\|A\xi\|e^{-\gamma_1 t} + \|\tilde{A}\sigma\|_1), \\ \left\|\dfrac{\partial}{\partial\nu}\Gamma_2\right\|_{L^2(\partial\Omega)} \le 2\epsilon C_N(\|A\rho\| + \|\tilde{A}\sigma\|) \le 4\epsilon C_N\|A\xi\|e^{-\gamma_1 t}. \end{cases} \tag{3.28}$$

Next we estimate $\Gamma_1, \Gamma_2, \Gamma_{3_t}$, and $\partial\Gamma_2/\partial\nu$ by higher-order terms of ξ. To show it, we introduce the following lemma.

LEMMA 3.7. *Under the assumption of Lemma* 3.2,

(i) $\|A(\frac{\partial}{\partial u})^2 f_1(u)(\rho+\sigma)(\tilde{\rho}+\tilde{\sigma})\| \le (C+\epsilon C_N)(\|A\rho\| + \|\tilde{A}\sigma\|)(\|A\tilde{\rho}\| + \|\tilde{A}\tilde{\sigma}\|)$,

(ii) $\|\tilde{A}(\frac{\partial}{\partial u})^2 f_2(u)(\rho+\sigma)(\tilde{\rho}+\tilde{\sigma})\| \le (C+\epsilon C_N)(\|A\rho\| + \|\tilde{A}\sigma\|)(\|A\tilde{\rho}\| + \|\tilde{A}\tilde{\sigma}\|)$,

(iii) $\|\frac{\partial}{\partial\nu}(\frac{\partial}{\partial u})^2 f_2(u)(\rho+\sigma)(\tilde{\rho}+\tilde{\sigma})\|_{L^2(\partial\Omega)} \le \epsilon C_N(\|A\rho\| + \|\tilde{A}\sigma\|)(\|A\tilde{\rho}\| + \|\tilde{A}\tilde{\sigma}\|)$,

(iv) $\|(\frac{\partial}{\partial u})^2 h(u)(\rho+\sigma)(\tilde{\rho}+\tilde{\sigma})\|_{L^2(\partial\Omega)}$
$\le C_N\{(\|A\rho\| + \|\tilde{A}\sigma\|_1)(\|A\tilde{\rho}\| + \|\tilde{A}\tilde{\sigma}\|) + \|\tilde{A}\tilde{\sigma}\|\|\tilde{A}\sigma\|_1\}$.

The proof of this lemma will be carried out in the Appendix.

Apply Lemma 3.7(i) to (3.27). Then

$$\|A\Gamma_1\| \le (C+\epsilon C_N)(\|A(p^\xi - p)\| + \|\tilde{A}(q^\xi - q)\|)(\|A\rho\| + \|\tilde{A}\sigma\|).$$

Similarly, we can get the estimates of $\|\tilde{A}\Gamma_2\|$, $\|\frac{\partial}{\partial\nu}\Gamma_2\|_{L^2(\partial\Omega)}$, and $\|\Gamma_{3_t}\|_{L^2(\partial\Omega)}$ by using Lemma 3.7 (ii), (iii), and (iv) respectively. Hence we obtain the following:

$$\begin{cases} \|A\Gamma_1\| \le (C+\epsilon C_N)(\|A(p^\xi - p)\| + \|\tilde{A}(q^\xi - q)\|)(\|A\rho\| + \|\tilde{A}\sigma\|), \\ \|\tilde{A}\Gamma_2\| \le (C+\epsilon C_N)(\|A(p^\xi - p)\| + \|\tilde{A}(q^\xi - q)\|)(\|A\rho\| + \|\tilde{A}\sigma\|), \\ \|\Gamma_{3_t}\|_{L^2(\partial\Omega)} \le \epsilon C_N(\|A(p^\xi - p)\| + \|\tilde{A}(q^\xi - q)\|_1)(\|A\rho\| + \|\tilde{A}\sigma\|) \\ \qquad\qquad + \epsilon C_N\|\tilde{A}(q^\xi - q)\|\|\tilde{A}\sigma\|_1, \\ \left\|\frac{\partial}{\partial\nu}\Gamma_2\right\|_{L^2(\partial\Omega)} \le \epsilon C_N(\|A(p^\xi - p)\| + \|\tilde{A}(q^\xi - q)\|)(\|A\rho\| + \|\tilde{A}\sigma\|). \end{cases}$$

Using the estimates of $\|A\rho\|, \|\tilde{A}\sigma\|, \|A(p^\xi - p)\|$, and $\|\tilde{A}(q^\xi - q)\|$ as in Lemma 3.4, we have

(3.29)
$$\begin{cases} \|A\Gamma_1\| \le 4(C+\epsilon C_N)\|A\xi\|^2 e^{-2\gamma_1 t}, \\ \|\tilde{A}\Gamma_2\| \le 4(C+\epsilon C_N)\|A\xi\|^2 e^{-2\gamma_1 t}, \\ \|\Gamma_{3_t}\|_{L^2(\partial\Omega)} \le 2\epsilon C_N\|A\xi\|^2 e^{-2\gamma_1 t} + \epsilon C_N\|A\xi\|e^{-\gamma_1 t}(2\|\tilde{A}(q^\xi - q)\|_1 + \|\tilde{A}\sigma\|_1), \\ \left\|\frac{\partial}{\partial\nu}\Gamma_2\right\|_{L^2(\partial\Omega)} \le 4\epsilon C_N\|A\xi\|^2 e^{-2\gamma_1 t}. \end{cases}$$

Next we prove

$$\int_{-\infty}^{t} \|\tilde{A}\sigma\|_1^2 e^{2\gamma s} ds \le C_N\|A\xi\|^2 e^{2(\gamma-\gamma_1)t}. \tag{3.30}$$

The argument used in getting (3.6) leads us to

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\tilde{A}\sigma\|^2 + \|\tilde{A}\sigma\|_1^2 \le & 2\{(K_2+\epsilon C_N)\|\tilde{A}\sigma\| + \epsilon C_N\|\tilde{A}\sigma\|_1\}(\|A\rho\| + \|\tilde{A}\sigma\|) \\ & + \epsilon C_N\|\tilde{A}\sigma\|_1(\|A\rho\| + \|\tilde{A}\sigma\|_1), \end{aligned}$$

hence

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\tilde{A}\sigma\|^2 + \left(\frac{\lambda_{N+1}-\gamma_2}{\lambda_{N+1}} + \frac{\gamma_2}{\lambda_{N+1}}\right)(1-\epsilon C_N)\|\tilde{A}\sigma\|_1^2 & \\ \le 2K_2\|\tilde{A}\sigma\|(\|A\rho\| + \|\tilde{A}\sigma\|) + \epsilon C_N(\|A\rho\| + \|\tilde{A}\sigma\|)^2 & \\ \le 4(K_2+\epsilon C_N)\|A\xi\|^2 e^{-2\gamma_1 t}. & \end{aligned}$$

Since $\gamma < (1-\epsilon C_N)\gamma_2$ for sufficiently small ϵ, we have

$$\frac{1}{2}\frac{d}{dt}\left(e^{2\gamma t}\|\tilde{A}\sigma\|^2\right) + (1-\epsilon C_N)\frac{\lambda_{N+1}-\gamma_2}{\lambda_{N+1}}\|\tilde{A}\sigma\|_1^2 e^{2\gamma t} \le 4(C+\epsilon C_N)\|A\xi\|^2 e^{2(\gamma-\gamma_1)t}.$$

Integrate this inequality over $(-\infty, t]$. Then we obtain (3.30). Similarly, we get

$$\int_{-\infty}^{t} \|\tilde{A}(q^\xi - q)\|_1^2 e^{2\gamma s} ds \le C_N\|A\xi\|^2 e^{2(\gamma-\gamma_1)t}. \tag{3.31}$$

We note that (3.30) and (3.31) hold for an arbitrary constant γ satisfying $\gamma_1 - \epsilon C_N \leq \gamma \leq \frac{\gamma_1+\gamma_2}{2} + \epsilon C_N$.

Putting

$$\text{(3.32)} \qquad \begin{cases} a_1(t;\xi) = \|A\xi\|e^{-\gamma_1 t}, \\ b_1(t;\xi) = \|\tilde{A}(q^\xi - q)\|_1 + \|\tilde{A}\sigma\|_1, \\ a(t;\xi) = (C + \epsilon C_N)a_1(t;\xi), \\ b(t;\xi) = \min(2a_1(t;\xi) + 2b_1(t;\xi), 2a_1(t;\xi)^2 + 2a_1(t;\xi)b_1(t;\xi)), \end{cases}$$

and substituting (3.32) into (3.28) and (3.29) we obtain

$$\begin{cases} \|A\Gamma_1\| \leq a(t;\xi), \\ \|\tilde{A}\Gamma_2\| \leq a(t;\xi), \\ \|\Gamma_{3_t}\|_{L^2(\partial\Omega)} \leq b(t;\xi), \\ \left\| \dfrac{\partial}{\partial\nu}\Gamma_2 \right\|_{L^2(\partial\Omega)} \leq a(t;\xi). \end{cases}$$

Finally we prove (3.17). A simple calculation yields

$$\begin{aligned} b(t;\xi)^2 \leq & \{2(a_1(t;\xi) + b_1(t;\xi))\}^{2-2\delta}\{2a_1(t;\xi)(a_1(t;\xi) + b_1(t;\xi))\}^{2\delta} \\ \leq & C\{a_1(t;\xi)^{2+2\delta} + a_1(t;\xi)^{2\delta}b_1(t;\xi)^2\}. \end{aligned}$$

Hence we obtain

$$\begin{aligned} & \int_{-\infty}^t b(s;\xi)^2 e^{2\gamma t} ds \\ & \leq C_N\|A\xi\|^{2+2\delta}\int_\infty^t e^{2\gamma s - 2(1+\delta)\gamma_1 s} ds + C_N\|A\xi\|^2\int_{-\infty}^t e^{-2\gamma_1\delta s}(\|\tilde{A}(q^\xi - q)\|_1^2 + \|\tilde{A}\sigma\|_1^2) ds \\ & \leq C_N\|A\xi\|^{2+2\delta}e^{2(\gamma-(1+\delta)\gamma_1)t} + C_N\|A\xi\|^2\int_{-\infty}^t \|\tilde{A}\sigma_1\|_1^2 e^{2(\gamma-\gamma_1\delta)s} ds \\ & \leq C_N\|A\xi\|^{2+2\delta}e^{2(\gamma-(1+\delta)\gamma_1)t} + C_N\|A\xi\|^{2+2\delta}e^{2(\gamma-(1+\delta)\gamma_1)t}, \end{aligned}$$

where we used (3.30) and (3.31), but replaced γ by $\gamma - \gamma_1\delta$ to get the last one. This concludes the proof of Lemma 3.6. □

3.4. Convergence. We estimate the difference between $\Phi_\epsilon(p)$ and $\Phi_0(p)$ (for the case of $\epsilon = 0$). Let $p^\epsilon(t), q^\epsilon(t)$ $(\epsilon \geq 0)$ be the solution of

$$\begin{cases} p_t + Ap = f_1(p+q), \\ q_t + \tilde{A}q = f_2(p+q), \\ \dfrac{\partial}{\partial\nu}q = \epsilon g(p+q), \\ p(0) = p_0, \quad q \text{ is } H^4\text{-bounded}. \end{cases}$$

We set

$$\tilde{p}(t) = p^\epsilon(t) - p^0(t), \qquad \tilde{q}(t) = q^\epsilon(t) - q^0(t).$$

It is easily shown that

$$\begin{cases} \dfrac{1}{2}\dfrac{d}{dt}\|A\tilde{p}\|^2 \geq -\lambda_N\|A\tilde{p}\|^2 - (K_1+\epsilon C_N)(\|A\tilde{p}\|+\|\tilde{A}\tilde{q}\|)\|A\tilde{p}\|, \\ \dfrac{1}{2}\dfrac{d}{dt}\|\tilde{A}\tilde{q}\|^2 \leq -\|A\tilde{q}\|_1^2 + \left|\displaystyle\int_{\partial\Omega}\tilde{A}\tilde{q}\Big(\epsilon h(p^\epsilon+q^\epsilon)+\epsilon\frac{\partial f_2}{\partial\nu}(p^\epsilon+q^\epsilon)g(p^\epsilon+q^\epsilon)\Big)dS\right| \\ \qquad\qquad + (K_2+\epsilon C_N)(\|A\tilde{p}\|+\|\tilde{A}\tilde{q}\|)\|A\tilde{p}\|, \end{cases}$$

from which it follows that

$$\begin{cases} \dfrac{1}{2}\dfrac{d}{dt}\|A\tilde{p}\|^2 \geq -\lambda_N\|A\tilde{p}\|^2 - (K_1+\epsilon C_N)(\|A\tilde{p}\|+\|\tilde{A}\tilde{q}\|)\|A\tilde{p}\|, \\ \dfrac{1}{2}\dfrac{d}{dt}\|\tilde{A}\tilde{q}\|^2 \leq -\|A\tilde{q}\|_1^2 + \epsilon C_N\|\tilde{A}\tilde{q}\|_1 + (K_2+\epsilon C_N)(\|A\tilde{p}\|+\|\tilde{A}\tilde{q}\|)\|A\tilde{p}\|. \end{cases} \tag{3.33}$$

Set

$$X(t) = \|A\tilde{p}(t)\|e^{\gamma t}, \qquad Y(t) = \|\tilde{A}\tilde{q}(t)\|e^{\gamma t}.$$

Apply the argument of Step 1 in the proof of Proposition 3.5 to this case. We easily obtain for all $\delta \in [0,1]$ that

$$\begin{cases} \|A\tilde{p}(t)\| \leq \min(\epsilon C_N e^{-\gamma t}, 4R_2) \leq \epsilon^\delta C_N e^{-\delta\gamma t} & (t \leq 0), \\ \|\tilde{A}\tilde{q}(t)\| \leq \min(\epsilon C_N e^{-\gamma t}, 4R_2) \leq \epsilon^\delta C_N e^{-\delta\gamma t} & (t \leq 0). \end{cases} \tag{3.34}$$

In particular, if we take $t = 0$,

$$\|\tilde{A}(\Phi_\epsilon(p_0) - \Phi_0(p_0))\| \leq \epsilon C_N.$$

Next we investigate the Fréchet derivative of the difference. Put

$$\tilde{\rho}(t;p_0,\xi) = \rho^\epsilon(t;p_0,\xi) - \rho^0(t;p_0,\xi), \qquad \tilde{\sigma}(t;p_0,\xi) = \sigma^\epsilon(t;p_0,\xi) - \sigma^0(t;p_0,\xi),$$

and simply write

$$u^\epsilon = p^\epsilon(t;p_0) + q^\epsilon(t;p_0), \qquad \eta^\epsilon = \rho^\epsilon(t;p_0,\xi) - \sigma^\epsilon(t;p_0,\xi),$$

where $\rho^\epsilon + \sigma^\epsilon$ is the solution of

$$\begin{cases} \rho_t + A\rho = \dfrac{\partial}{\partial u}f_1(p^\epsilon(t;p_0)+q^\epsilon(t;p_0))(\rho+\sigma), \\ \sigma_t + \tilde{A}\sigma = \dfrac{\partial}{\partial u}f_2(p^\epsilon(t;p_0)+q^\epsilon(t;p_0))(\rho+\sigma), \\ \dfrac{\partial}{\partial\nu}\sigma = \epsilon g(p^\epsilon(t;p_0)+q^\epsilon(t;p_0))(\rho+\sigma). \end{cases} \tag{3.35}$$

Then $\tilde{\rho}, \tilde{\sigma}$ satisfy

$$\begin{cases} \tilde{\rho}_t + A\tilde{\rho} = \dfrac{\partial}{\partial u}f_1(u^\epsilon)\eta^\epsilon - \dfrac{\partial}{\partial u}f_1(u^0)\eta^0, \\ \tilde{\sigma}_t + \tilde{A}\tilde{\sigma} = \dfrac{\partial}{\partial u}f_2(u^\epsilon)\eta^\epsilon - \dfrac{\partial}{\partial u}f_2(u^0)\eta^0, \\ \dfrac{\partial}{\partial\nu}\tilde{\sigma} = \epsilon g(u^\epsilon)\eta^\epsilon. \end{cases} \tag{3.36}$$

The same computation for (3.30) can apply to (3.35) and yield

$$\int_{-\infty}^{t}\|\tilde{A}\sigma^\epsilon\|_1^2 e^{2\gamma s}ds \leq C_N\|A\xi\|^2 e^{2(\gamma-\gamma_1)t}. \tag{3.37}$$

By an argument similar to that in §3.3 and (3.34), (3.36) leads to

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|A\tilde{\rho}\|^2 \geq & -\lambda_N\|A\tilde{\rho}\|^2 - \epsilon^\delta(C+\epsilon C_N)\|A\xi\|e^{-\gamma_1 t-\delta\gamma t}\|A\tilde{\rho}\| \\
& -(K_1+\epsilon C_N)(\|A\tilde{\rho}\|+\|\tilde{A}\tilde{\sigma}\|)\|\tilde{\rho}\|, \\
\frac{1}{2}\frac{d}{dt}\|\tilde{A}\tilde{\sigma}\|^2 \leq & -\|\tilde{A}\tilde{\sigma}\|_1^2 + \left|\int_{\partial\Omega}\tilde{A}\tilde{\sigma}\frac{\partial}{\partial\nu}\left(-\tilde{\sigma}_t + \frac{\partial}{\partial u}f_2(u^\epsilon)\eta^\epsilon - \frac{\partial}{\partial u}f_2(u^0)\eta^0\right)dS\right| \\
& +\epsilon^\delta(C+\epsilon C_N)\|A\xi\|e^{-\gamma_1 t-\delta\gamma t}\|A\tilde{\sigma}\| + (K_2+\epsilon C_N)(\|A\tilde{\rho}\|+\|\tilde{A}\tilde{\sigma}\|)\|\tilde{\sigma}\|,
\end{aligned}$$

from which

(3.38)

$$\begin{cases}
\dfrac{1}{2}\dfrac{d}{dt}\|A\tilde{\rho}\|^2 \geq -\lambda_N\|A\tilde{\rho}\|^2 - \epsilon^\delta(C+\epsilon C_N)\|A\xi\|e^{-\gamma_1 t-\delta\gamma t}\|A\tilde{\rho}\| \\
\qquad\qquad -(K_1+\epsilon C_N)(\|A\tilde{\rho}\|+\|\tilde{A}\tilde{\sigma}\|)\|\tilde{\rho}\|, \\
\dfrac{1}{2}\dfrac{d}{dt}\|\tilde{A}\tilde{\sigma}\|^2 \leq -\|\tilde{A}\tilde{\sigma}\|_1^2 + (K_2+\epsilon C_N)(\|A\tilde{\rho}\|+\|\tilde{A}\tilde{\sigma}\|)\|\tilde{\sigma}\| \\
\qquad\qquad +\epsilon^\delta(C+\epsilon C_N)\|A\xi\|e^{-\gamma_1 t-\delta\gamma t}\|A\tilde{\sigma}\| + \epsilon C_N(\|A\xi\|e^{\gamma t}+\|\tilde{A}\sigma^\epsilon\|_1)\|\tilde{A}\tilde{\sigma}\|_1.
\end{cases}$$

We define

$$\begin{cases}
\tilde{X}(t)^2 = \displaystyle\int_{-\infty}^t \|A\tilde{\rho}(s;p_0,\xi)\|^2 e^{2\gamma s}ds, \\
\tilde{Y}(t)^2 = \displaystyle\int_{-\infty}^t \|\tilde{A}\tilde{\sigma}(s;p_0,\xi)\|^2 e^{2\gamma s}ds, \\
\tilde{Y}_1(t)^2 = \displaystyle\int_{-\infty}^t \|\tilde{A}\tilde{\sigma}(s;p_0,\xi)\|_1^2 e^{2\gamma s}ds.
\end{cases}$$

By the squeezing property

$$0 \leq \tilde{X}(t) \leq \tilde{K}, \quad 0 \leq \tilde{Y}(t) \leq \tilde{K} \quad (t \leq 0),$$

where

$$\tilde{K} = \frac{2(\epsilon^\delta(C+\epsilon C_N)+\epsilon C_N)}{\gamma-\gamma_1}\|A\xi\| \leq \epsilon^\delta C_N\|A\xi\|.$$

Then we have

$$\begin{cases}
\|A\tilde{\rho}\| \leq \epsilon^\delta C_N\|A\xi\|e^{-\gamma t}, \\
\|\tilde{A}\tilde{\sigma}\| \leq \epsilon^\delta C_N\|A\xi\|e^{-\gamma t},
\end{cases}$$

whereby

$$\left\|\left(\frac{\partial}{\partial p}\Phi_\epsilon(p_0)\xi - \frac{\partial}{\partial p}\Phi_0(p_0)\xi\right)\right\| \leq \epsilon^\delta C_N\|A\xi\|.$$

The proof of Theorem A (iii) is now complete. □

Proof of Corollary C. It is enough to show the derivation of the inertial form

(1.8). By (1.6) and the Taylor expansion of F and G, we have

$$p_t = PF(p+\Phi_\epsilon(p)) + \epsilon \sum_{j=1}^{m+1} \phi_j \int_{\partial\Omega} DG(p+\Phi_\epsilon(p))\phi_j dS$$
$$= PF(p) + P\frac{\partial}{\partial u}F(p)\Phi_\epsilon(p) + \epsilon \sum_{j=1}^{m+1} \phi_j \int_{\partial\Omega} DG(p)\phi_j dS + O(\epsilon^2).$$

Since the mean value of $DF(p)\Phi_\epsilon(p)$ over Ω vanishes, we have

$$p_t = PF(p) + \epsilon\frac{|\partial\Omega|}{|\Omega|}DG(p) + O(\epsilon^2).$$

Next we estimate the remaining term given by

$$\epsilon R(D,\epsilon,p) = PF(p+\Phi_\epsilon(p)) - PF(p) + \epsilon \sum_{j=1}^{m+1} \phi_j \int_{\partial\Omega} DG(p+\Phi_\epsilon(p))\phi_j dS - \epsilon\frac{|\partial\Omega|}{|\Omega|}DG(p).$$

The mean value theorem implies that

$$\epsilon R(D,\epsilon,p)$$
$$= \int_0^1 \left(PF'(p+\theta\Phi_\epsilon(p))\Phi_\epsilon(p) + \epsilon \sum_{j=1}^{m+1} \phi_j \int_{\partial\Omega} DG'(p+\theta\Phi_\epsilon(p))\Phi_\epsilon(p)\phi_j dS \right) d\theta$$
$$= \int_0^1 \left(\int_0^1 PF''(p+\theta\theta'\Phi_\epsilon(p))\theta\Phi_\epsilon(p)^2 d\theta' \right.$$
$$\left. + \epsilon \sum_{j=1}^{m+1} \phi_j \int_{\partial\Omega} DG'(p+\theta\Phi_\epsilon(p))\Phi_\epsilon(p)\phi_j dS \right) d\theta.$$

Then we obtain

$$R(D,\epsilon,p) = \int_0^1 \Big(\frac{1}{\epsilon} \int_0^1 PF''(p+\theta\theta'\Phi_\epsilon(p))\theta\Phi_\epsilon(p)^2 d\theta' d\theta$$
$$+ \int_0^1 \sum_{j=1}^{m+1} \phi_j \int_{\partial\Omega} DG'(p+\theta\Phi_\epsilon(p))\Phi_\epsilon(p)\phi_j dS \Big) d\theta,$$

by which

$$|R(D,\epsilon,p)| = O(\epsilon d_*), \qquad \left\| \frac{\partial R(D,\epsilon,p)}{\partial p} \right\| = O(\epsilon^\delta d_*). \qquad \square$$

Proof of Corollary D. Changing the variable $t = \epsilon s$, we have

$$\frac{1}{\epsilon}u_s = D\Delta u + F(u),$$

that is,

$$u_s = \tilde{D}\Delta u + \epsilon F(u). \tag{3.39}$$

Set

$$\tilde{A} = -\tilde{D}\Delta + cI, \qquad \tilde{F}(u) = \epsilon F(u) + cu,$$

where c is sufficiently small. In (3.39) the gap condition holds because K_1 and K_2 are close to 0 for sufficiently small ϵ and c. Hence Theorem A is applicable in this case. □

4. Application. (a) First consider the case $n = 1$ and $m = 1$ in (1.1) and $(1.2)_\epsilon$, that is,

$$\frac{\partial u}{\partial t} = d\frac{\partial^2 u}{\partial x^2} + F(u), \tag{4.1}$$

$$\begin{cases} \left.\dfrac{\partial u}{\partial x}\right|_{x=0} = -\epsilon G_0(u(t,0)), \\ \left.\dfrac{\partial u}{\partial x}\right|_{x=1} = \epsilon G_1(u(t,0)), \end{cases} \tag{4.2}_\epsilon$$

where F and G_i $(i = 0, 1)$ are sufficiently smooth functions and d is a positive constant. It is known that if $\epsilon = 0$ and if F satisfies

$$\lim_{|u|\to\infty} \frac{F(u)}{u} < 0,$$

then the equation has a global attractor. Moreover, under the assumption that every equilibrium solution is hyperbolic, the flow on the global attractor is Morse–Smale (see [2], [12], and [15]). Here we assume for $|u| \geq R$,

$$\begin{cases} \dfrac{F(u)}{u} < 0, \\ \dfrac{G_i(u)}{u} < 0 \qquad (i = 0, 1). \end{cases} \tag{4.3}$$

Then (4.1) with $(4.2)_\epsilon$ possesses a family of global attractors $\mathcal{A}_\epsilon$ $(\epsilon > 0)$ and for $u \in \mathcal{A}_\epsilon$, $\|u\|_{L^\infty} \leq R$. Indeed, if there is an x such that

$$|u(0,x)| > R,$$

consider solutions $\tilde{u}_\pm(t)$ satisfying

$$\frac{d}{dt}\tilde{u}_\pm = F(\tilde{u}_\pm), \tag{4.4}$$

$$\max_{0\leq x\leq 1} u(0,x) < \tilde{u}_+(0), \qquad \min_{0\leq x\leq 1} u(0,x) > \tilde{u}_-(0)$$

and apply the comparison theorem to $u(t,x)$ and $\tilde{u}_\pm(t)$. Then we see that there is a $t_0 > 0$ such that

$$\sup_x |u(t,x)| \leq R \quad \text{for } t \geq t_0.$$

Recall the fact that

$$\lambda_j = O(j^2) \quad \text{as } j \to \infty.$$

By this and Corollary C we have an inertial manifold $\mathcal{M}_\epsilon$ for any d. Then the inertial form $(\epsilon > 0)$ of Theorem A is C^1-perturbed from the one of $\epsilon = 0$. Assume that any equilibrium solution for $\epsilon = 0$ is hyperbolic. Considering the Morse–Smale property of the flow for the case $\epsilon = 0$, we can assert that the structure of the dynamics on the manifold $\mathcal{M}_\epsilon$ is equivalent to that on $\mathcal{M}_0$. (Note that the flow of Morse–Smale is structurally stable; see [12].)

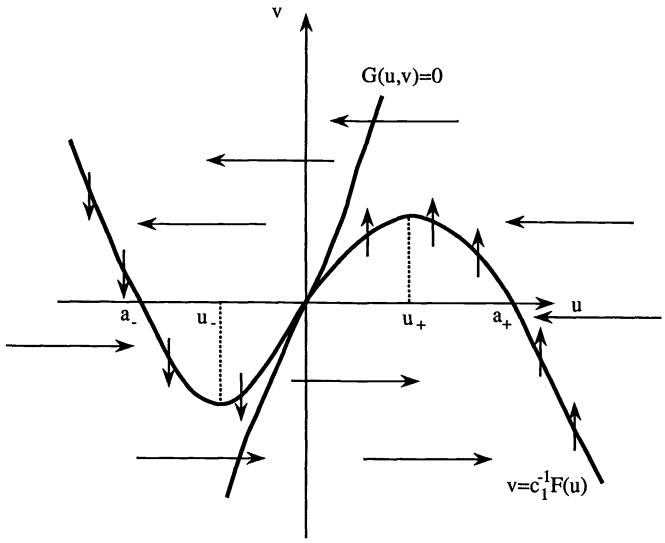


FIG. 3. *The graph of $F(u)$ and nullcline of $G(u,v)$. The arrows indicate schematic directions of the dynamics given by $u' = F(u) - c_1 v$, $v' = \epsilon c_2 G(u,v)$.*

(b) We consider the case of Corollary D. Let $\bar{p}$ be the hyperbolic equilibrium solution of

$$p_t = F(p) + \alpha \tilde{D} G(p). \tag{4.5}$$

Then for small ϵ there is a hyperbolic equilibrium solution $\bar{p}_\epsilon$ of (1.10) satisfying $\bar{p}_\epsilon \to \bar{p}$ as $\epsilon \to 0$. If, in addition, $\bar{p}$ is asymptotically stable for (4.5), then $\bar{p}_\epsilon$ is also for (1.10).

(c) Let us consider the following system of reaction-diffusion equations:

$$\begin{cases} u_t = d_1 \Delta u + F(u) \\ v_t = d_2 \Delta v \end{cases} \quad \text{in } \Omega, \tag{4.6}$$

and the boundary conditions

$$\begin{cases} \dfrac{\partial}{\partial \nu} u = \epsilon G_1(u,v) \\ \dfrac{\partial}{\partial \nu} v = \epsilon G_2(u,v) \end{cases} \quad \text{on } \partial\Omega, \tag{4.7}_\epsilon$$

where $F : \mathbb{R} \to \mathbb{R}$ and $G_i : \mathbb{R}^2 \to \mathbb{R}$ $(i = 1,2)$ are smooth functions. Suppose that F satisfies

$$\begin{cases} F'(u) > 0 \quad \text{in } (u_-, u_+) \quad (u_- < 0 < u_+), \\ F'(u) < 0 \quad \text{in } (-\infty, u_-) \cup (u_+, \infty), \\ F''(u_-) > 0, \qquad F''(u_+) < 0, \\ F(a_-) = F(0) = F(a_+) = 0, \qquad a_- < u_-, u_+ < a_+ \end{cases} \tag{4.8}$$

(see Fig. 3). The functions $G_i(u,v)$ $(i = 1,2)$ will be specified later. The system (4.6) with the homogeneous Neumann boundary condition (i.e., $\epsilon = 0$ in $(4.7)_\epsilon$) is a gradient system and every solution converges to an equilibrium solution $(u,v) = (\bar{u}, c)$, where $\bar{u}$ is a solution of

$$\begin{cases} d_1 \Delta u + F(u) = 0 & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.9}$$

and c is a constant. Nevertheless we can show that (4.6) and $(4.7)_\epsilon$ have a (relaxation-oscillating) periodic solution for each small $\epsilon > 0$, if we choose the functions $G_i(u, v)$ and d_i appropriately. Let us observe it below. First modify F outside

$$\{u \in \mathbb{R}; |u| \le R\} \qquad (R > 3|a_\pm|),$$

as satisfying

$$F(u) = -u, \qquad |u| > R. \tag{4.10}$$

Let G_i $(i = 1, 2)$ be functions such that

$$\begin{cases} G_1(u, v) = -v, \\ G_2(u, v) = u - h(v) & \left(|u| \le \dfrac{2}{3}R\right), \end{cases} \tag{4.11}$$

where

$$\begin{cases} h(0) = 0, \quad h'(v) \ge 0, \quad \dfrac{|\partial\Omega|}{|\Omega|} F'(0)h'(0) < 1, \\ |h(v)| < |F^{-1}(v)| \quad \text{if } v \in (F(u_-), 0) \cup (0, F(u_+)); \end{cases} \tag{4.12}$$

moreover,

$$G_i(u, v) = 0 \qquad (|u| \ge R, \quad i = 1, 2). \tag{4.13}$$

PROPOSITION 4.1. *Under the conditions* (4.8) *and* (4.10)–(4.13), *consider the equation* (4.6) *with* $(4.7)_\epsilon$. *For any sufficiently large* d_1 *and* d_2, *there is* ϵ_1 *such that for each* $\epsilon \in (0, \epsilon_1)$ *there exists a two-dimensional inertial manifold. In addition to the above conditions, assume that*

$$d_1 = \frac{1}{\epsilon}\,. \tag{4.14}$$

Then the equation has an asymptotically stable periodic solution.

Proof. The first part of the theorem immediately follows from a direct application of Corollary C of §1. To show the latter part we have to investigate the inertial form

$$\begin{cases} y_t = F(y) - c_1 z + r_1(\epsilon, y, z) = U(\epsilon, y, z), \\ z_t = \epsilon c_2 (y - h(z)) + \epsilon r_2(\epsilon, y, z) = \epsilon V(\epsilon, y, z), \end{cases} \tag{4.15}$$

where we put

$$\begin{cases} (y, z) = \dfrac{1}{|\Omega|} \left(\displaystyle\int_\Omega u dx, \int_\Omega v dx \right), \\ c_1 = \dfrac{|\partial\Omega|}{|\Omega|}, \qquad c_2 = d_2 c_1. \end{cases} \tag{4.16}$$

Applying the implicit function theorem to the equation

$$U(\epsilon, y, z) = 0, \qquad V(\epsilon, y, z) = 0, \tag{4.17}$$

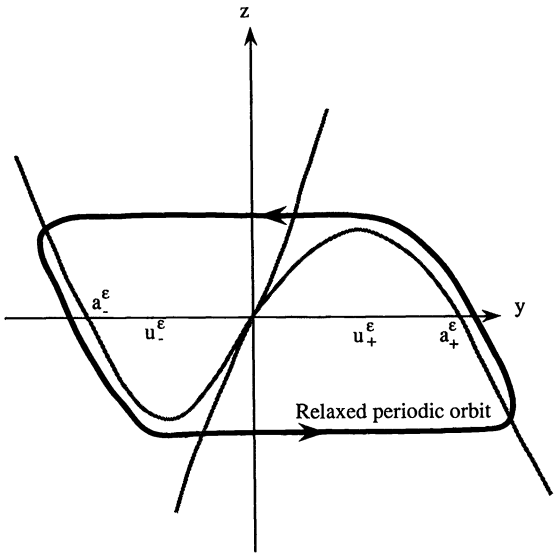


FIG. 4. *The bold line indicates the relaxed periodic solution to the reduced equation* (4.15).

and considering the conditions (4.8) and (4.11), there exist C^1-functions

$$z = \frac{F(y;\epsilon)}{c_1}, \qquad y = h(z;\epsilon),$$

which solve (4.17) and satisfy

$$F(y;\epsilon) \to F(y), \quad h(z;\epsilon) \to h(z) \quad \text{as } \epsilon \to 0.$$

Then condition (4.8) is also satisfied for $F(.;\epsilon)$ by replacing u_-, u_+, a_-, a_+ by suitable ones, say $u_-^\epsilon, u_+^\epsilon, a_-^\epsilon, a_+^\epsilon$, respectively. Note that

$$u_\pm^\epsilon \to u_\pm, \qquad a_\pm^\epsilon \to a_\pm.$$

The condition (4.12) implies that the Jacobian matrix of the right-hand side of (4.15) around $(y,z) = (0,0)$ has two eigenvalues with negative real part. With the phase plane analysis we easily show the existence of a periodic solution $(y,z) = (y_\epsilon(t), z_\epsilon(t))$ (see Fig. 4).

The asymptotic stability of the periodic solution can be proved by the sign of

$$\int_0^{T(\epsilon)} \left\{ \frac{\partial}{\partial y} U(\epsilon, y_\epsilon(t), z_\epsilon(t)) + \epsilon \frac{\partial}{\partial y} V(\epsilon, y_\epsilon(t), z_\epsilon(t)) \right\} dt$$

$$= \int_0^{T(\epsilon)} \{F'(y_\epsilon(t)) + O(\epsilon)\} dt,$$

where $T(\epsilon)$ denotes the period. If this value is negative, the periodic solution is asymptotically stable and only one periodic orbit exists in a neighborhood of it. This will be shown by the facts that the period $T(\epsilon)$ has a uniform bound in ϵ and that the periodic orbit $(y_\epsilon(t), z_\epsilon(t))$, $t \in [0, T(\epsilon)]$ converges to a set

$$\{(y,z); z = F(y), a_0 \le y \le u_-, u_+ \le y \le a_3\},$$

$$\cup \{(y,z); a_0 \le y \le u_+, z = F(u_+)\} \cup \{(y,z); u_- \le y \le a_3, z = F(u_-)\},$$

where a_0 and a_3 are such that

$$a_0 \in F^{-1}(F(u_+)), \qquad a_0 \neq u_+,$$
$$a_3 \in F^{-1}(F(u_-)), \qquad a_0 \neq u_-.$$

For a detailed discussion, see [4] and [23].

(d) In the fields of population biology, physiology, and biochemical reaction (e.g., [16] and [21]), one can find a compartment model of ODEs which describes a spatial model between species (respectively, chemical substances) distributed in several habitats (respectively, cells). Here we will introduce the compartment model from a system of reaction-diffusion equations (1.1) with $(1.2)_\epsilon$ by an application of Corollary C.

Let us consider the equations

$$\frac{\partial v^j}{\partial t} = D^j \frac{\partial v^j}{\partial x^2} + F^j(v^j) \quad \text{in } (L_{j-1}, L_j), \quad j = 1, \dots, l, \tag{4.18}$$

$$\begin{cases} \dfrac{\partial v^j}{\partial x}(t, L_{j-1}) = -\epsilon G^{j-1}(v^{j-1}(t, L_{j-1}), v^j(t, L_{j-1})), \\ \dfrac{\partial v^j}{\partial x}(t, L_j) = \epsilon G^j(v^j(t, L_j), v^{j+1}(t, L_j)), \end{cases} \quad j = 1, \dots, l, \tag{4.19}$$

where

$$v^j = (v_1^j, \dots, v_s^j)^T, \quad L_0 = 0, \quad L_j < L_{j+1},$$
$$D^j = \text{ diag } (d_1^j, \dots, d_s^j),$$
$$F^j : \mathbb{R}^s \longrightarrow \mathbb{R}^s,$$

and it is assumed that

$$G^0 = 0, \qquad G^l = 0.$$

Put

$$u^j(t,x) = \begin{cases} v^j\left(t, L_{j-1} + \dfrac{x}{L_1}(L_j - L_{j-1})\right) & (j\text{: odd}), \\ v^j\left(t, L_{j-1} - \dfrac{x - L_1}{L_1}(L_j - L_{j-1})\right) & (j\text{: even}). \end{cases}$$

Then (4.18) and (4.19) are equivalent to

$$\frac{\partial u^j}{\partial t} = c_j^2 D^j \frac{\partial u^j}{\partial x^2} + F^j(u^j) \quad \text{in } (0, L_1), \quad j = 1, \dots, l, \tag{4.20}$$

$$\begin{cases} \dfrac{\partial u^j}{\partial x}(t,0) = \begin{cases} -\epsilon c_j G^{j-1}(u^{j-1}(t,0), u^j(t,0)) & (j\text{: odd}), \\ -\epsilon c_j G^j(u^j(t,0), u^{j+1}(t,0)) & (j\text{: even}), \end{cases} \\ \dfrac{\partial u^j}{\partial x}(t, L_1) = \begin{cases} \epsilon c_j G^j(u^j(t, L_1), u^{j+1}(t, L_1)) & (j\text{: odd}), \\ \epsilon c_j G^{j-1}(u^{j-1}(t, L_1), u^j(t, L_1)) & (j\text{: even}), \end{cases} \end{cases} \tag{4.21}$$

where

$$c_j = \frac{L_j - L_{j-1}}{L_1}.$$

Hence this system of equations is in the class of (1.1) and $(1.2)_\epsilon$. Under the conditions (C.1) and (C.2'), we have an inertial manifold, and its inertial form is written as

$$\frac{d\bar{u}_j}{dt} = F^j(\bar{u}^j) + \frac{\epsilon c_j^3}{L_1} D^j\{G^{j-1}(\bar{u}^{j-1}, \bar{u}^j) + G^j(\bar{u}^j, \bar{u}^{j+1})\} + O(\epsilon^2). \tag{4.22}$$

Appendix: Proofs of Lemmas 2.4, 3.2, and 3.7. First recall the definitions of f_1, f_2, and g:

$$\begin{cases} f_1(p+q) = \chi(\|Ap\|^2 + \|q\|_1^2) \displaystyle\sum_{j=1}^{N} \left\{ (\tilde{F}(p+q), \phi_j) + \epsilon \int_{\partial\Omega} DG(p+q)\phi_j dS \right\} \phi_j, \\ f_2(p+q) = \chi(\|Ap\|^2 + \|q\|_1^2)\tilde{F}(p+q) - f_1(p+q), \\ g(p+q) = \chi(\|Ap\|^2 + \|q\|_1^2)G(p+q). \end{cases}$$

Let us introduce the following lemmas.

LEMMA A.1. *For* $\|APu\| \le 2R_1$, $\|\tilde{A}Qu\| \le 2R_1$, *and* $u_1, u_2, u_3 \in H^2$,

$$\left| \frac{\partial}{\partial u} \chi(\|APu\|^2 + \|Qu\|_1)u_1 \right| \le K_3(\|\tilde{A}Pu_1\| + \|Qu_1\|_1),$$

$$\left| \frac{\partial^2}{\partial u^2} \chi(\|APu\|^2 + \|Qu\|_1)u_1u_2 \right| \le K_3(\|\tilde{A}Pu_1\| + \|Qu_1\|_1)(\|\tilde{A}Pu_2\| + \|Qu_2\|_1),$$

$$\left| \frac{\partial^3}{\partial u^3} \chi(\|APu\|^2 + \|Qu\|_1)u_1u_2u_3 \right| \le K_3(\|\tilde{A}Pu_1\| + \|Qu_1\|_1) \times (\|\tilde{A}Pu_2\| + \|Qu_2\|_1)(\|\tilde{A}Pu_3\| + \|Qu_3\|_1),$$

where $K_3 = C(1+R_1)^3$.

Proof. Differentiate χ, and we have

$$\frac{\partial \chi}{\partial u}(\cdot)u_1 = \chi'(\cdot)\{(APu, APu_1) + (Qu, Qu_1)_1\},$$

$$\frac{\partial^2 \chi}{\partial u^2}(\cdot)u_1u_2 = \chi''(\cdot)\{(APu, APu_1) + (Qu, Qu_1)_1\}\{(APu, APu_2) + (Qu, Qu_2)_1\} + \chi'(\cdot)\{(APu_2, APu_1) + (Qu_2, Qu_1)_1\},$$

$$\begin{aligned} \frac{\partial^3 \chi}{\partial u^3}(\cdot)u_1u_2u_3 = {} & \chi'''(\cdot)\{(APu, APu_1) + (Qu, Qu_1)_1\}\{(APu, APu_2) + (Qu, Qu_2)_1\} \\ & \times \{(APu, APu_3) + (Qu, Qu_3)_1\} \\ & + \chi''(\cdot)\{(APu_3, APu_1) + (Qu_3, Qu_1)_1\}\{(APu, APu_2) + (Qu, Qu_2)_1\} \\ & + \chi''(\cdot)\{(APu, APu_1) + (Qu, Qu_1)_1\}\{(APu_3, APu_2) + (Qu_3, Qu_2)_1\} \\ & + \chi''(\cdot)\{(APu_2, APu_1) + (Qu_2, Qu_1)_1\}\{(APu, APu_3) + (Qu, Qu_3)_1\}. \end{aligned}$$

Taking the moduli, we can obtain the desired inequalities. □

LEMMA A.2. *Suppose the assumption of Lemma* A.1. *Then*

$$\begin{cases} \left\| \dfrac{\partial g}{\partial u}(u)u_1 \right\|_{L^2(\partial\Omega)} \le K_4(\|APu_1\| + \|Qu_1\|_1), \\ \left\| \dfrac{\partial^2 g}{\partial u^2}(u)u_1u_2 \right\|_{L^2(\partial\Omega)} \le K_4(\|APu_1\| + \|Qu_1\|_1)(\|APu_2\| + \|Qu_2\|_1) \\ \qquad\qquad + K_4\|u_1\|_1 \ \|Au_2\|, \\ \left\| \dfrac{\partial^3 g}{\partial u^3}(u)u_1u_2u_3 \right\|_{L^2(\partial\Omega)} \le K_4(\|APu_1\| + \|Qu_1\|_1)(\|APu_2\| + \|Qu_2\|_1) \\ \qquad\qquad \times (\|APu_3\| + \|Qu_3\|_1) + K_4\|u_1\|_1 \ \|\tilde{A}u_2\| \ \|\tilde{A}u_3\|, \end{cases}$$

where u_1, u_2, *and* u_3 *belong to* H^2 *with* $\|\partial u_i/\partial\nu\|_{H^1(\partial\Omega)} \le C\epsilon\|u_i\|_{H^1(\partial\Omega)}$ $(i = 1, 2, 3)$ *and* $K_4 = C|G|_3(1 + R_1)^3$.

First differentiate the mapping $g = \chi G$:

$$\frac{\partial g}{\partial u}(u)u_1 = \chi G'(\cdot)u_1 + G(\cdot)\frac{\partial \chi}{\partial u}(\cdot)u_1,$$

$$\frac{\partial^2 g}{\partial u^2}(u)u_1u_2 = \chi G''(\cdot)u_1u_2 + \frac{\partial \chi}{\partial u}(\cdot)u_2G'(\cdot)u_1 + \frac{\partial^2 \chi}{\partial u^2}(\cdot)u_1u_2G + \frac{\partial \chi}{\partial u}(\cdot)u_1G'(\cdot)u_2,$$

$$\begin{aligned} \frac{\partial^3 g}{\partial u^3}(u)u_1u_2u_3 &= \chi G'''(\cdot)u_1u_2u_3 + \frac{\partial \chi}{\partial u}(\cdot)u_3G''(\cdot)u_1u_2 \\ &+ \frac{\partial^2 \chi}{\partial u}(\cdot)u_2u_3G'(\cdot)u_1 + \frac{\partial \chi}{\partial u}(\cdot)u_2G''(\cdot)u_1u_3 + \frac{\partial^2 \chi}{\partial u^2}(\cdot)u_1u_2G'(\cdot)u_3 \\ &+ \frac{\partial^3 \chi}{\partial u^3}(\cdot)u_1u_2u_3G + \frac{\partial^2 \chi}{\partial u^2}(\cdot)u_1u_3G'(\cdot)u_2 + \frac{\partial \chi}{\partial u}(\cdot)u_1G''(\cdot)u_2u_3. \end{aligned}$$

By Lemma 2.1,

$$\|u_1u_2\|_{L^2(\partial\Omega)} \le C\|u_1\|_1 \ \|\tilde{A}u_2\|,$$
$$\|u_1u_2u_3\|_{L^2(\partial\Omega)} \le C\|u_1\|_1 \ \|\tilde{A}u_2\| \ \|\tilde{A}u_3\|.$$

From these inequalities and Lemma A.1, we see the result of the lemma. □

Before the proofs of Lemmas 2.4, 3.2, and 3.8, we prepare the next lemma.

LEMMA A.3. *Under* $\|APu\| \le 2R_1$, $\|\tilde{A}Qu\| \le 2R_1$, *and* $\|\partial u/\partial\nu\|_{H^1(\partial\Omega)} \le C\epsilon$,

(i) $\|\tilde{A}\tilde{F}\| \le C|\tilde{F}|_2(1 + R_1)^2$,

(ii) $\|Af_1\| \le C|\tilde{F}|_2(1 + R_1)^2 + \epsilon C_N|G|(1 + |\tilde{F}|_1)$,

(iii) $\|\tilde{A}f_2\| \le C|\tilde{F}|_2(1 + R_1)^2 + \epsilon C_N|G|(1 + |\tilde{F}|_1)$,

(iv) $\|\nabla\Delta\tilde{F}\| \le C|\tilde{F}|_3(1 + R_1)^3(1 + \lambda_N^{\frac{1}{2}} + \|\tilde{A}q\|_1)$.

Proof. We can easily get (i) by Leibniz's formula. The inequality (iii) is shown by (i) and (ii). Here we consider (ii) and (iv). From the definition of f_1,

$$\begin{aligned} \|Af_1(p + q)\|^2 &= \left\| \sum_{j=1}^{N} \chi \left\{ (\tilde{F}, A\phi_j) + \epsilon \int_{\partial\Omega} DG\phi_j dS \right\} \phi_j \right\|^2 \\ &\le 2 \left\| \sum_{j=1}^{N} \chi(\tilde{F}, A\phi_j)\phi_j \right\|^2 + 2 \left\| \sum_{j=1}^{N} \phi_j \int_{\partial\Omega} \epsilon\chi DG\phi_j dS \right\|^2. \end{aligned}$$

By

$$\begin{aligned}(\tilde{F}, A\phi_j) =&(\tilde{A}\tilde{F}, \phi_j) + \int_{\partial\Omega} \phi_j D\frac{\partial}{\partial\nu}\tilde{F}dS \\ =&(\tilde{A}\tilde{F}, \phi_j) + \int_{\partial\Omega} \phi_j D\frac{\partial\tilde{F}}{\partial u}\epsilon\chi G dS,\end{aligned}$$

we have

$$\|Af_1(p+q)\|^2 \leq 3\left(|\chi|\sum_{j=1}^{N}(\tilde{A}\tilde{F}, \phi_j)^2 + \epsilon C\lambda_N^2(1+N)^2|G|^2(1+|\tilde{F}|_1)^2\right).$$

Hence

$$\|Af_1(p+q)\| \leq C\|\tilde{A}\tilde{F}\| + \epsilon C_N|G|(1+|\tilde{F}|_1).$$

Next we obtain

$$\left\|\frac{\partial^3}{\partial x_1^3}\tilde{F}\right\| \leq C|\tilde{F}'|\left\|\frac{\partial^3 u}{\partial x_1^3}\right\| + 3|\tilde{F}''|\left\|\frac{\partial^2 u}{\partial x_1^2}\frac{\partial u}{\partial x_1}\right\| + |\tilde{F}'''|\left\|\frac{\partial u}{\partial x_1}\right\|_{L^6}^3, \tag{A.1}$$

by Leibniz's formula. From Sobolev embeddings in §2.1,

$$\begin{aligned}\|p+q\|_{H^3} \leq& C(\|p\|_{H^3} + \|q\|_{H^3}) \\ \leq& C\left(\lambda_N^{\frac{1}{2}}\|Ap\| + \|\tilde{A}q\|_1 + \|p\| + \|q\| + \left\|\frac{\partial q}{\partial\nu}\right\|_{H^2(\partial\Omega)}\right), \\ \|p+q\|_{W^{2,4}}\|p+q\|_{W^{1,4}} \leq& C\|p+q\|_{H^3}\|p+q\|_{H^2} \\ \leq& C\left(\lambda_N^{\frac{1}{2}}\|Ap\| + \|\tilde{A}q\|_1 + \|p\| + \|q\| + \left\|\frac{\partial q}{\partial\nu}\right\|_{H^2(\partial\Omega)}\right) \\ &\times\left(\|Ap\| + \|\tilde{A}q\| + \|p\| + \|q\| + \left\|\frac{\partial q}{\partial\nu}\right\|_{H^1(\partial\Omega)}\right), \\ \|p+q\|_{W^{1,6}}^3 \leq& C\left(\|Ap\|^3 + \|\tilde{A}q\|^3 + \|p\|^3 + \|q\|^3 + \left\|\frac{\partial q}{\partial\nu}\right\|_{H^1(\partial\Omega)}^3\right).\end{aligned}$$

Applying these three inequalities to (A.1), we have

$$\left\|\frac{\partial^3}{\partial x_1^3}\tilde{F}\right\| \leq C|\tilde{F}|_3(1+R_1)^3(1+\lambda_N^{\frac{1}{2}} + \|\tilde{A}q\|_1).$$

By this we can easily obtain (iv). □

We now prove Lemma 2.4.

Proof of Lemma 2.4. We first prove (i). We see that

$$\begin{aligned}\|\tilde{A}f_2(p+q)\|_1 \leq&\|\chi\tilde{A}\tilde{F}(p+q)\|_1 + \|Af_1(p+q)\|_1 \\ \leq& C\|\chi D^{\frac{3}{2}}\nabla\Delta\tilde{F}\| + C\|\chi\Delta\tilde{F}\| + \lambda_N^{\frac{1}{2}}\|Af_1(p+q)\|.\end{aligned} \tag{A.2}$$

The inequality (i) immediately follows from Lemma A.3.

Now we prove (ii) of the lemma. Since

$$\begin{aligned}\frac{\partial}{\partial\nu}(f_2(p+q)_t) =&\frac{\partial}{\partial\nu}(\chi\tilde{F}-f_1)_t = \frac{\partial}{\partial\nu}(\chi\tilde{F})_t\\ =&\frac{\partial}{\partial\nu}\left(\frac{\partial\chi}{\partial u}(\cdot)u_t\tilde{F}+\chi\tilde{F}'(\cdot)u_t\right)\\ =&\frac{\partial\chi}{\partial u}(\cdot)u_t\tilde{F}'\frac{\partial q}{\partial\nu}+\chi\tilde{F}''\frac{\partial q}{\partial\nu}(\cdot)(p_t+q_t)+\chi\tilde{F}'\frac{\partial q_t}{\partial\nu},\end{aligned}$$

we have

$$\begin{aligned}&\left\|\frac{\partial}{\partial\nu}(f_2(p+q)_t)\right\|_{L^2(\partial\Omega)}\\ &\quad\le\epsilon K_3(\|Ap_t\|+\|q_t\|_1)|\tilde{F}'|\|g\|_{L^2(\partial\Omega)}+\epsilon|\chi||\tilde{F}''|\|\chi DG\|_{L^\infty(\partial\Omega)}\|p_t+q_t\|_{L^2(\partial\Omega)}\\ &\qquad+\epsilon|\chi||\tilde{F}'|\|g_t\|_{L^2(\partial\Omega)}\\ &\quad\le\epsilon C|\tilde{F}|_2|G|_2(\|Ap_t\|+\|q_t\|_1)\\ &\quad\le\epsilon C_N(1+\|\tilde{A}q\|_1),\end{aligned}$$

where we used (i) of this lemma. Next consider

$$\begin{aligned}\frac{\partial}{\partial\nu}q_{tt}=\epsilon g(u)_{tt}&=\epsilon\Big(\frac{\partial g}{\partial u}(\cdot)u_t\Big)_t\\ &=\epsilon\frac{\partial^2 g}{\partial u^2}(\cdot)u_tu_t+\epsilon\frac{\partial g}{\partial u}(\cdot)u_{tt}.\end{aligned}$$

Applying Lemma A.2 to the above equation yields

$$\begin{aligned}\|D\frac{\partial}{\partial\nu}q_{tt}\|\le&\epsilon K_4(\|Ap_t\|+\|q_t\|_1)^2+\epsilon K_4(\|Ap_{tt}\|+\|q_{tt}\|_1)\\ \le&\epsilon C_N\|\tilde{A}^2q\|_1+\epsilon C(1+R_1)^2.\end{aligned}$$

□

Let us set

$$\eta=\rho+\sigma,\qquad\tilde{\eta}=\tilde{\rho}+\tilde{\sigma},$$

throughout the rest of this paper.

Proof of Lemma 3.2. We first consider (i). By the definition of f_1, we have

$$\begin{aligned}\frac{\partial}{\partial u}f_1(u)\eta=&\frac{\partial\chi}{\partial u}(\cdot)\eta\Big(P\tilde{F}+\epsilon\sum\phi_j\int_{\partial\Omega}DG\phi_j dS\Big)\\ &+\chi(\cdot)\left(P\tilde{F}'(\cdot)\eta+\epsilon\sum\phi_j\int_{\partial\Omega}DG'(\cdot)\eta\phi_j dS\right).\end{aligned}$$

Lemma A.1 implies

$$\begin{aligned}\left\|A\frac{\partial}{\partial u}f_1(u)\eta\right\|\le&K_4(\|A\rho\|+\|\sigma\|_1)\left\|AP\tilde{F}+\epsilon\sum\lambda_j\phi_j\int_{\partial\Omega}DG\phi_j dS\right\|\\ &+|\chi|\|AP\tilde{F}'(\cdot)\eta\|+\epsilon\lambda_N|\chi||DG'|\ \|\eta\|_{L^2(\partial\Omega)}\|\phi_j\|_{L^\infty}.\end{aligned}\tag{A.3}$$

By Lemma A.3 we have

$$\|AP\tilde{F}(u)\|\le C(1+\epsilon|G|_1)|\tilde{F}|_2(1+R_1)^4\quad\text{for }\|Ap\|\le 2R_1,\quad\|\tilde{A}q\|\le 2R_1.$$

Similarly, we obtain

$$\|\tilde{A}\tilde{F}'(\cdot)\eta\| \le C|\tilde{F}|_3 R_1^2(\|A\rho\| + \|\tilde{A}\sigma\|)$$

by using the integration by parts. Substituting these inequalities into (A.3) yields

$$\begin{aligned}\left\|A\frac{\partial}{\partial u}f_1(u)\eta\right\| \le& C(1+\epsilon|G|_1)|\tilde{F}|_2(1+R_1)^2(\|A\rho\| + \|\sigma\|_1)\\ &+ C(|\tilde{F}|_3 R_1^2 + \epsilon|G|_1)(\|A\rho\| + \|\tilde{A}\sigma\|).\end{aligned} \tag{A.4}$$

Then we obtain

$$\left\|A\frac{\partial}{\partial u}f_1(u)(\rho+\sigma)\right\| \le \{C|\tilde{F}|_3(1+R_1)^4 + \epsilon C_N\}(\|A\rho\| + \|\tilde{A}\sigma\|).$$

The inequality (ii) immediately follows from the fact that

$$\frac{\partial}{\partial u}f_2(\cdot)\eta = \frac{\partial}{\partial u}(\chi(\cdot)\tilde{F}(\cdot)\eta) - \frac{\partial}{\partial u}f_1(\cdot)\eta.$$

We turn to (iii). It is easily obtained that

$$\frac{\partial}{\partial\nu}\frac{\partial}{\partial u}(\chi\tilde{F})(\rho+\sigma) = \frac{\partial\chi}{\partial u}\eta\tilde{F}'\frac{\partial q}{\partial\nu} + \chi\tilde{F}''\eta\frac{\partial q}{\partial\nu} + \epsilon\chi\tilde{F}'\frac{\partial g}{\partial u}(v)\eta,$$

where we used

$$\frac{\partial\rho}{\partial\nu} = 0, \quad \frac{\partial\sigma}{\partial\nu} = \epsilon\frac{\partial g}{\partial u}(v)\eta \quad \text{for some } v \in H^2.$$

By Lemmas A.1 and A.2, we can easily obtain

$$\left\|\frac{\partial}{\partial\nu}\frac{\partial}{\partial u}f_2(\cdot)\eta\right\|_{L^2(\partial\Omega)} \le \epsilon C|\tilde{F}|_2|G|_1(\|Ap\| + \|\tilde{A}\sigma\|).$$

Now we calculate the Fréchet derivative of $h(u)$. Since

$$h(u) = \frac{\partial g}{\partial u}(u)k(u), \qquad k(u) = -\tilde{A}u + f_1 + f_2, \tag{A.5}$$

we have

$$\begin{aligned}\frac{\partial}{\partial u}h(u)\eta =& \frac{\partial^2 g}{\partial u^2}(u)\eta k(u) + \frac{\partial g}{\partial u}(u)\frac{\partial k}{\partial u}(u)\eta\\ =& \frac{\partial^2 g}{\partial u^2}(u)\eta k(u) + \frac{\partial g}{\partial u}(u)\left(-\tilde{A}\eta + \frac{\partial f_1}{\partial u}(u)\eta + \frac{\partial f_2}{\partial u}(u)\eta\right).\end{aligned} \tag{A.6}$$

By using (A.6), Lemmas A.2 and A.3, and the condition $\|\tilde{A}^2 q\| \le 2R_3$, we have

$$\begin{aligned}\left\|\frac{\partial}{\partial u}h(u)\eta\right\|_{L^2(\partial\Omega)} \le& K_4(\|A\rho\| + \|\sigma\|_1)(\|APk(u)\| + \|Qk(u)\|_1)\\ &+ K_4\left(\left\|-A^2\rho + A\frac{\partial f_1}{\partial u}(\cdot)\eta\right\| + \left\|-A\sigma + \frac{\partial f_2}{\partial u}(\cdot)\eta\right\|_1\right),\end{aligned}$$

which immediately implies (iv). □

Next we prove Lemma 3.7.

Proof of Lemma 3.7. All the inequalities (i)–(iv) can be shown in the same manner as obtained above. To avoid repeating a lengthy computation, we only give a remark on the case of (iv). It is apparently difficult to handle (iv) because it includes the higher-order derivatives.

Differentiating (A.6) yields

$$\begin{aligned}\frac{\partial^2 h}{\partial u^2}(\cdot)\eta\tilde{\eta} =& \frac{\partial^3 g}{\partial u^3}(\cdot)\eta\tilde{\eta}k(u) + \frac{\partial^2 g}{\partial u^2}(\cdot)\eta\left(-\tilde{A}\tilde{\eta} + \frac{\partial f_1}{\partial u}(\cdot)\tilde{\eta} + \frac{\partial f_2}{\partial u}(\cdot)\tilde{\eta}\right)\\ &+ \frac{\partial^2 g}{\partial u^2}(\cdot)\tilde{\eta}\left(-\tilde{A}\eta + \frac{\partial f_1}{\partial u}(\cdot)\eta + \frac{\partial f_2}{\partial u}(\cdot)\eta\right)\\ &+ \frac{\partial g}{\partial u}(\cdot)\left(\frac{\partial^2 f_1}{\partial u^2}(\cdot)\tilde{\eta}\eta + \frac{\partial^2 f_2}{\partial u^2}(\cdot)\tilde{\eta}\eta\right).\end{aligned} \tag{A.7}$$

We can apply Lemma A.2 to this case and use the inequalities (i) and (ii) of Lemma 3.2 and (i) and (ii) of Lemma 3.7 to obtain the desired one. Since it is an easy exercise, we omit the details. □

Acknowledgments. The authors would like to thank Prof. V. Kalantarov (Academy of Sciences, Azerbaidzhan) for his useful comments.

REFERENCES

[1] S. AGMON, A. DOUGLIS, AND L. NIRENBERG, *Estimates near the boundary for the solutions of elliptic partial differential equations satisfying general boundary conditions* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] S. B. ANGENENT, *The Morse–Smale property for a semi-linear parabolic equation*, J. Differential Equations, 62 (1986), pp. 427–442.

[3] P. W. BATES AND C. K. R. T. JONES, *Invariant manifolds for semilinear partial differential equations*, Dynamics Reported, Vol. 2, 1989, pp. 1–38.

[4] C. BONET, *Singular perturbation of relaxed periodic orbits*, J. Differential Equations, 66 (1987), pp. 301–339.

[5] S. N. CHOW AND K. LU, *Invariant manifolds for flows in Banach spaces*, J. Differential Equations, 74 (1988), pp. 285–317.

[6] S.-N. CHOW, K. LU, AND G. R. SELL, *Smoothness of inertial manifolds*, IMA Preprint.

[7] P. CONSTANTIN, C. FOIAS, B. NICOLAENCO, AND R. TEMAM, *Spectral barriers and inertial manifolds for dissipative partial differential equations*, J. Dynamics and Differential Equations, 1 (1989), pp. 45–73.

[8] E. CONWAY, D. HOFF, AND J. SMOLLER, *Large time behavior of solutions of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math, 35 (1978), pp. 1–16.

[9] C. FOIAS, G. R. SELL, AND R. TEMAM, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309–353.

[10] A. FRIEDMAN, *Partial differential equations of parabolic type*, Prentice Hall, Englewood Cliffs, NJ, 1964.

[11] J. K. HALE, *Large diffusivity and asymptotic behavior in parabolic systems*, J. Math. Anal. Appl., 118 (1986), pp. 455–466.

[12] J. K. HALE, *Asymptotic Behavior of Dissipative Systems*, Amer. Math. Soc., Providence, RI, 1988.

[13] J. K. HALE AND C. ROCHA, *Varying boundary conditions with large diffusivity*, J. Math. Pures Appl., 66 (1987), pp. 139–158.

[14] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[15] D. HENRY, *Some infinite-dimensional Morse–Smale systems defined by parabolic partial differential equations*, J. Differential Equations, 59 (1985), pp. 165–205.

[16] K. KAWASAKI AND E. TERAMOTO, *Spatial pattern formation of prey-predator populations*, J. Math. Biology, 8 (1979), pp. 33–46.

[17] M. KWAK, *Finite dimensional inertial forms for the 2D Navier–Stokes equations*, Indiana Univ. Math. J., 42 (1992), pp. 927–982.

[18] J. MALLET-PARET AND G. R. SELL, *Inertial manifolds for reaction diffusion equations in higher space dimensions*, J. Amer. Math. Soc., 1 (1988), pp. 805–866.

[19] Y. MORITA, *Reaction-diffusion systems in nonconvex domains: Invariant manifold and reduced form*, J. Dynamics and Differential Equations, 2 (1990), pp. 69–115.

[20] Y. MORITA AND S. JIMBO, *ODE's on inertial manifolds for ration-diffusion systems in a singularly perturbed domain with several thin channels*, J. Dynamics and Differential Equations, 4 (1992), pp. 65–93.

[21] J. NEU, *Coupled chemical oscillators*, SIAM J. Appl. Math., 37 (1979), pp. 307–315.

[22] H. NINOMIYA, *Some remarks on inertial manifolds*, J. Math. Kyoto Univ., 32 (1992), pp. 667–688.

[23] J. J. STOKER, *Nonlinear Vibration*, Wiley, New York, 1950.

[24] R. TEMAM, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Applied Mathematics Series, Vol 68, Springer-Verlag, Berlin, 1988.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1357–1374, September 1994

006

A FREE BOUNDARY PROBLEM MODELING THERMAL INSTABILITIES: WELL-POSEDNESS*

MICHAEL L. FRANKEL† AND VICTOR ROYTBURD‡

Abstract. In this paper, the authors analyze a simple free boundary model associated with solid combustion and some phase transition processes. There is strong evidence that this "one-phase" model captures many salient features of dynamical behavior of more realistic (and complicated) combustion and phase transition models. The main result is a global existence and uniqueness theorem whose proof is based on a uniform a priori estimate on the growth of solutions. The techniques employed are quite elementary and involve some maximum principle type estimates as well as parabolic potential estimates for the equivalent integral equation.

Key words. free boundary problems, gasless combustion

AMS subject classifications. 35R35, 35B40, 80A25

1. Introduction.

In the present paper we study a simple free boundary model problem associated with solid combustion and some phase transition processes.

The model has been introduced by one of the authors in [1] as an ad hoc mathematical construction in an attempt to delineate the salient mechanism of thermal instability in solid state combustion. In the case of *one space dimension* (*only one-dimensional problems* are considered in this paper), the free boundary problem is posed for the heat equation in the semi-infinite domain

$$u_t = u_{xx}, \quad -\infty < x < s(t) \tag{1.1}$$

with the Stefan type conditions at the free boundary which in the simplest setup read

$$u|_{x=s(t)} = g_1(\dot{s}(t)), \quad (\partial u/\partial x)_{x=s(t)} = -\dot{s}(t), \tag{1.2}$$

where $s(t)$ is the position of the free boundary and $\dot{s}(t)$ is its velocity. We should point out right away a very important distinction from classical Stefan problems (see [2]). For an equilibrium Stefan problem the temperature at the free boundary is given as a constant, usually 0, or as a prescribed function of $s(t)$. In our opinion, this represents a serious deficiency of the Stefan problem model as applied to unstable phenomena. In many cases the simple boundary condition in (1.2) is more appropriate than the classical Stefan condition. In the context of solid combustion, the first boundary condition in (1.2) expresses the dependence of the propagation velocity on the temperature of the flame front due to the implicit presence of the second so-called deficient chemical reactant. In the context of solidification of overcooled liquids (see, e.g., [10]) or the amorphous→crystalline transition [8], the condition corresponds to the interface attachment kinetics that are due to various microscopic mechanisms responsible for the incorporation of the product at the interface into the crystalline lattice. In both cases the correct choice of the kinetic function g_1 provides the necessary feedback mechanism that enables the model to sustain global

* Received by the editors June 24, 1992; accepted for publication (in revised form) May 18, 1993.

† Department of Mathematical Sciences, Indiana University—Purdue University at Indianapolis, Indianapolis, Indiana 46205. The work of this author was partially supported by National Science Foundation grant DMS-9305228 and by Department of Energy grant DE-FG2-88ER1382.

‡ Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12180-3590 (roytbu@rpi.edu). The work of this author was partially supported by National Science Foundation grants DMS-8911888 and DMS-9311659.

in time and, it seems, even uniformly bounded evolution, no matter how complex the dynamical evolution may be. It turns out that the boundary condition in (1.2) is more natural for the rigorous mathematical analysis as well and leads to an easier integral equation than the classical Stefan condition (cf. §4).

We believe that, in spite of its simplicity, *the model captures essential features of dynamical behavior pertinent to more complicated and presumably more realistic mathematical models.* Abundant computational evidence in support of this belief is presented in [3]. It appears that the free boundary problem (1.1)–(1.2) is archetypical for the whole class of models with thermal instabilities. This motivates the thorough analytical investigation of the problem we undertake in this paper and its sequel [4]. We also note parenthetically that while introducing in [1] the model (1.1)–(1.2) as a pure mathematical "truncation" of a two-phase solid combustion model, the author was quite unaware that a similar free boundary problem had been introduced earlier to describe a real physical phenomenon—laser induced evaporation of solid materials [5].

A few remarks are in order to put the model in (1.1)–(1.2) in the context of standard models of mathematical combustion. It is well known that for certain ranges of parameters, uniformly traveling modes of flame propagations are unstable and undergo transition to self-oscillatory regimes. These transitions have been observed in experiments for both condensed phase and premixed gaseous fuels [6], [7]. A similar and physically closely related transition has been observed in the processes of solidification of thin film where a rapid solidification wave is initiated by a laser beam and in the laser-induced evaporation of solid materials (see [8], [5]). Auto-oscillatory and more complex (including chaotic) modes of propagation have also been demonstrated in numerical simulations on mathematical models of different degrees of complexity.

Probably the simplest example of a physical system of the kind just discussed is provided by gasless condensed phase combustion. For this type of combustion the solid fuel mixture is transformed directly into a solid product. In addition to its theoretical interest, gasless solid phase combustion currently finds technological applications as a method of synthesizing certain ceramics and metallic alloys [9].

The most primitive model of gasless combustion involves a system of differential equations for the temperature T and the limiting concentration of the fuel C (see [6]). In the one-dimensional formulation it takes the form

$$T_t = (\kappa T_x)_x + qW(C,T), \tag{1.3}$$

$$C_t = -W(C,T), \tag{1.4}$$

where κ is the thermal diffusivity, W is the chemical reaction rate, and q is the heat release.

For physically relevant values of parameters, the system is characterized by the strong temperature sensitivity of the rate and by rather sharply defined regions of dramatic change in the state variables that are usually associated with propagating fronts. It should be noted that as with the famous definition of pornography, although the fronts can be easily identified in the results of numerical simulations, their formal definition is rather vague. Numerical simulations on the models with *distributed* chemical kinetics require a sharp resolution of very thin reaction zones.

An attractive alternative to the models with distributed kinetics is provided by those with concentrated kinetics (so-called flame sheet approximation). The dis-

tributed reaction rate in (1.3)–(1.4) is replaced by the δ-function,

$$W = w(T)\delta(x - s(t)) \tag{1.5}$$

supported on the interface $x = s(t)$ between the fresh ($C = 1$) and burnt ($C = 0$) material (see, for example, [11]). The strength of the δ-function $w(T)$ is determined through an asymptotic analysis by matching relevant outer solutions. Of course, all the intricacies of the behavior in the reaction zone are lost in this approximation.

The system (1.3)–(1.4) with the δ-function source is easily transformed to the system of two heat equations coupled at the interface

$$\begin{aligned} &T_t^- = (\kappa T_x^-)_x, \qquad T_t^+ = (\kappa T_x^+)_x, \\ &T^-|_{x=s(t)} = T^+|_{x=s(t)}, \qquad (\kappa T_x^+ - \kappa T_x^-)_{x=s(t)} = -w(T)_{x=s(t)}, \\ &\frac{ds}{dt} = -w(T)|_{x=s(t)} \end{aligned} \tag{1.6}$$

where

$$\begin{aligned} T^-(x,t) &= T(x,t) \quad \text{for } x < s(t), \\ T^+(x,t) &= T(x,t) \quad \text{for } x > s(t). \end{aligned}$$

This is the free interface two-phase problem of gasless combustion. The heat conductivity coefficient is usually considered to be a constant. But, in principle, the heat conductivities of the fuel and of the product may be drastically different. For example, if the product is a foamlike material [12], then $\kappa_{product} \ll \kappa_{fuel}$. By setting $\kappa_{product} = 0$ in the equation and the boundary condition for T^+ in (1.6) we arrive at the model problem (1.1)–(1.2).

The principal result of the present paper is the global in time well-posedness of the free boundary problem (1.1)–(1.2). There is substantial literature treating, in particular, questions of well-posedness for combustion models with distributed kinetics (see, for example, [13] and [14]). It should be noted that corresponding problems for concentrated kinetics are in general harder. Indeed, while distributed kinetics problems contain nonlinearities of the type $f(u)$, the corresponding free boundary problems formulated in the front attached coordinate frame demonstrate nonlinearities of the type $u_x f(u(0,t))$. The two-phase solid combustion model has been considered by one of the authors [15]. More recently, Brauner et al. [16] have investigated a one-phase model of the classical Stefan type that arises from the consideration of very special perturbations in the equidiffusional model of gaseous combustion. Chow and Shen [17] have treated the relevant free boundary problem in a very general setting. All the aforementioned papers [15]–[17] employ similar techniques. First it is proved that the linearization about a traveling wave solution defines an analytic semigroup in an appropriate weighted Banach space. Then it is demonstrated that the nonlinear system defines a local nonlinear semigroup in an appropriate scale of interpolation spaces.

The single most important element allowing us to obtain global results of the paper is a *physically correct choice of conditions applied to the kinetic functions.* With this choice, the global well-posedness becomes physically feasible and therefore amenable to mathematical treatment. In view of relative simplicity of our problem we are able to establish the global well-posedness by rather elementary, classical methods. The rest of the paper is organized as follows. In §2 we collect the assumptions on kinetic functions and formulate the main result. In §3, by using an energy estimate for u and maximum principle type estimates for u and u_x we establish the crucial

a priori estimate. In §4 we derive an integral equation for the free boundary which, because of the nature of the boundary condition for u in (1.2), is much simpler than the analogous equation for the classical Stefan problem [18]. In §5 we prove that the integral operator is a contraction for small times and thus we establish the local existence and uniqueness for solutions of the integral equation. Again, the estimates are easier than for the classical case [18]. It is a simple matter to derive the global existence and uniqueness results from the a priori estimates and the local existence and uniqueness (§5.4). In §6 we conclude the proof of well-posedness by showing that solutions depend continuously on initial data.

2. Statement of the problem. We will be concerned with the following free boundary problem: Find $s(t)$ and $u(x,t)$ such that

$$u_{xx} = u_t \quad \text{for } x < s(t),\ s(0) = 0, \tag{2.1}$$

$$u(x,0) = u^0(x) \text{ for } x \leq 0, \quad \text{where } u^0(x) \geq 0, \tag{2.2}$$

$$u(s(t),t) = g_1(V(t)) \quad \text{for } t > 0, \tag{2.3}$$

$$u_x(s(t),t) = g_2(V(t)) \quad \text{for } t > 0, \tag{2.4}$$

where

$$V(t) = \frac{ds}{dt} \tag{2.5}$$

is the propagation velocity of the free boundary. In the context of solid fuel combustion, $s(t)$ represents the boundary between the unburnt and burnt material, and u is the nondimensionalized temperature. The temperature at the free boundary controls its velocity, $V(t) = g_1^{-1}(u(s(t),t))$. The model in (2.1)–(2.5) describes a one-phase system since we neglect the heat transfer in the burnt material. The heat exchange between the unburnt ($x < s(t)$) and burnt material is modeled by the boundary condition in (2.4) which, in principle, may be nonlinear.

We now introduce a set of rather general requirements on kinetic functions g_1 and g_2. We assume that

(A1) $g_1^{-1}(u)$ is a continuously differentiable, monotone decreasing, nonpositive function on $(0,\infty)$ with $g_1^{-1}(0) = v_0$ for some velocity $v_0 \leq 0$;

(A2) $g_2(V) > 0$ for $V \leq v_0$, g_2 is a continuously differentiable function on $(-\infty, 0)$;

(A3) there exists a limiting propagation velocity, $V_0 < 0$, such that

$$g_1^{-1}(u) > V_0 \quad \text{for any } u > 0.$$

The properties of the kinetic function g_1 listed in (A1), (A3) mimic those of the Arrhenius kinetics. The latter is usually written in the form $V \sim A\exp(-R/T)$. The assumption in (A2) is merely a generalization of the appropriate linear boundary condition that occurs in phase transition problems.

The main result of the paper is the following global existence and uniqueness theorem.

THEOREM 2.1. *Consider the problem in* (2.1)–(2.5). *Suppose that the kinetic functions* g_1 *and* g_2 *satisfy the assumptions in* (A1)–(A3), *and that the following conditions hold for the initial data* $u^0(x)$:

(1) $u^0(x)$ *has a continuous bounded derivative*;

(2) $u^0(x)$ *is integrable*;

(3) u_x^0 *approaches* 0 *at* $-\infty$ *and is consistent with boundary conditions at* $x = 0$, $u_x^0(0) = g_2 g_1^{-1} u^0(0)$.

Then there exists one and only one classical solution of the free boundary problem (2.1)–(2.5) *for all* $t > 0$.

(We say that $u(x,t), s(t)$ form a *classical solution* of (2.1)–(2.5) if (i) u_{xx} and u_t are continuous for $x < s(t), t > 0$; (ii) u and u_x are continuous for $x \leq s(t), t > 0$; (iii) u is continuous also for $t \geq 0$; (iv) $s(t)$ is continuously differentiable; (v) the equations in (2.1)–(2.5) are satisfied.)

The proof of Theorem 2.1 contains two major ingredients: a local existence theorem, and a priori estimated based mainly on the boundedness assumption (A3).

Remark. If the boundedness assumption in (A3) is dropped, one can obtain a local existence result with the existence interval determined by a norm of initial data.

3. A priori estimates. Results of this section are based on the maximum principle for the heat equation. We will always assume that (A1)–(A3) hold.

THEOREM 3.1. *Let* $u(x,t), s(t)$ *be a classical solution of the system* (2.1)–(2.5) *on the time interval* $0 < t < T_0$. *Then* $u(x,t) \geq 0$ *and* $V(t) < v_0$ *for* $0 < t < T_0$.

Remark. The theorem claims that the behavior of $u(x,t)$ is consistent with its interpretation as the temperature.

Proof. We will prove the theorem under an additional technical assumption that probably can be lifted. In the combustion context, it relates to the "cold boundary difficulty." We assume that

$$u^0(0) > 0. \tag{3.1}$$

Since $u(x,t)$ is a classical solution, it is continuous for $t \geq 0$. By continuity, $u(s(t),t) > 0$ for $0 \leq t < T$, and in this interval $V(t) < v_0$ (see (A1)). In a standard fashion define

$$T^* = \sup \{T | V(t) < v_0 \text{ for } 0 < t < T\}. \tag{3.2}$$

In the domain $D_{T^*} = \{(x,t) | x < s(t), 0 < t < T^*\}$ $u(x,t)$ is a solution of the initial value problem with the Dirichlet boundary conditions which is obtained from (2.1)–(2.5) by dropping the extra boundary condition (2.4). Both the initial conditions, $u^0(x)$, and the boundary conditions $u(s(t),t)$ are nonnegative. By the maximum principle,

$$u(x,t) \geq 0 \quad \text{in } \bar{D}_{T^*}. \tag{3.3}$$

We want to prove that $T^* = T_0$. Let $T^* < T_0$; then $V(T^*) = v_0$, and therefore $u(s(T^*), T^*) = 0$. On the other hand, $u_x(s(T^*), T^*) = g_2(v_0) > 0$ (see (A2)). This yields that $u(x, T^*) < u(s(T^*), T^*) = 0$ for some $x < s(T^*)$ which contradicts (3.3). □

Remark. The proof of Theorem 3.1 does not make use of the assumption in (A3). This boundedness assumption will be crucial for our next result.

THEOREM 3.2. *Let $u(x,t), s(t)$ be a classical solution of* (2.1)–(2.5). *Assume that the initial data are integrable,*

$$\int_{-\infty}^{0} u^0(x)dx = I_0 < \infty, \tag{3.4}$$

and that u_x^0 approaches 0 *at $-\infty$ and is consistent with the boundary conditions,*

$$u_x^0(0) = g_2 g_1^{-1} u^0(0). \tag{3.5}$$

Then the following estimates hold:

$$V_0 < V(t) < v_0, \tag{3.6}$$

$$U_1 \le u_x(x,t) \le U_2, \tag{3.7}$$

$$0 \le u(x,t) < \sqrt{2U_2(I_0 + U_2 t)}, \tag{3.8}$$

where

$$U_1 = \min\left(\inf_{x\le 0} u_x^0, \min_{V_0\le V\le v_0} g_2(V)\right), \qquad U_2 = \max\left(\sup_{x\le 0} u_x^0, \max_{V_0\le V\le v_0} g_2(V)\right). \tag{3.9}$$

Proof. The proof is based on the maximum principle and an energy type estimate.

The right inequality in (3.6) and the left one in (3.8) have been proved in Theorem 3.1, and in view of the inequality $g_1^{-1} > V_0$, the left inequality in (3.6) is a direct consequence of the boundary condition (2.3), which can be transformed to the form $V = g_1^{-1}(u(s(t),t))$. We turn now to the proof of the rest of the estimates.

It will be demonstrated later on that if u, s is a classical solution, then u may be represented in the form

$$\begin{aligned} u(x,t) = &\int_0^t \Big[G(t-\tau, x-s(\tau))\,\{u_x(s(\tau),\tau) + u(s(\tau),\tau)V(\tau)\} \\ &+ \frac{\partial G}{\partial x}(t-\tau, x-s(\tau))u(s(\tau),\tau)\Big]d\tau \\ &+ \int_{-\infty}^{0} G(x-\xi,t)u^0(\xi)d\xi, \end{aligned} \tag{3.10}$$

where G is the fundamental solution of the heat equation. It is clear from (3.10) that $w(x,t) = u_x(x,t)$ solves the heat equation for $t > 0$, $x < s(t)$. Also it is easy to check that $w(x,t) \to u_x^0(x)$ as $t \to 0$ for $x < 0$ (just from the integral representation (3.10) via integration by parts in the last term) and for $x = 0$ because of the consistency condition (3.2). Thus u_x is a solution of the problem

$$w_t = w_{xx},$$

$$w(s(t),t) = g_2(V(t)),$$

$$w(x,0) = u_x^0(x),$$

and the maximum principle for w yields the inequalities in (3.7).

Now consider the rate of change of the energy integral

$$\begin{aligned}\frac{d}{dt}I &= \frac{d}{dt}\int_{-\infty}^{s(t)} u(x,t)dx \\ &= \int_{-\infty}^{s(t)} u_t dx + V(t)u(s(t),t) \\ &= \int_{-\infty}^{s(t)} u_{xx} dx + V(t)u(s(t),t) \\ &= g_2(V(t)) + V(t)g_1(V(t)) \le g_2(V(t)) \le U_2,\end{aligned}$$

since $V \le 0$ and $g_1 \ge 0$. It follows that

$$I(t) \le I_0 + U_2 t. \tag{3.11}$$

On the other hand, $I(t)$ is estimated below as follows. Let $u^*(t) = \sup_{x \le s(t)} u(x,t)$ and x_1 such that $u(x_1,t) > u^* - \epsilon$. Since $u_x < U_2$,

$$u(x,t) \ge u(x_1,t) + U_2(x - x_1)$$

for

$$x_0 \le x \le x_1,$$

where

$$x_0 = x_1 - u(x_1,t)/U_2.$$

Therefore,

$$\begin{aligned}I(t) &> \int_{x_0}^{x_1} u(x,t)dx \ge \int_{x_0}^{x_1} [u(x_1,t) + U_2(x - x_1)]\, dx \\ &= (u^* - \epsilon)^2/2U_2.\end{aligned}$$

Thus,

$$\sup_{x \le s(t)} u(x,t) \le (2I(t)U_2)^{1/2} \le \sqrt{2U_2(I_0 + U_2 t)}. \qquad \square$$

4. The integral equation. Let G be the fundamental solution of the heat equation

$$G(x,t,\xi,\tau) = \exp\left\{-\frac{(x-\xi)^2}{4(t-\tau)}\right\}[4\pi(t-\tau)]^{-1/2}. \tag{4.1}$$

If $u(x,t), s(t)$ is a classical solution of (2.1)–(2.5), then by integrating Green's identity

$$\frac{\partial}{\partial \xi}\left(G\frac{\partial u}{\partial \xi} - u\frac{\partial G}{\partial \xi}\right) - \frac{\partial}{\partial \tau}(Gu) = 0 \tag{4.2}$$

over the domain $\xi < s(\tau)$, $0 < \tau < t$ and using the Stokes formula we obtain

$$
\begin{aligned}
u(x,t) &= \int_{\xi=s(\tau)} \left(G\frac{\partial u}{\partial \xi} - \frac{\partial G}{\partial \xi} u \right) d\tau + \int_{\xi=s(\tau)} G u d\xi + \int_{\tau=0} G u d\xi \\
&= \int_{\xi=s(\tau)} G\left[g_2(V) + g_1(V)V\right] d\tau - \int_{\xi=s(\tau)} \frac{\partial G}{\partial \xi} g_1(V) d\tau \\
&\quad + \int_{\tau=0} G u^0 d\xi.
\end{aligned} \tag{4.3}
$$

We note that in the limit $x \to s(t)-$ the first integral in (4.3) (with kernel G) is continuous while the second integral experiences a jump equal to $\frac{1}{2}g_1$ (this is a well-known jump property of the normal derivative of the heat potential; see Friedman [18]). Taking into account the boundary condition for u (2.3) we obtain in the limit the following integral equation for $V(t)$:

$$
\begin{aligned}
\frac{1}{2} g_1(V(t)) &= \int_0^t G(s(t), t, s(\tau), \tau) \left[g_2(V(\tau)) + g_1(V(\tau))V(\tau)\right] d\tau \\
&\quad - \int_0^t \frac{\partial G}{\partial \xi}(s(t), t, s(\tau), \tau) g_1(V(\tau)) d\tau \\
&\quad + \int_{-\infty}^0 G(s(t), t, \xi, 0) u^0(\xi) d\xi,
\end{aligned} \tag{4.4}
$$

where

$$
s(t) = \int_0^t V(\tau) d\tau. \tag{4.5}
$$

Thus, if u, s is a solution of the free boundary problem (2.1)–(2.5), then $V(t)$ solves the integral equation in (4.4) with $s(t)$ given by (4.5).

Conversely, let V be a continuous solution of (4.4). Then the integral representation in (4.3) defines a solution of the heat equation. It is easily seen from (4.3) that $u(x,t) \to u^0(x)$ as $t \to 0$. Also $u(x,t) \to g_1(V(t))$ as $x \to s(t)$ since V is a solution of the integral equation (4.4), which is defined by this limit. If we show that

$$
\lim_{x \to s(t)} u_x(x,t) = g_2(V(t)), \tag{4.6}
$$

then the proof of equivalence between the free boundary problem (2.1)–(2.5) and the integral equation in (4.4) will be complete.

By applying to u Green's formula we obtain the representation

$$
u(x,t) = \int_{\xi=s(\tau)} \left(G\frac{\partial u}{\partial \xi} - \frac{\partial G}{\partial \xi} g_1 + G g_1 V \right) d\tau + \int_{\tau=0} G u^0 d\xi, \tag{4.7}
$$

which differs from (4.3) only in the first integral. We subtract (4.3) from (4.7) to obtain

$$
\int_{\xi=s(\tau)} G\left(\frac{\partial u}{\partial \xi} - g_2(V) \right) d\tau = 0. \tag{4.8}
$$

We differentiate (4.8) with respect to x and use the above-mentioned jump property of the single-layer heat potential to get the integral equation for the difference $u_x(s(t),\tau) - g_2(V(\tau))$:

$$\frac{1}{2}[u_x(s(t),t) - g_2(V(t))] + \int_0^t G_\xi(s(t),t,s(\tau),\tau)\,[u_x - g_2(V(\tau))]\,d\tau = 0. \tag{4.9}$$

Since

$$|G_\xi(s(t),t,s(\tau),\tau)| < C/(t-\tau)^{1/2}$$

we see that the integral operator in (4.9) is a contraction (for small t) and the only solution is

$$u_x(x(\tau),\tau) - g_2(V(\tau)) \equiv 0.$$

Remark. If the method of this section is applied to the classical Stefan problem, then one arrives at a Fredholm integral equation of the first kind. In order to get a decent integral equation one should differentiate the analogue of (4.3) with respect to x and pass to the limit $x \to s(\tau)$, taking into account the jump property of parabolic potentials (cf. [18]). The resulting integral equation has integral kernels that are more singular than the kernels in (4.4). Consequently, some of the estimates in the next section are less involved than their counterparts for the classical Stefan problem.

5. Existence and uniqueness. In this section we show that a version of the integral equation (4.4) defines a contraction mapping for $0 < t < \sigma$ if σ is sufficiently small. The proof is based on rather coarse estimates. It follows very closely the argument of Friedman [18, pp. 506–511] but is less involved since our integral equation is simpler than the one in [18]. Then we use a priori estimates of §2 to establish global existence.

5.1. Modified integral equation. It is convenient to rewrite (4.4) in terms of the temperature at the front $\psi(t) = g_1(V(t))$:

$$\begin{aligned}\frac{1}{2}\psi(t) &= \int_0^t G(s(t),t,s(\tau),\tau)g_2 g_1^{-1}(\psi(\tau))d\tau \\ &\quad + \int_0^t G(s(t),t,s(\tau),\tau)\psi(\tau)g_1^{-1}\psi(\tau)d\tau \\ &\quad - \int_0^t \frac{\partial G}{\partial \xi}(s(t),t,s(\tau),\tau)\psi(\tau)d\tau \\ &\quad + \int_{-\infty}^0 G(s(t),t,\xi,0)u^0(\xi)d\xi,\end{aligned} \tag{5.1}$$

where

$$s(t) = \int_0^t g_1^{-1}(\psi(\tau))d\tau. \tag{5.2}$$

Let $K\psi$ denote the nonlinear integral operator on the right-hand side of (5.1). We will show that the transformation

$$w = 2Kv$$

is a contraction of an appropriate subset of $C[0,\sigma]$ for some σ and therefore has a unique fixed point $\psi = 2K\psi$.

Remark. Based on their physical interpretation, the kinetic functions g_1 and g_2 in (5.1)–(5.2) are defined for $V_0 < V < v_0$. It is not clear a priori why the integral operator K preserves the "physical" cone of positive temperatures. To avoid complications caused by using an iteration scheme in the cone $\{\psi \geq 0\}$, we extend functions g_1 and g_2 in (5.1)–(5.2) to the interval $(V_0, -V_0)$ (recall that $V_0 < 0$). We require the extension of g_1 to be monotone decreasing, the extension of g_2 to be positive, and both g_1^{-1} and g_2 to be twice differentiable with bounded first derivatives. We abuse the notation slightly and keep the same notation for the extensions.

The following simple proposition demonstrates that any solution with nonnegative initial data u^0 of the integral equation (5.1) *with extended kinetic functions* is positive. Therefore, for $u^0 \geq 0$ the fixed point of K *is positive* and is not affected by the choice of particular extensions of g_1 and g_2.

PROPOSITION 5.1. *Let g_1 and g_2 in the integral equation* (5.1) *be extensions of the original kinetic functions and let the initial data be nonnegative with $u^0(0) > 0$. Then any solution of* (5.1) *is positive.*

Proof. Let ψ be a solution of (5.1). As was demonstrated in §4 it corresponds to a solution of the free boundary problem (with extended kinetic functions). We note that specific monotonicity properties of g_1 and g_2 have not been used in §4.

To demonstrate that $V(t) < v_0$ and therefore that $\psi(t) = g_1(V(t)) > 0$, we observe that the proof of Theorem 3.1 is valid for the extended kinetic functions as well. Indeed, the only properties of kinetic functions used in the proof are:

$$g_1(V) > 0 \text{ for } V < v_0 \quad \text{and} \quad g_2(v_0) > 0.$$

They obviously hold for the extended kinetic functions and therefore Theorem 2 yields the desired result. □

5.2. A ball mapped into itself. In the Banach space $C_\sigma = C[0,\sigma]$ with uniform norm we consider the closed ball $B_{M,\sigma} = \{v \epsilon C_\sigma, \|v\| = \sup|v| \leq M\}$ with M to be specified later on. We will estimate separate terms in (5.1). The least trivial term is estimated as follows:

$$\left|\int_0^t \frac{\partial G}{\partial \xi}(s(t), t, s(\tau), \tau)\psi(\tau)d\tau\right| \tag{5.3}$$

$$= \frac{1}{2}\left|\int_0^t \frac{s(t) - s(\tau)}{t - \tau} G\psi(\tau)d\tau\right| \leq C_3|V_0|M\sqrt{t}$$

since

$$\left|\frac{s(t) - s(\tau)}{t - \tau}\right| \leq |V_0|$$

by (5.2) and since $|G| \leq C|t - \tau|^{-1/2}$. Other terms in (5.1) are estimated even easier. We note only that

$$|g_2 g^{-1}(\psi(t))| \leq V_2 = \max_{[V_0, -V_0]} g_2.$$

As the final estimate we obtain

$$\frac{1}{2}\|w\| \le C_1 V_2\sqrt{\sigma} + C_2|V_0|M\sqrt{\sigma} + C_3|V_0|M\sqrt{\sigma} + \|u^0\|, \tag{5.4}$$

where the constants C_1, C_2, C_3 are simple combinations of powers of 2 and π.

If M is now taken to be

$$M = 4\sup_{x\le 0}|u^0(x)|, \tag{5.5}$$

then for

$$\sqrt{\sigma} \le [C_1V_2 + (C_2+C_3)|V_0|M]^{-1}\, M/4 \tag{5.6}$$

the ball $B_{M,\sigma}$ is mapped by $2M$ into itself.

5.3. K is a contraction on $B_{M,\sigma}$. Let, $w = K\psi, w' = K\psi'$; then

$$\begin{aligned} w - w' &= \left[\int_0^t G g_2 g_1^{-1}(\psi)d\tau - \int_0^t G g_2 g_1^{-1}(\psi')d\tau\right] \\ &\quad + \left[\int_0^t G\psi g_1^{-1}(\psi)d\tau - \int_0^t G\psi' g_1^{-1}(\psi')d\tau\right] \\ &\quad + \left[-\int_0^t \frac{\partial G}{\partial \xi}\psi d\tau + \int_0^t \frac{\partial G}{\partial \xi}\psi' d\tau\right] \\ &\quad + \left[\int_{-\infty}^0 \{G(s,t,\xi,0) - G(s',t,\xi,0)\}\, u^0(\xi)d\xi\right] \\ &= W_1 + W_2 + W_3 + W_4. \end{aligned} \tag{5.7}$$

The estimations are quite elementary and are based on the mean value theorem. First we note that

$$\begin{aligned} |\Delta G| &\equiv |G(s(t),t,s(\tau),\tau) - G(s'(t),t,s'(\tau),\tau)| \\ &= |G(s(t)-s(\tau),t-\tau,0,0) - G(s'(t)-s'(\tau),t-\tau,0,0)| \\ &= |s(t)-s(\tau)-(s'(t)-s'(\tau))|\left|\frac{\partial G}{\partial x}(\tilde{s},t-\tau,0,0)\right| \\ &= \left|\frac{s(t)-s'(t)-(s(\tau)-s'(\tau))}{2(t-\tau)}\right| |\tilde{s}G(\tilde{s},t-\tau,0,0)| \\ &= \frac{1}{2}\left|\frac{ds}{dt}(\tilde{\tau}) - \frac{ds'}{dt}(\tilde{\tau})\right| |\tilde{s}G(\tilde{s},t-\tau,0,0)|, \end{aligned}$$

where $\tau \le \tilde{\tau} \le t$ and $\tilde{s}$ is between $s'(t)-s'(\tau)$ and $s(t)-s(\tau)$. From (5.2)

$$\begin{aligned} \left|\frac{ds}{dt}(\tilde{\tau}) - \frac{ds'}{dt}(\tilde{\tau})\right| &= |g_1^{-1}\psi(\tilde{\tau}) - g_1^{-1}\psi'(\tilde{\tau})| \\ &\le L_1\|\psi - \psi'\|, \end{aligned} \tag{5.8}$$

where L_1 is the Lipschitz constant for g_1^{-1}. Also

$$|\tilde{s}| \le \max\{|s'(t) - s'(\tau)|, |s(t) - s(\tau)|\} \le V_0(t-\tau).$$

Since $|G| \le C_0(t-\tau)^{-1/2}$ we get the estimate

$$|\Delta G| \le C_0 V_0 L_1 \|\psi - \psi'\|(t-\tau)^{1/2} = A_1\|\psi - \psi'\|(t-\tau)^{1/2}. \tag{5.9}$$

In a similar fashion we obtain that

$$\left|\Delta\frac{\partial G}{\partial \xi}\right| = \left|\frac{\partial G}{\partial \xi}(s(t), t, s(\tau), \tau) - \frac{\partial G}{\partial \xi}(s'(t), t, s'(\tau), \tau)\right| \tag{5.10}$$

$$\le A_2\|\psi - \psi'\|(t-\tau)^{-1/2}.$$

Now we are able to estimate the first three terms in (5.7):

$$|W_1| = \left|\int_0^t \Delta G g_2 g_1^{-1}(\psi)d\tau + \int_0^t G(s'(t), t, s'(\tau), \tau)[g_2 g_1^{-1}(\psi') - g_2 g_1^{-1}(\psi)]d\tau\right|$$

$$\le A_1\|\psi - \psi'\|t^{3/2}\frac{2}{3}V_2 + cL_2\|\psi - \psi'\|t^{1/2} \tag{5.11}$$

$$= (A_3 t^{3/2} + A_4 t^{1/2})\|\psi - \psi'\|,$$

where L_2 is the Lipschitz constant for $g_2 g_1^{-1}$. Similarly,

$$|W_2| \le A_1\|\psi - \psi'\|t^{3/2}\frac{2}{3}|V_0| + cL_1\|\psi - \psi'\|t^{1/2} \tag{5.12}$$

$$= (A_5 t^{3/2} + A_6 t^{1/2})\|\psi - \psi'\|,$$

and

$$|W_3| \le sA_2\|\psi - \psi'\|t^{1/2}\cdot M + \left|\int_0^t \frac{\partial G}{\partial \xi}(\psi - \psi')d\tau\right| \tag{5.13}$$

$$\le 2A_2\|\psi - \psi'\|t^{1/2}M + A_7 t^{1/2}\|\psi - \psi'\|$$

since

$$\left|\frac{\partial G}{\partial \xi}(s'(t), t, s'(\tau), \tau)\right| = \frac{1}{2}\left|\frac{s'(t) - s'(\tau)}{t-\tau}G\right|$$

$$\le \frac{1}{2}|V_0|\,|G| \le \frac{1}{2}V_0(t-\tau)^{-1/2}.$$

The estimation for the last term in (5.7) is a little different. Suppose $s(t) < s'(t) < 0$ and split the integral for W_4 into three integrals:

$$W_4 = \int_{-\infty}^{s} \delta G u^0 d\xi + \int_s^{s'} \delta G u^0 d\xi + \int_{s'}^{0} \delta G u^0 d\xi, \tag{5.14}$$

where

$$\delta G = G(s(t), t, \xi, 0) - G(s'(t), t, \xi, 0).$$

By the mean value theorem,

$$\delta G = (s - s')\frac{\partial G}{\partial x}(\tilde{s} - \xi, t, 0, 0) = (s - s')\frac{\tilde{s} - \xi}{2t} G(\tilde{s} - \xi, t, 0, 0),$$

where

$$\tilde{s} = \tilde{s}(t, \xi), \quad s(t) \le \tilde{s} \le s'(t).$$

If $\xi < s < \tilde{s} < s'$, then

$$\begin{aligned}(\tilde{s} - \xi)G(\tilde{s} - \xi, t, 0, 0) &= c_0(\tilde{s} - \xi)t^{-1/2}e^{-(\tilde{s}-\xi)^2/4t} \\ &= c_0(8t)^{1/2}\frac{\tilde{s} - \xi}{(8t)^{1/2}}e^{-(s-\xi)^2/8t}2^{1/2}(2t)^{-1/2}e^{-(\tilde{s}-\xi)^2/8t} \qquad (5.15) \\ &\le 4t^{1/2}c_1 G(\tilde{s} - \xi, 2t, 0, 0) \le 4c_1 t^{1/2} G(s - \xi, 2t, 0, 0),\end{aligned}$$

where $c_1 = \max(xe^{-x^2})$. Thus

$$\begin{aligned}\left|\int_{-\infty}^{s} \delta G u^0 d\xi\right| &\le \frac{s - s'}{2t} 4c_1 t^{1/2} \int_{-\infty}^{s} G(s - \xi, 2t, 0, 0)|u^0(\xi)|d\xi \\ &\le 2c_1 t^{1/2}\frac{1}{t}\int_0^t \left[g_1^{-1}\psi - g_1^{-1}\psi'\right] d\tau \int_{-\infty}^{0} G(s - \xi, 2t, 0, 0)|u^0(\xi)|d\xi \qquad (5.16) \\ &\le 2c_1 L_1 t^{1/2} \sup|u^0(\xi)|\, \|\psi - \psi'\|.\end{aligned}$$

The second integral in (5.14) is simpler:

$$\begin{aligned}\int_s^{s'} \delta G u^0 d\xi &\le (s' - s)2\sup G \cdot \|u^0\| \\ &= (s' - s)2c_0 t^{-1/2}\|u^0\| \\ &\le 2c_0\|\psi - \psi'\|t^{1/2}\|u^0\|.\end{aligned}$$

The integral over $(s', 0)$ is estimated similarly to (5.16). Finally, by combining the preceding estimates we get

$$|W_4| \le A_8 t^{1/2}\|\psi - \psi'\|\, \|u^0\|. \qquad (5.17)$$

The same estimate holds if $s < 0 < s'$ or $0 < s < s'$ (recall that we do not assume a priori that $s(t) < 0$).

The results in (5.11)–(5.13) and (5.17) yield the following contraction estimate:

$$\begin{aligned}\|K\psi - K\psi'\| \le{}& A_9\sigma^{3/2}\|\psi - \psi'\| + A_{10}\sigma^{1/2}\|\psi - \psi'\| \qquad (5.18) \\ &+ A_{11}\sigma^{1/2}\|\psi - \psi'\|M + A_{12}\sigma^{1/2}\|\psi - \psi'\|\, \|u^0\|,\end{aligned}$$

where all the constants A_9, A_{10}, etc. (as well as all the A constants introduced previously) depend only on bounds and Lipschitz constants of g_1^{-1} and $g_2 g_1^{-1}$. If $\sigma < 1$ and such that

$$\sigma^{1/2}\left[A_9 + A_{10} + M(A_{11} + A_{12})\right] < 1/2, \tag{5.19}$$

then $2K$ is a contraction on $B_{M,\sigma}$. Therefore, it has a fixed point $\psi(t)$ in $B_{M\sigma}$, which is unique. It is easy to show (see Friedman [18]) that any solution of the integral equation in (5.1), regardless of whether it is bounded by M or not, must coincide with ψ in their common interval of existence.

5.4. Global existence. We have proved the existence and uniqueness of a solution $\psi(t)$ of (5.1) for $0 \le t < \sigma$. If $t_1 < \sigma$, then by repeating the argument of §§5.2–5.3 with the initial data $u^0(\xi) = u(\xi, t_1)$ (where u is computed according to (4.3) with $V = g_1^{-1}(\psi)$) we can construct a solution of the integral equation for $t_1 \le t \le t_2$. Note that the equation should be modified in the obvious way: the spatial integration must be taken from $-\infty$ to $s(t_1)$.

Both solutions for $0 \le t \le t_1$ and for $t_1 \le t \le t_2$ generate solutions of the free boundary problem (2.1)–(2.5) on these intervals. The solution on $[0, t_1]$ is actually extendable to $[0, t_1+\epsilon], t_1+\epsilon < \sigma$, and because of uniqueness the solutions on $[0, t_1+\epsilon]$ and $[t_1, t_2]$ coincide in the overlapping interval. Therefore, the extended solution is differentiable with respect to t across the line $t = t_1$. Thus, we obtain a unique solution defined for $0 \le t \le t_2$. This process can be repeated indefinitely.

To establish the global existence stated in Theorem 2.1, we need to show that if $u(x,t), V(t)$ is a classical solution of (2.1)–(2.5), which exists and is unique for $0 \le t < t_0$, then it exists and is unique for $0 \le t < t_0 + \epsilon$ for some $\epsilon > 0$.

From Theorem 3.2, (3.8) it follows that

$$0 \le u(\xi, t_0) \le \sqrt{2U_2(I + U_2 t_0)}.$$

We select M in (5.5) to be

$$M = 1 + \sqrt{2U_2(I + U_2 t_0)};$$

then if σ satisfies the inequalities in (5.6) and (5.19), the solution of the integral equation (5.1) exists and is unique for $t_0 < t < t_0 + \sigma$. The previous argument shows that solutions for $0 < t < t_1 = t_0 - \delta$ and for $t_1 \le t \le t_2 = t_1 + \sigma$, with $\delta < \sigma/2$, agree and form a solution for $0 \le t < t_0 + \epsilon, \quad \epsilon = \sigma - \delta$. This concludes the proof of Theorem 2.1.

6. Continuous dependence on initial conditions. By employing estimates similar to those in §5, it is not hard to prove that solutions of the free boundary problem (2.1)–(2.5) depend continuously on initial data. The proof is relatively routine, and we include it only for completeness.

THEOREM 6.1. *Let S be the class of initial data satisfying conditions* (1)–(3) *of the existence and uniqueness theorem (Theorem* 2.1). *For any $u^0 \in S$, there exists $\sigma > 0$ (which depends only on the uniform norm $\|u^0\|$) so that solutions of the problem* (2.1)–(2.5) *depend on initial conditions continuously at u^0. More precisely, if $\{u(x,t), s(t)\}$, $\{\tilde{u}(x,t), \tilde{s}(t)\}$, $0 < t < \sigma$ are solutions of the problem* (2.1)–(2.5) *with initial data $u^0, \tilde{u}^0 \in S$, then for $x < 0, 0 < t < \sigma$*

$$|V(t) - \tilde{V}(t)| < c\|u^0 - \tilde{u}^0\|, \tag{6.1}$$

$$|u(x-s(t),t)-\tilde{u}(x-\tilde{s}(t),t)|<c\|u^0-\tilde{u}^0\|. \tag{6.2}$$

Remark. We state and prove continuous dependence on initial conditions only locally in time. The argument extending this result to any fixed time follows closely the proof of global existence in §5.4.

Proof. We will establish first the estimate in (6.1) and then use it to derive (6.2). Let ψ and $\tilde{\psi}$ be solutions of the integral equation in (5.1) with initial data u^0 and $\tilde{u}^0$, respectively. Since $\psi = 2K\psi$, $\tilde{\psi} = 2K\tilde{\psi}$ where K is the integral operator, we have

$$\begin{aligned}\frac{1}{2}(\psi-\tilde{\psi}) &= \left[\int_0^t Gg_2g_1^{-1}(\psi)d\tau - \int_0^t Gg_2g_1^{-1}(\tilde{\psi})d\tau\right]\\ &+\left[\int_0^t G\psi g_1^{-1}(\psi)d\tau - \int_0^t G\tilde{\psi}g_1^{-1}(\tilde{\psi})d\tau\right]\\ &+\left[-\int_0^t \frac{\partial G}{\partial\xi}\psi d\tau + \int_0^t \frac{\partial G}{\partial\xi}\tilde{\psi}d\tau\right]\\ &+\left[\int_{-\infty}^0 \{G(s,t,\xi,0)-G(\tilde{s},t,\xi,0)\}u^0(\xi)d\xi\right]\\ &+\int_{-\infty}^0 G(\tilde{s},t,\xi,0)\{u^0(\xi)-\tilde{u}(\xi)\}d\xi.\end{aligned} \tag{6.3}$$

This expression differs from (5.7) by the last integral only. By literally repeating estimates in §5.3 we obtain the following analogue of (5.18):

$$\begin{aligned}\frac{1}{2}\|\psi-\tilde{\psi}\| &\le A_9\sigma^{3/2}\|\psi-\tilde{\psi}\| + A_{10}\sigma^{1/2}\|\psi-\tilde{\psi}\|\\ &+A_{11}\sigma^{1/2}\|\psi-\tilde{\psi}\|M + A_{12}\sigma^{1/2}\|\psi-\tilde{\psi}\|\,\|u^0\| + \|u^0-\tilde{u}^0\|,\end{aligned} \tag{6.4}$$

where A_9, A_{10}, etc. depend only on bounds and Lipschitz constants of g_1^{-1} and $g_2g_1^{-1}$. The constant M may be taken larger than the a priori bounds for u and $\tilde{u}$ in (3.8) with $t=\sigma$. By now taking σ so small that the coefficient of $\|\psi-\tilde{\psi}\|$ in (5.4) is smaller than $\frac{1}{4}$, we get

$$\|\psi-\tilde{\psi}\|<4\|u^0-\tilde{u}^0\|. \tag{6.5}$$

Since $V=g_1^{-1}(\psi)$, (6.5) yields (6.1):

$$\|V-\tilde{V}\|<C\|u^0-\tilde{u}^0\|$$

for $0<t<\sigma$, where C depends only on the Lipschitz constant of g_1^{-1}.

To obtain the estimate in (6.2) we note that $u(x,t)$ and $\tilde{u}(x,t)$ solve the heat equation in the domains $\{x<s(t), 0<t<\sigma\}$ and $\{x<\tilde{s}(t), 0<t<\sigma\}$ with initial and boundary conditions

$$u(x,0)=u^0(x), \quad u(t,s(t))=\psi(t); \tag{6.6}$$

$$\tilde{u}(x,0)=\tilde{u}^0(x), \quad \tilde{u}(t,\tilde{s}(t))=\tilde{\psi}(t). \tag{6.7}$$

Let $\tilde{w}(x,t)$ be a solution of the auxiliary problem

$$\tilde{w}_t - \tilde{w}_{xx} = 0, \quad 0 < t < \sigma, \quad x < \tilde{s}(t), \tag{6.8}$$

$$\tilde{w}(x,0) = u^0(x), \quad \tilde{w}(\tilde{s}(t),t) = \psi(t).$$

Then, by the maximum principle,

$$|\tilde{u}(x,t) - \tilde{w}(x,t)| \le \max(\|u^0 - \tilde{u}^0\|, \|\psi - \tilde{\psi}\|) \tag{6.9}$$

$$\le C\|u^0 - \tilde{u}^0\|.$$

Now, through the change of variables $w(x,t) = u(x - (s(t) - \tilde{s}(t)), t)$ the domain $\{x < s(t), 0 < t < \sigma\}$ is transformed into the domain $\{x < \tilde{s}(t), 0 < t, < \sigma\}$ where w solves the problem

$$w_t - w_{xx} = \left[V(t) - \tilde{V}(t)\right] w_x, \tag{6.10}$$

$$w(x,0) = u^0(x), \quad w(\tilde{s}(t),t) = \psi(t).$$

Thus, the difference $W = w - \tilde{w}$ solves the equation

$$W_t - W_{xx} = \left(V - \tilde{V}\right) w_x \tag{6.11}$$

with zero initial and boundary conditions. If Γ is the Green's function of the heat operator in the domain $\{x < \tilde{s}(t), \quad 0 < t < \sigma\}$ with zero boundary conditions, then

$$W(x,t) = \int_0^t \int_{-\infty}^{\tilde{s}(t)} \Gamma(x,\xi,t,\tau)(V - \tilde{V}) w_x d\xi d\tau. \tag{6.12}$$

By the maximum principle,

$$\left| \int_{-\infty}^{s(t)} \Gamma(x,\xi,t,\tau) \left[V_\tau - \tilde{V}(\tau)\right] w_x(\xi,\tau) d\xi \right| \tag{6.13}$$

$$\le \|V - \tilde{V}\| \sup_{\xi} |w_x(\xi,\tau)|.$$

And since $w_x = u_x$, we can use the a priori estimate (3.8), $|u_x| \le U_j$, to continue the estimate (6.13) as follows:

$$\le \|V - \tilde{V}\| U_j.$$

Upon integration of (6.13) with respect to τ we obtain the estimate

$$|W(x,t)| \le \sigma \|V - \tilde{V}\| \cdot U_j, \tag{6.14}$$

with U_j determined by $\|u_x^0\|$.

We combine the estimates in (6.9), (6.14):

$$|\tilde{u}(x,t) - u(x - [s(t) - \tilde{s}(t)], t)|$$

$$= |\tilde{u}(x,t) - w(x,t)| \le |\tilde{u}(x,t) - \tilde{w}(x,t)| + |\tilde{w}(x,t) - w(x,t)|$$

$$= |\tilde{u} - \tilde{w}| + |W| \le C\|u^0 - \tilde{u}^0\| + \sigma\|V - \tilde{V}\| U_j$$

$$\le C\|u^0 - \tilde{u}^0\|,$$

which becomes the estimate in (6.2) via a change of variables $x \to x - \tilde{s}(t)$. □

7. Concluding remarks. This paper is a part of a broader project aimed at analytical and numerical investigation of the free boundary model of thermal instabilities. We are convinced that this one-phase model is generic in the sense that it captures the principal nonlinearity responsible for the rich dynamical behavior of many systems with thermal instabilities.

Mathematically the model is so transparent and simple that it lends itself to a pure elementary treatment: a substantial part of the technical apparatus developed for two-phase models [15], [17] proves to be unnecessary. We have derived a simple uniform estimate on the growth of solution ($\sim \sqrt{t}$) and proved the existence and uniqueness of classical solutions globally in time for any physically meaningful initial data. These results provide a firm theoretical ground for numerical simulations on the model.

It should be noted, however, that some conditions in (A1)–(A3) imposed on the kinetic functions may be too strong. Our numerical experiments with various kinetics indicate that the global existence, perhaps in a more restricted sense, should be valid for a broader class of kinetic functions. It is unclear whether the very direct method of the present paper works for this broader class of kinetics.

Another remark concerns two-phase problems. We believe that the global in time existence of classical solutions can be established in this case as well. However, there is a difficulty here related to the fact that the "latent heat" of the transition diffuses into both phases. Apparently, for solutions of the resulting system of two partial differential equations there is no a priori estimate based on the maximum principle. The problem requires a more subtle treatment which, at the moment, we are unable to carry out.

The sequel of this paper [4] is devoted to the dynamic study of the model. We investigate traveling wave solutions, their stability and bifurcations.

REFERENCES

[1] M. L. FRANKEL, *Free boundary problems and dynamical geometry associated with flames*, in Dynamical Issues in Combustion Theory, P. Fife, A. Liñán, and F. Williams, eds., IMA Volumes in Mathematics and Its Applications, 35 (1991), pp. 107–127.

[2] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[3] M. FRANKEL, V. ROYTBURD, AND G. SIVASHINSKY, *A sequence of period doublings and chaotic pulsations in a free boundary problem modeling thermal instabilities*, SIAM J. Appl. Math., 54 (1994).

[4] M. FRANKEL AND V. ROYTBURD, *Stability and bifurcations for a model of thermal instabilities*, J. Dynamics Differential Equations, submitted.

[5] S. M. GOL'BERG AND M. I. TRIBELSKII, *On laser induced evaporation of nonlinear absorbing media*, Zh. Tekh. Fiz. (Sov. Phys.-J. Tech. Phys.), 55 (1985), pp. 848-857. (In Russian.)

[6] K. G. SHKADINSKY, B. I. KHAIKIN, AND A. G. MERZHANOV, *Propagation of a Pulsating Exothermic Reaction Front in the Condensed Phase*, Combust. Expl. Shock Waves, 7 (1971), pp. 15–22.

[7] A. G. MERZHANOV, A. K. FILONENKO, AND I. P. BOROVINSKAYA, *New Phenomena in Combustion of Condensed Systems*, Dokl. Akad. Nauk USSR, 208 (1973), pp. 892–894; Soviet Phys. Dokl., 208 (1973), pp. 122–125.

[8] W. VAN SAARLOS AND J. WEEKS, *Surface undulations in explosive crystallization: a nonlinear analysis of a thermal instability*, Physica D, 12 (1984), pp. 279–294.

[9] A. G. MERZHANOV, SHS *Processes: Combustion theory and practice*, Arch. Combustionis, 1 (1981), pp. 23–28.

[10] M. FRANKEL, *On the nonlinear evolution of a solid-liquid interface*, Physics Letters A, 128 (1988), pp. 57–60.

[11] B. J. MATKOWSKY AND G. I. SIVASHINSKY, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Appl. Math, 35 (1978), pp. 230–255.

[12] G. I. SIVASHINSKY, private communication, 1991.
[13] B. LARROUTUROU, *The equations of one-dimensional unsteady flame propagation: existence and uniqueness*, SIAM. J. Math. Anal., 19 (1988), pp. 32–59.
[14] J. BEBERNES AND D. EBERLEY, *Mathematical problems from combustion theory*, Springer-Verlag, New York, 1989.
[15] V. ROYTBURD, *A Hopf bifurcation for a reaction-diffusion equation with concentrated chemical kinetics*, J. Differential Equations, 56 (1985), pp. 40–62.
[16] C. M. BRAUNER, S. NOOR EBAD, AND CL. SCHMIDT-LAINE, *Nonlinear stability analysis of singular traveling waves in combustion—a one-phase problem*, Nonlinear Analysis Theory Meth. Appl., 16 (1991), pp. 881–892.
[17] S.-N. CHOW AND W. SHEN, *A free boundary problem related to condensed two-phase combustion*, preprint, 1991.
[18] A. FRIEDMAN, *Free boundary problems for parabolic equations* I. *Melting of solids*, J. Math. Mech., 8 (1959), pp. 161–184.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1375–1392, September 1994

007

ON A NONLINEAR INTEGRODIFFERENTIAL DRIFT-DIFFUSION SEMICONDUCTOR MODEL*

JIN LIANG†

Abstract. The author considers a drift-diffusion semiconductor system coming from a model of n-GaAs related to the Gunn effect, which can be transformed into a nonlinear integrodifferential equation with integral boundary condition. The global existence, uniqueness, and regularity of the solution are obtained.

Key words. drift-diffusion, semiconductor, nonlinear integrodifferential equations, integral boundary condition, global existence, regularity

AMS subject classifications. 35K60, 35M99, 35Q60

1. Introduction. In 1963, Gunn reported the discovery of current oscillation at low microwave frequencies which were produced when the semiconductor n-GaAs was subjected to an electric field of a few kV/cm [7]. This "Gunn effect" could be tracked to the fact that the velocity of electrons in n-GaAs, $\widetilde{V}(\widetilde{E})$, had a maximum as a function of the local electric field. The subsequent region of negative slope (negative differential mobility) of $\widetilde{V}(\widetilde{E})$ causes the Gunn instability [12], [13]. $\widetilde{V}(\widetilde{E})$ is described in Remark#. As indicated in [2], [3], and [4] (see also [9], [11]–[13]), the two equations governing charge transport in n-GaAs are the following:

- Poisson's law for the electric field, $\widetilde{E}(\xi,\tau)$,

$$\widetilde{E}_\xi = e(N - N_0)/\epsilon; \tag{1.1}$$

- the continuity equation for the electron concentration, $N(\xi,\tau)$,

$$N_\tau + [\widetilde{V}(\widetilde{E})N - DN_\xi]_\xi = 0. \tag{1.2}$$

Here $-e$ is electron charge, ϵ is the permittivity of the semiconductor, D is the diffusion constant, and N_0 is the concentration of donor impurities that we assume to be uniform. These equations are a reduced form of the well-known drift-diffusion equations [9], where transport by the holes is neglected.

We eliminate N by substituting (1.1) to (1.2) and then integrate the result with respect to ξ,

$$\epsilon\widetilde{E}_\tau + (eN_0 + \epsilon\widetilde{E}_\xi)\widetilde{V}(\widetilde{E}) - D\epsilon\widetilde{E}_{\xi\xi} = J_{tot}(\tau). \tag{1.3}$$

This is Ampère's law, the sum of the displacement current and the electron current at a point of the semiconductor is equal to the total current, $J_{tot}(\tau)$ (the integration constant). The bias determines $J_{tot}(\tau)$. For a purely resistive external circuit, we have

$$\int_0^L \widetilde{E}(\xi,\tau)d\xi + J_{tot}(\tau)\widetilde{R} = \widetilde{\Phi}, \tag{1.4}$$

* Received by the editors October 8, 1992; accepted for publication (in revised form) April 8, 1993.

† Department of Mathematics, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom. The work of this author was supported by Fundação Oriente in Portugal during a stay in the University of Lisbon.

where $\widetilde{\Phi}$ is the voltage and the constant $\widetilde{R}$ is proportional to the resistance. For voltage bias, $\widetilde{R} = 0$ in (1.4).

In ideal contacts, the resistivity is zero and Ohm's law implies Dirichlet boundary condition as follows:

$$\widetilde{E}(0,\tau) = 0, \quad \widetilde{E}(L,\tau) = 0. \tag{1.5}$$

But more realistically, on the boundary, the displacement current plus $J_{0C}(\widetilde{E})$ equals the total current, where $J_{0C}(\widetilde{E}) = \rho_{0C}^{-1}\widetilde{E}$ for a purely Ohmic contact with resistivity ρ_{0C}. That is, we have the boundary condition (see also [2]–[4] and [13])

$$\epsilon\widetilde{E}_\tau + J_{0C}(\widetilde{E}) = J_{tot}(\tau). \tag{1.6}$$

By changing to dimensionless variables, we can transform problems (1.3), (1.4), and (1.6) into the following form (see also [2]–[4]):

$$\begin{cases} E_t + V(E)(1+E_x) - \delta E_{xx} = J(t), & \text{in } Q_T = (0,l)\times(0,T), \\ \displaystyle\int_0^l E(x,t)\,dx + RJ(t) = \Phi(t), & \text{in } [0,T], \\ E(x,0) = E_0(x), & \text{on } [0,l], \\ \displaystyle E_t(0,t) + \frac{E(0,t)}{\rho_0} = J(t), & \text{on } [0,T], \\ \displaystyle E_t(l,t) + \frac{E(l,t)}{\rho_l} = J(t), & \text{on } [0,T], \end{cases} \tag{1.7}$$

where E and J are unknowns related to electrical field and total current, respectively, $V(\cdot)$ is a smooth function with linear growth, Φ is a given function related to the voltage, and $R \geq 0$, $\delta, l, T, \rho_0, \rho_l > 0$ are given constants.

Remark$^{\#}$. In semiconductor theory, $V(E)$ is related to the drift velocity of electrons in n-GaAs versus electric field E. It satisfies $V(0) = 0$, $V'(0) > 0$, and $V(E)$ has at most one maximum when $E > 0$, and when $E \to \infty$, $V(E)$ grows at most linearly. Two examples of $V(E)$ are $V(E) = \mu_0 E((1+BE^4)/(1+E^4))$ and $V(E) = \mu_0 E((1+BE^3)/(1+E^4))$, where μ_0 and $B < 1$ are positive constants, (see [2]–[4]). In this paper, we consider a more general class of functions V (see assumption (A7) below).

Problem (1.7) is a nonlinear parabolic differential equation coupled to an integral equation with nonstandard differential boundary conditions. This problem can be written as an integrodifferential problem (see §2). In this way, we can change the system into an integrodifferential equation so that we can get help from the linear integrodifferential theory for the Dirichlet problem. We also can write the boundary conditions in the form of integral equation (see §3). It is then possible for us to use the results for the Dirichlet problem (in §2) and a suitable map to discuss the original problem (1.7). That is, the problem is equivalent to a nonlinear problem consisting of integrodifferential equation with an integral boundary condition, which will be discussed in the following.

In this paper, we consider global existence, uniqueness, regularity, and continuous dependence with respect to given data of the solution for problem (1.7). To the best of our knowledge, we obtain the first such results for this model, which was already known as early as the 1960s (see [7] and [13]).

In order to discuss problem (1.7), we will first consider an integrodifferential problem with a Dirichlet boundary condition, which is also of interest because of (1.5) and physical relevance. The main difficulty is to obtain estimates for $\|E\|_{C^0}$ or $\|E\|_{L_p}$. In [5], the authors have discussed a class of semilinear integrodifferential problems whose integral term also grows linearly. They have obtained a priori estimates in $L_\infty\big([0,T];W^2_\infty(\Omega)\big)$ by using a Green's representation, but their method is not well fitted to our nonlinear case. We therefore have to find a new way of obtaining estimates. The cases of $R>0$ and $R=0$ will be treated separately. For $R>0$, we can directly use maximum principle techniques, but unfortunately all estimates will depend on R. It is not clear how to pass the limit as $R\to 0$ and to get the results for the case of $R=0$ from $R>0$, so we consider the case of $R=0$ by a different method. This case is technically more difficult, since after the transformation, the integral term contains nonlinear high-order differential terms and the maximum principle cannot be used. To overcome the difficulties, we use some special test functions to get the necessary estimates. Then, with the results for the Dirichlet problem, we can define a map to consider the original problem (1.7). We also discuss it in both cases of $R>0$ and $R=0$. For the case of $R=0$, we overcome the difficulties of obtaining an a priori estimate and the compactness of the map by making a suitable transformation.

The outline of this paper is as follows. In §2, as a preliminary step in investigating the original problem (1.7), we discuss (1.7) with Dirichlet boundary conditions. Existence, uniqueness, and regularity of the solution are obtained, as well as continuous dependence of the solution on the given data, Φ, E_0, and boundary conditions. The main results of the existence, uniqueness, and regularity of the solution for the original problem (1.7) are obtained in §3 from the results on the preliminary problems considered in §2.

Throughout, the standard space notations, such as $C^k(I)$, $C^{k+\alpha}(I)$, $C^{k,l}(Q_T)$, $C^{k+\alpha,\frac{k+\alpha}{2}}(Q_T)$, $W_p^{k,l}(Q_T)$, and $L_p(0,T;B(I))$ are used (see [8]), and $C^\infty(\Omega)=\{u: u\in C^k(\Omega)$, for all $k\geq 0\}$.

2. Dirichlet integrodifferential problem. In preparation for investigating problem (1.7), in this section, we consider existence, uniqueness, regularity, and continuous dependence with respect to given data of the solution for the following two systems with Dirichlet boundary conditions:

$$\text{(2.1)}\qquad \begin{cases} E_t+V(E)(1+E_x)-\delta E_{xx}=J(t),\\ \displaystyle\int_0^l E(x,t)dx+RJ(t)=\Phi(t),\\ E(x,0)=E_0(x),\quad E(0,t)=E_1(t),\quad E(l,t)=E_2(t), \end{cases}$$

$$\text{(2.2)}\qquad \begin{cases} \displaystyle E_t+V\left(E+\frac{1}{l}\left(\Phi(t)-\int_0^l E(x,t)dx\right)\right)(1+E_x)-\delta E_{xx}=0,\\ E(x,0)=E_0(x),\qquad E(0,t)=E_1(t),\quad E(l,t)=E_2(t). \end{cases}$$

Incidentally, problem (2.1) also comes from semiconductor theory (see §1 or [2]–[4] and [13] for more details).

In these two problems, the given functions $\Phi(t)$, $E_i(t)$, $(i=1,2)$, and $E_0(x)$ are assumed to satisfy one of (A1)–(A3) in the following:

(A1) $E_i(t)\in C^1([0,T])$, $E_0(x)\in C^2([0,l])$, $\Phi(t)\in C^1([0,T])$; or

(A2) $E_i(t)\in C^{1+\frac{\alpha}{2}}([0,T])$, $E_0(x)\in C^{2+\alpha}([0,l])$, $\Phi(t)\in C^{1+\frac{\alpha}{2}}([0,T])$; or

(A3) $E_i(t) \in C^\infty([0,T]), \quad E_0(x) \in C^\infty([0,l]), \quad \Phi(t) \in C^\infty([0,T]).$

We also assume they satisfy the following compatibility assumptions:

(A4) $$\int_0^l E_0(x)dx = \Phi(0),$$

and

(A5) $$E_1(0) = E_0(0), \qquad E_2(0) = E_0(l).$$

And for problem (2.2) and problem (2.1) when $R > 0$,

(A6) $$\begin{cases} E_1'(0) + V\big(E_0(0)\big)\big(1 + E_{0x}(0)\big) - \delta E_{0xx}(0) = 0, \\ E_2'(0) + V\big(E_0(l)\big)\big(1 + E_{0x}(l)\big) - \delta E_{0xx}(l) = 0; \end{cases}$$

or for problem (2.1) when $R = 0$,

(A6′) $$\begin{cases} E_1'(0) + V\big(E_0(0)\big)\big(1 + E_{0x}(0)\big) - \delta E_{0xx}(0) \\ = \dfrac{1}{l}\left(\Phi'(0) + \displaystyle\int_0^l V(E_0)(1 + E_{0x})dx - \delta \int_0^l E_{0xx}dx\right), \\ E_2'(0) + V\big(E_0(l)\big)\big(1 + E_{0x}(l)\big) - \delta E_{0xx}(l) \\ = \dfrac{1}{l}\left(\Phi'(0) + \displaystyle\int_0^l V(E_0)(1 + E_{0x})dx - \delta \int_0^l E_{0xx}dx\right); \end{cases}$$

The function $V(E)$, which is related to the nonlinear term, is assumed to satisfy

(A7) $V(\cdot)$ is a smooth function satisfying $|V(E)| \le C_v|E|$, where C_v is a positive constant.

Problem (2.1) can be changed into an integrodifferential problem according to the different cases $R > 0$ or $R = 0$. In these two cases, we obtain nonlinear integrodifferential equations in different forms. We will consider them and obtain existence of the solutions in the following two sections. For problem (2.2), we use a similar argument, which is discussed in §2.3. The further properties of uniqueness and continuous dependence on the given data of the solutions are considered in the last section.

2.1. Problem (2.1): The existence for $R > 0$. In this case, we can simply replace $J(t)$ by the second equation in (2.1) and get the following equivalent form of problem (2.1) when $R > 0$,

(2.3) $$\begin{cases} E_t + V(E)(1 + E_x) - \delta E_{xx} = \dfrac{1}{R}\left(\Phi(t) - \displaystyle\int_0^l E dx\right), \\ E(x,0) = E_0(x), \quad E(0,t) = E_1(t), \quad E(l,t) = E_2(t). \end{cases}$$

In order to obtain the existence of the solution, we first obtain a priori estimates. We suppose $\Phi(t)$, $E_i(t)$, $(i = 1,2)$, and $E_0(x)$ satisfy (A5) and one of the assumptions (A1)–(A3).

Let $F = Ee^{-Mt}$, where M will be determined later; then F satisfies

(2.4) $$\begin{cases} F_t + MF + e^{-Mt}V(Fe^{Mt})(1 + F_x e^{Mt}) - \delta F_{xx} = \dfrac{1}{R}\left(e^{-Mt}\Phi(t) - \displaystyle\int_0^l F dx\right), \\ F(x,0) = E_0(x), \quad F(0,t) = E_1 e^{-Mt}, \quad F(l,t) = E_2 e^{-Mt}. \end{cases}$$

If F takes the positive maximum value $F_{\max}$ and the negative minimum value $F_{\min}$ in the interior of the domain, then

$$\begin{aligned} &MF_{\max} + e^{-Mt}V(F_{\max}e^{Mt}) \\ &\quad < \frac{1}{R}\left(|\Phi(t)| + \int_0^l |F|dx\right) \le \frac{1}{R}(|\Phi| + l|F|_{\max}), \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} &MF_{\min} + e^{-Mt}V(F_{\min}e^{Mt}) \\ &\quad > -\frac{1}{R}\left(|\Phi(t)| + \int_0^l |F|dx\right) \ge -\frac{1}{R}(|\Phi| + l|F|_{\max}). \end{aligned} \tag{2.6}$$

If $F_{\max} = |F|_{\max}$, from (2.5) and (A7) we have

$$M|F|_{\max} - C_v|F|_{\max} \le \frac{1}{R}(|\Phi| + l|F|_{\max}).$$

Now choose $M = (2l/R) + C_v$, and we obtain that

$$|F|_{\max} \le C(\Phi).$$

If $-F_{\min} = |F|_{\max}$, from (2.6), we can use the same argument to obtain the same result.

In the other cases, if F takes the maximum or minimum on the boundary or at the initial time, the argument is similar. Therefore, from the relationship of E and F we obtain that

$$\|E\|_{C^0} \le (\|E_0\|_{C^0} + \|E_1\|_{C^0} + \|E_2\|_{C^0} + C\|\Phi\|_{C^0}), \tag{2.7}$$

where C depends only on δ, R, T, and l.

Now for $\sigma \in [0,1]$, consider a map $\mathbf{T}_\sigma : \mathbf{B} \to \mathbf{B}$, where $\mathbf{B}$ is the Banach space $C^0(\overline{Q}_T)$. For any $H \in \mathbf{B}$, define $\mathbf{T}_\sigma(H)$ as the solution E of the following problem:

$$\begin{cases} E_t + V(E)(1 + E_x) - \delta E_{xx} = \dfrac{\sigma}{R}\left(\Phi(t) - \displaystyle\int_0^l H dx\right), \\ E(x,0) = \sigma E_0(x), \quad E(0,t) = \sigma E_1(t), \quad E(l,t) = \sigma E_2(t). \end{cases} \tag{2.8}$$

It is well known that the solution of problem (2.8) exists and belongs to $W_\infty^{2,1}(Q_T)$, which then belongs to $C^{1+\lambda,\frac{1+\lambda}{2}}(Q_T)$ for any $\lambda \in [0,1)$ (see [10]), and so the mapping is well defined and is compact. With the estimate (2.7), we know $\mathbf{T}_\sigma(H) = 0$, when $\sigma = 0$. Again with the estimate (2.7), we can easily verify that the map is continuous and each fixed point of the map is uniformly bounded with respect to σ. For any bounded set of $\mathbf{B}$, the map is uniformly continuous with respect to σ. Thus, we know that $\mathbf{T}_\sigma$ has at least one fixed point when $\sigma = 1$ in $C^0(\overline{Q}_T)$ by the Leray–Schauder fixed point theorem. By regularity, the fixed point is in $W_\infty^{2,1}(Q_T)$ and so in $C^{1+\lambda,\frac{1+\lambda}{2}}(Q_T)$; moreover, the solution can be continued up to the boundary and the corners. Its first derivatives can be continued up to the boundary, too, but not on the corners.

Therefore, we have the following.

THEOREM 2.1. *If assumptions* (A1), (A5), *and* (A7) *are satisfied, problem* (2.3) *admits at least one weak solution in* $W_\infty^{2,1}(Q_T)$, *then in* $C^0(\overline{Q}_T) \cap C^{1+\lambda,\frac{1+\lambda}{2}}(\overline{Q}_T \backslash P_c)$ *for any* $\lambda \in [0,1)$. *It satisfies the estimates* (2.7) *and*

$$\|E\|_{W_\infty^{2,1}(Q_T)} \le C\left(\|E_0\|_{C^2}, \sum_{i=1,2}\|E_i\|_{C^1}, \|\Phi\|_{C^1}\right),$$

where C *also depends on* δ, l, T, *and* R; *and* $P_c = \{x = 0, t = 0\} \cup \{x = l, t = 0\}$.

Suppose (A4) and (A6) are satisfied, if we increase the regularity of the solution of the problem in the standard method (see [6]) we can obtain the following regularity result, which also holds on the corners.

THEOREM 2.2. *If assumptions* (A2) *and* (A4)–(A7) *are satisfied, problem* (2.3) *admits at least one classical solution in* $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)$, *and it satisfies the estimates* (2.7) *and*

$$\|E\|_{C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)} \le C\left(\|E_0\|_{C^{2+\alpha}}, \sum_{i=1,2}\|E_i\|_{C^{1+\frac{\alpha}{2}}}, \|\Phi\|_{C^{1+\frac{\alpha}{2}}}\right),$$

where C *also depends on* δ, l, T, *and* R.

THEOREM 2.3. *If assumptions* (A3)–(A7) *are satisfied and* V *is sufficiently smooth, problem* (2.3) *has at least one classical solution in* $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T) \cap C^\infty(\overline{Q}_T \backslash P_c)$, *where* P_c *is defined in Theorem* 2.1.

Remark. For Dirichlet problems (2.3) and (2.2), (A4) is not necessarily essential. Indeed, we can obtain the same regularity result with a weaker assumption by modifying (A6). For convenience to the latter, however, we do not consider the more general problem.

2.2. Problem (2.1): The existence for $R = 0$. In this case, differentiating the second equation of (2.1) with respect to t, we have

$$\int_0^l E_t(x,t)\,dx = \Phi'(t), \tag{2.9}$$

then integrating the first equation of (2.1) with respect to x from 0 to l, we obtain

$$\int_0^l E_t dx + \int_0^l V(E)(1+E_x)dx - \delta\int_0^l E_{xx}dx = J(t)l. \tag{2.10}$$

Substituting (2.10) into the first equation in (2.1), we have

$$\begin{cases} E_t + V(E)(1+E_x) - \delta E_{xx} \\ \quad = \dfrac{1}{l}\left(\Phi' + \displaystyle\int_0^l V(E)(1+E_x)dx - \delta\int_0^l E_{xx}dx\right), \\ E(x,0) = E_0(x), \quad E(0,t) = E_1(t), \quad E(l,t) = E_2(t). \end{cases} \tag{2.11}$$

If (A4) is satisfied, problem (2.11) is equivalent to problem (2.1) when $R = 0$, and so we can consider problem (2.11) instead of problem (2.1).

Remark. (A4) insures the equivalence of (2.11) and (2.1) when $R = 0$, which is not difficult to verify. If Φ is a constant, it will not appear in (2.11), but the solution

of problem (2.11), which is the same as the one of problem (2.1), still depends on Φ via E_0 by (A4). However, (A4) is not necessary for the existence of the solution of problem (2.11).

Suppose E_0, E_i, $i = 1, 2$ and Φ satisfy (A5) and one of (A1)–(A3), V satisfies (A7). We first obtain a priori estimates.

Set $F(x,t) = E(x,t) - (1/l)\big(E_1(t)(l-x) + E_2(t)x\big)$, then F satisfies

$$
(2.12)\quad \begin{cases} F_t + V\left(F + \dfrac{1}{l}\big(E_1(l-x) + E_2x\big)\right)\left(1 + F_x + \dfrac{1}{l}(E_2 - E_1)\right) - \delta F_{xx} \\ \qquad = J(t) - \dfrac{1}{l}\big(E_1'(l-x) + E_2'x\big), \\ \displaystyle\int_0^l F(x,t)dx = \Phi(t) - \frac{l}{2}(E_1(t) + E_2(t)), \\ F(x,0) = E_0(x) - \dfrac{1}{l}\big(E_1(0)(l-x) + E_2(0)x\big), \quad F(0,t) = F(l,t) = 0. \end{cases}
$$

Take a test function $g_1 = F - (6/l^3)x(l-x)\int_0^l F dx$, then g_1 satisfies

$$
(2.13)\qquad g_1|_{x=0,x=l} = 0, \qquad \int_0^l g_1\, dx = 0.
$$

Note

$$
(2.14)\qquad \int_0^l J(t)\left(F - \frac{6}{l^3}x(l-x)\int_0^l F dx\right)dx = 0
$$

and the second equation in (2.12), then multiply both sides of the first equation in (2.12) by g_1 and integrate it by parts from 0 to l to get

$$
\begin{aligned}
(2.15)\qquad & \frac{1}{2}\frac{d}{dt}\int_0^l |F|^2 dx - \frac{d}{dt}\int_0^l F\left(\frac{6}{l^3}x(l-x)\int_0^l F dx\right)dx \\
& + \int_0^l F\left(\frac{6}{l^3}x(l-x)\left(\Phi' - \frac{l}{2}(E_1' + E_2')\right)\right)dx \\
& + \int_0^l V(F + \frac{1}{l}(E_1(l-x) + E_2x)) \\
& \quad \cdot \left(1 + F_x + \frac{1}{l}(E_2 - E_1)\right)\cdot\left(F - \frac{6}{l^3}x(l-x)\int_0^l F dx\right)dx \\
& + \delta\int_0^l F_x\left(F_x - \frac{6}{l^3}(l-2x)\int_0^l F dx\right)dx \\
& = -\int_0^l \frac{1}{l}\big(E_1'(l-x) + E_2'x\big)\left(F - \frac{6}{l^3}x(l-x)\int_0^l F dx\right)dx.
\end{aligned}
$$

Using (A7), the nonlinear term of (2.15) can be estimated as follows:

$$
\left|\int_0^l V\left(F + \frac{1}{l}\big(E_1(l-x) + E_2x\big)\right)\left(F_x + \frac{1}{l}(E_2 - E_1)\right)F dx\right|
$$

$$
\begin{aligned}
&= \left| \int_0^l V(E) E_x \left(E - \frac{1}{l}(E_1(l-x) + E_2 x) \right) dx \right| \\
&\le \left| \int_0^l V(E) E E_x dx \right| + C(C_v, E_1, E_2) \int_0^l (1 + |E|^2) dx + \epsilon \int_0^l |E_x|^2 dx \\
&= \left| \int_{E_1}^{E_2} V(E) E dE \right| + C(C_v, E_1, E_2) \int_0^l (1 + |F|^2) dx + \frac{\delta}{4} \int_0^l |F_x|^2 dx \\
&= C(E_1, E_2) + C \int_0^l (1 + |F|^2) dx + \frac{\delta}{4} \int_0^l |F_x|^2 dx,
\end{aligned}
$$

where C depends only on the given data and C_v is defined in (A7). By (A7), we have that $|C(E_1, E_2)| \le C(|E_1|^3 + |E_2|^3)$.

Thus, from (2.15), we have

$$
\frac{1}{2} \frac{d}{dt} \int_0^l |F|^2 dx + \frac{\delta}{2} \int_0^l |F_x|^2 dx \le C \left(1 + \int_0^l |F|^2 dx \right), \tag{2.16}
$$

where C depends only on l, T, δ, Φ, Φ', E_i, E_i', $i = 1, 2$. Integrating (2.16) from 0 to t and using Gronwall's inequality, we have

$$
\begin{aligned}
&\max_{0 \le t \le T} \int_0^l |F|^2 dx + \int_0^T \int_0^l |F_x|^2 dx dt \\
&\quad \le C(\|\phi|\|_{H^1}, \|E_0\|_{L_2}, \|E_1\|_{H^1}, \|E_2\|_{H^1}).
\end{aligned} \tag{2.17}
$$

Now, take $g_2 = F^3 - (6/l^3) x(l-x) \int_0^l F^3 dx$, which also satisfies (2.13), as a test function. Multiply both sides of the first equation in (2.12) by g_2 and integrate it by parts on $[0, l] \times [0, t]$, then use the same argument as above to obtain

$$
\begin{aligned}
&\max_{0 \le t \le T} \int_0^l |F|^4 dx + \int_0^T \int_0^l |F|^2 |F_x|^2 dx dt \\
&\quad \le C(\|\Phi\|_{W_4^1}, \|E_0\|_{L_4}, \|E_1\|_{W_4^1}, \|E_2\|_{W_4^1}).
\end{aligned} \tag{2.18}
$$

Now, take $g_3 = F_t - (6/l^3) x(l-x) \int_0^l F_t \, dx$, which also satisfies (2.13), as a test function. Multiply both sides of the equation in (2.12) by g_3 and integrate it by parts on $[0, l] \times [0, t]$. Noting that

$$
\int_0^T \int_0^l F F_x F_t dx dt \le \epsilon \int_0^T \int_0^l |F_t|^2 dx dt + C_\epsilon \int_0^T \int_0^l |F|^2 |F_x|^2 dx dt,
$$

and using a similar argument as above with (2.18), we obtain

$$
\begin{aligned}
&\int_0^T \int_0^l |F_t|^2 dx dt + \max_{0 \le t \le T} \int_0^l |F_x|^2 dx \\
&\quad \le C \left(\|\Phi\|_{W_4^1}, \|E_0\|_{H^1}, \|E_1\|_{W_4^1}, \|E_2\|_{W_4^1} \right).
\end{aligned} \tag{2.19}
$$

From (2.19), by the Sobolev imbedding theorem (see [1]), we have

$$
\|F\|_{L_\infty(0,T;C^{\frac{1}{2}}([0,l]))} \le C(\|\Phi\|_{W_4^1}, \|E_0\|_{H^1}, \|E_1\|_{W_4^1}, \|E_2\|_{W_4^1}), \tag{2.20}
$$

where C depends only on the given data.

On the other hand, as we pointed out before, problem (2.12) can be rewritten as follows:

$$
\begin{cases}
F_t - \delta F_{xx} \\
\quad = -V\left(F + \frac{1}{l}\big(E_1(l-x) + E_2 x\big)\right)\left(1 + F_x + \frac{1}{l}(E_2 - E_1)\right) \\
\qquad + \frac{1}{l}\left(\Phi' + \int_0^l V\left(F + \frac{1}{l}\big(E_1(l-x) + E_2 x\big)\right)\left(1 + F_x + \frac{1}{l}(E_2 - E_1)\right)dx\right. \\
\qquad\qquad \left. -\delta\int_0^l F_{xx}dx - \big(E_1'(l-x) + E_2' x\big)\right), \\
F(x,0) = F_0(x) = E_0(x) - \frac{1}{l}\big(E_1(0)(l-x) + E_2(0)x\big), \\
F(0,t) = F(l,t) = 0.
\end{cases} \tag{2.21}
$$

Thus from (2.21), we have, by Green's representation, that $F(x,t)$ can be written as follows:

$$
\begin{aligned}
F(x,t) = {} & \int_0^l G(x,y;t,0)F_0(y)dy \\
& + \int_0^t\int_0^l \Big\{G(x,y;t,\tau)\Big[-V(F + \tfrac{1}{l}(E_1(l-y) + E_2 y))(1 + F_y(y,\tau)) \\
& \qquad + \frac{1}{l}(E_2 - E_1)(\tau) + \frac{1}{l}\Big(\Phi'(\tau) + \int_0^l V(F + \frac{1}{l}(E_1(l-z) + E_2 z)) \\
& \qquad \cdot \Big(1 + F_z(z,\tau) + \frac{1}{l}(E_2 - E_1)(\tau)\Big)dz\Big)\frac{\delta}{l}\int_0^l F_{zz}(z,\tau)dz \\
& \qquad - \frac{1}{l}\big(E_1'(\tau)(l-y) + E_2'(\tau)y\big)\Big]\Big\}dyd\tau,
\end{aligned} \tag{2.22}
$$

where the Green function $G(x,y;t,\tau)$ satisfies the estimates (see [5], [8], and [14] for details),

$$
\begin{cases}
\left|\int_0^l G_x(x,y;t,\tau)dy\right| \le C(t-\tau)^{-\frac{1}{2}}, \\
\left|\int_0^l G_{xx}(x,y;t,\tau)dy\right| \le C(t-\tau)^{-\beta},
\end{cases} \tag{2.23}
$$

for $t > \tau \ge 0$, where $\beta \in (0,1)$ is some constant and C depends only on the given data.

From (2.22) along with the estimate (2.20), we can control the nonlinear term and obtain the estimate

$$
\begin{aligned}
&\|F\|_{L_\infty([0,l])}(t)+\|F_x\|_{L_\infty([0,l])}(t)+\|F_{xx}\|_{L_\infty([0,l])}(t)\\
&\le C\Bigg[\Big(\|\Phi\|_{C^1}+\|E_0\|_{C^2}+\|E_1\|_{C^1}+\|E_2\|_{C^1}\Big)\\
&\quad+\int_0^t b(t-\tau)\Big(\|F\|_{L_\infty([0,l])}(\tau)\\
&\quad+\|F_x\|_{L_\infty([0,l])}(\tau)+\|F_{xx}\|_{L_\infty([0,l])}(\tau)\Big)d\tau\Bigg],
\end{aligned}
\tag{2.24}
$$

where $b(t-\tau)=1+(t-\tau)^{-\frac{1}{2}}+(t-\tau)^{-\beta}$ and C depends only on the given data.

We will use the following lemma (see [5]).

LEMMA 2.4. *Let $f(t)$ be a nondecreasing function and $C>0$, $\gamma\in(0,1)$. If*

$$
y(t)\le f(t)+C\int_0^t\frac{y(\tau)}{|t-\tau|^\gamma}\,d\tau,
$$

then

$$
y(t)\le Cf(t),
$$

where C depends only on γ and the upper bound of T.

Now, after differentiating (2.22), we use Lemma 2.4, formula (2.24), estimates (2.23), and the first equation in (2.21) to obtain the estimate

$$
\|F\|_{L_\infty(0,T;W^2_\infty([0,l]))}+\|F_t\|_{L_\infty(Q_T)}\le C(\|\Phi\|_{C^1},\|E_0\|_{C^2},\|E_1\|_{C^1},\|E_2\|_{C^1}),
\tag{2.25}
$$

where C depends only on the given data.

Hence, E satisfies the estimate

$$
\|E\|_{L_\infty(0,T;W^2_\infty([0,l]))}+\|E_t\|_{L_\infty(Q_T)}\le C(\|\Phi\|_{C^1},\|E_0\|_{C^2},\|E_1\|_{C^1},\|E_2\|_{C^1}).
\tag{2.26}
$$

By the Sobolev imbedding theorem (see [1]), we have

$$
\|E\|_{L_\infty(0,T;C^{1+\lambda}([0,l]))}\le C(\|\Phi\|_{C^1},\|E_0\|_{C^2},\|E_1\|_{C^1},\|E_2\|_{C^1}),
\tag{2.27}
$$

where $\lambda\in(0,1)$.

Now, let $\sigma\in[0,1]$. In a Banach space $\mathbf{B}=L_\infty\big(0,T;C^1[0,l]\big)$ consider the map $\mathbf{T}_\sigma$ defined as follows: For $H\in\mathbf{B}$, let $\mathbf{T}_\sigma(H)$ be the $W^{2,1}_\infty(Q_T)$ solution E of the problem

$$
\begin{cases}
E_t+V(E)(1+E_x)-\delta E_{xx}=\dfrac{\sigma}{l}\left(\Phi'+\displaystyle\int_0^l V(H)(1+H_x)dx-\delta H_x\Big|_0^l\right),\\
E(x,0)=\sigma E_0(x),\quad E(0,t)=\sigma E_1,\quad E(l,t)=\sigma E_2.
\end{cases}
\tag{2.28}
$$

With estimates (2.27), we can easily verify the hypotheses of the Leray–Schauder fixed point theorem without difficulty and conclude that the map has a fixed point in $\mathbf{B}$. That is, there exists at least one solution of the problem in $L_\infty(0,T;C^1([0,l]))$, which is then in $W^{2,1}_\infty(Q_T)$ and hence in $C^{1+\lambda,\frac{1+\lambda}{2}}(Q_T)$ for any $\lambda\in[0,1)$. The continuity on the boundary can be discussed as we did in the last section.

THEOREM 2.5. *If assumptions* (A1), (A5), *and* (A7) *are satisfied, problem* (2.11) *admits at least one solution in $W^{2,1}_\infty(Q_T)$ and so in $C^0(\overline{Q}_T)\cap C^{1+\lambda,\frac{1+\lambda}{2}}(\overline{Q}_T\backslash P_c)$ for*

any $\lambda \in [0,1)$, *where* P_c *is defined in Theorem* 2.1. *It satisfies an estimate that is similar to the one in Theorem* 2.1 *with* C *independent of* R.

Increasing the regularity of the solution in the typical way, if (A6′) is assumed, we can obtain the following theorems.

THEOREM 2.6. *If assumptions* (A2), (A5), (A6′), *and* (A7) *are satisfied, problem* (2.11) *admits at least one classical solution in* $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)$. *It satisfies an estimate that is similar to the one in Theorem* 2.2 *with* C *independent of* R.

THEOREM 2.7. *If assumptions* (A3), (A5), (A6′), *and* (A7) *are satisfied and* V *is sufficiently smooth, problem* (2.11) *has at least one classical solution in* $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T) \cap C^{\infty}(\overline{Q}_T \backslash P_c)$, *where* P_c *is defined in Theorem* 2.1.

Combining the results for the cases of $R > 0$ and $R = 0$, we have the following.

THEOREM 2.8. *Under the assumptions* (A4), (A5), (A6) (or (A6′)), (A7) *and one of* (A1)–(A3), *problem* (2.1) *admits at least one solution in* $W_{\infty}^{2,1}(Q_T)$, *and the solution is smoother if the given functions are smoother.*

2.3. Problem (2.2). Arguing as we did in §2.1, we can investigate problem (2.2) and obtain the following theorems.

THEOREM 2.9. *If assumptions* (A1), (A5), *and* (A7) *are satisfied, problem* (2.2) *admits at least one solution in* $W_{\infty}^{2,1}(Q_T)$, *which is then in* $C^0(\overline{Q}_T) \cap C^{1+\lambda,\frac{1+\lambda}{2}}(\overline{Q}_T \backslash P_c)$ *for any* $\lambda \in [0,1)$, *where* P_c *is defined in Theorem* 2.1. *The solution satisfies an estimate which is similar to the one in Theorem* 2.1 *with* C *independent of* R.

THEOREM 2.10. *If assumptions* (A2), (A4)–(A7) *are satisfied, problem* (2.2) *admits at least one classical solution in* $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)$. *It satisfies an estimate which is similar to the one in Theorem* 2.2 *with* C *independent of* R.

THEOREM 2.11. *If assumptions* (A3)–(A7) *are satisfied and* V *is sufficiently smooth, problem* (2.2) *has at least one classical solution in* $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T) \cap C^{\infty}(\overline{Q}_T \backslash P_c)$, *where* P_c *is as in Theorem* 2.1.

2.4. Further properties of the solution to the Dirichlet problem. For the solution of problems (2.1) and (2.2) we have the following continuous dependence result.

THEOREM 2.12. *Let* $E(x,t)$ *and* $\overline{E}(x,t)$ *be the solutions of problem* (2.1) *or problem* (2.2) *corresponding to the boundary and initial conditions* $E_i(t)$, $E_0(x)$, $\Phi(t)$ *and* $\overline{E}_i(t)$, $\overline{E}_0(x)$, $\overline{\Phi}(t)$, *respectively, where* $i = 1,2$. *Then the following continuous dependence estimates hold. If* (A1), (A5), *and* (A7) *are assumed for the given functions, then*

$$(2.29)\quad \|E - \overline{E}\|_{W_{\infty}^{2,1}(Q_T)} \leq C\left(\|E_0 - \overline{E}_0\|_{C^2([0,l])} + \sum_{i=1,2} \|E_i - \overline{E}_i\|_{C^1([0,T])} + \|\Phi - \overline{\Phi}\|_{C^1([0,T])} \right);$$

or if (A2), (A4)–(A6) (*or* (A6′)), *and* (A7) *are assumed for the given functions, then*

$$(2.30)\quad \|E - \overline{E}\|_{C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)} \leq C\left(\|E_0 - \overline{E}_0\|_{C^{2+\alpha}([0,l])} + \sum_{i=1,2} \|E_i - \overline{E}_i\|_{C^{1+\frac{\alpha}{2}}([0,T])} + \|\Phi - \overline{\Phi}\|_{C^{1+\frac{\alpha}{2}}([0,T])} \right);$$

or if (A3) *and* (A7) *are assumed for the given functions and* V *is sufficiently smooth, then for any* $0 < k < \infty$,

$$\|E - \overline{E}\|_{C^k(\overline{Q}_T \setminus P_c)} \le C\Big(\|E_0 - \overline{E}_0\|_{C^k([0,l])} + \sum_{i=1,2} \|E_i - \overline{E}_i\|_{C^k([0,T])} + \|\Phi - \overline{\Phi}\|_{C^k([0,T])}\Big). \tag{2.31}$$

The constants C *in* (2.29)–(2.31) *depend only on the given data. In* (2.31) C *also depends on* k.

Proof. We only prove the results for problem (2.1) as the results for problem (2.2) can be obtained by a similar argument.

We already have the estimates that depend on the given data in Theorems 2.1–2.3 and 2.5–2.7. Let $F = E - \overline{E}$; then F will satisfy for $R > 0$

$$\begin{cases} F_t + V(E)F_x + \displaystyle\int_0^1 V'(\tau E + (1-\tau)\overline{E})d\tau(1 + \overline{E}_x)F - \delta F_{xx} \\ \quad = \dfrac{1}{R}\left((\Phi - \overline{\Phi}) - \displaystyle\int_0^l F dx\right), \\ F(x,0) = E_0 - \overline{E}_0, \quad F(0,t) = E_1 - \overline{E}_1, \quad F(l,t) = E_2 - \overline{E}_2; \end{cases} \tag{2.32}$$

or for $R = 0$,

$$\begin{cases} F_t + V(E)F_x + \displaystyle\int_0^1 V'\big(\tau E + (1-\tau)\overline{E}\big)d\tau F(1 + \overline{E}_x) - \delta F_{xx} \\ \quad = \dfrac{1}{l}\left((\Phi' - \overline{\Phi}') + \displaystyle\int_0^l V(E)F_x dx - \delta \int_0^l F_{xx} dx \right. \\ \qquad \left. + \displaystyle\int_0^l \int_0^1 V'\big(\tau E + (1-\tau)\overline{E}\big)d\tau(1 + \overline{E}_x)F dx\right), \\ F(x,0) = E_0 - \overline{E}_0, \quad F(0,t) = E_1 - \overline{E}_1, \quad F(l,t) = E_2 - \overline{E}_2. \end{cases} \tag{2.33}$$

Now using Green's representation, the estimates (2.7), (2.26), (2.27), and regularity results, then differentiating the first equation in (2.3) or (2.11) or (2.2), we obtain the following estimates for $R \ge 0$: For any integers $m \ge 2$ and $h \ge 1$,

$$\begin{aligned} &\sum_{0 \le i \le m, 0 \le j \le h} \left\| \frac{\partial^{i+j} F}{\partial x^i \partial t^j}(\cdot, t) \right\|_{L_\infty([0,l])} \\ &\le C \left(\|E_0 - \overline{E}_0\|_{C^m([0,l])} + \sum_{i=1,2} \|E_i - \overline{E}_i\|_{C^h([0,T])} + \|\Phi - \overline{\Phi}\|_{C^h([0,T])} \right. \\ &\qquad \left. + \sum_{0 \le i \le m, 0 \le j \le h} \int_0^t b(t-\tau) \left\| \frac{\partial^{i+j} F}{\partial x^i \partial t^j}(\cdot, \tau) \right\|_{L_\infty([0,l])} d\tau \right), \end{aligned} \tag{2.34}$$

where $b(t-\tau)$ is defined in (2.24), and C depends only on the given data and m, h. Applying Lemma 2.4 and the imbedding theorem to (2.34) yields (2.29)–(2.31). □

The uniqueness of the solution is a direct consequence of Theorem 2.12.

COROLLARY 2.13. *The solutions of problem* (2.1) *and* (2.2) *are unique.*

3. The integral boundary condition for the integrodifferential problem. In this section, we will discuss the existence, uniqueness, regularity, and continuous dependence on given data of the solution for problem (1.7) by using the results of §2. We transform the differential boundary conditions into integral ones. We then define a map on the unknown J in the boundary data. In this way, the boundary conditions become Dirichlet boundary conditions, which we have already considered in §2. We prove that the map has a fixed point. In §3.1, we consider the case $R > 0$. For the case $R = 0$, which is discussed in §3.2, we encounter the difficulties of the compactness of the map as well as obtaining an estimate when T is large. We will make a suitable transformation to overcome the difficulties.

In order to discuss the regularity, an additional compatibility assumption is required on the initial function $E_0(x)$,

$$\begin{aligned} &-\frac{E_0(0)}{\rho_0} + V(E_0(0))(1 + E_{0x}(0)) - \delta E_{0xx}(0) \\ &= -\frac{E_0(l)}{\rho_l} + V(E_0(l))(1 + E_{0x}(l)) - \delta E_{0xx}(l) = 0. \end{aligned} \tag{A8}$$

3.1. The case of $R > 0$. First of all, we point out that the boundary condition in problem (1.7) can be changed into an integral boundary condition. We are going to discuss the following equivalent form of problem (1.7):

$$\begin{cases} E_t + V(E)(1 + E_x) - \delta E_{xx} = J(t), \\ \displaystyle\int_0^l E(x,t)dx + RJ(t) = \Phi(t), \\ E(x,0) = E_0(x), \\ \displaystyle E(0,t) = E_0(0)\, e^{-\frac{t}{\rho_0}} + \int_0^t J(s)\, e^{\frac{s-t}{\rho_0}} ds, \\ \displaystyle E(l,t) = E_0(l)\, e^{-\frac{t}{\rho_l}} + \int_0^t J(s)\, e^{\frac{s-t}{\rho_l}} ds. \end{cases} \tag{3.1}$$

Suppose E_0 and Φ satisfy (A4) and one of (A1)–(A3) and V satisfies (A7). We start to obtain an a priori estimate in $L_\infty([0,T])$ for $J(t)$ in problem (3.1). From the second equation in (3.1) and estimate (2.7), we have for any $t \in [0,T)$,

$$\begin{aligned} |J(t)| &= \frac{1}{R}\left|\Phi(t) - \int_0^l E(x,t)dx\right| \\ &\le C\left(1 + \|E\|_{C^0([0,l]\times[0,t])}\right) \le C\left(1 + \int_0^t |J(\tau)|d\tau\right), \end{aligned} \tag{3.2}$$

where C depends on E_0, Φ, R, l, and C_v.

By Gronwall's inequality, we have

$$\|J(t)\|_{L_\infty([0,T))} \le C, \tag{3.3}$$

where C depends only on E_0, Φ, R, l, T, and C_v.

Consider the Banach space $\mathbf{B} = C^0([0,T))$, and a closed subset $\mathcal{M} = \{I \in \mathbf{B};\ I(0) = 0\}$. For $\sigma \in [0,1]$, consider the map $\mathbf{T}_\sigma$ defined as follows: For any $I \in \mathcal{M}$,

let $\mathbf{T}_\sigma(I)$ be the solution J of the following problem:

$$
\begin{cases}
E_t + V(E)(1+E_x) - \delta E_{xx} = \dfrac{1}{R}\left(\sigma\Phi - \displaystyle\int_0^l E dx\right), \\
E(x,0) = \sigma E_0(x), \\
E(0,t) = \sigma E_0(0)\, e^{-\frac{t}{\rho_0}} + \sigma \displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_0}} ds, \\
E(l,t) = \sigma E_0(l)\, e^{-\frac{t}{\rho_l}} + \sigma \displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_l}} ds, \\
J(t) = \dfrac{1}{R}\left(\sigma\Phi - \displaystyle\int_0^l E dx\right).
\end{cases} \tag{3.4}
$$

Observe that the first four equations in (3.4) give rise to the same problem as problem (2.3) with the boundary conditions

$$
\begin{cases}
E_1 = \sigma E_0(0)\, e^{-\frac{t}{\rho_0}} + \sigma \displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_0}} ds, \\
E_2 = \sigma E_0(l)\, e^{-\frac{t}{\rho_l}} + \sigma \displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_l}} ds.
\end{cases} \tag{3.5}
$$

It is obvious that $E_i \in C^1([0,T])$, $i = 1,2$, for any $I \in C^0([0,T))$. It is also easy to verify (A5). So by the result of Theorem 2.1, E exists and $E \in C^0(\overline{Q}_T) \cap C^{1+\lambda, \frac{1+\lambda}{2}}(\overline{Q}_T \backslash P_c)$ for any $\lambda \in (0,1)$. Thus, from the last equation in problem (3.4), $J(t)$ exists and belongs to $C^{\frac{1+\lambda}{2}}([0,T]) \subset \mathbf{B}$ as well as to the set $\mathcal{M}$ since $J(0) = 0$ by (A4); therefore, the map is well defined and compact. Continuity of the map follows from Theorem 2.12. Using estimate (2.7), it is easy to verify that when $\sigma = 0$, $T_\sigma(I) = 0$. For each fixed point we have estimate (3.3) and so by the Leray–Schauder fixed point theorem, the map has at least one fixed point in $\mathbf{B}$, that is, the following.

THEOREM 3.1. *For the case of $R > 0$, if $E_0 \in C^2([0,l])$, $\Phi \in C^1([0,T])$ and* (A4), (A7) *are satisfied, then problem* (1.7) *admits at least one solution belonging to* $W^{2,1}_\infty(Q_T)$.

Moreover, we have the following.

THEOREM 3.2. *For the case of $R > 0$, the solution of problem* (1.7) *is unique.*

Proof. If there are two solutions E and $\overline{E}$, corresponding to J and $\overline{J}$, respectively, then the function $F = E - \overline{E}$ satisfies

$$
\begin{cases}
F_t + V(E)F_x + \displaystyle\int_0^1 V'(\tau E + (1-\tau)\overline{E}) d\tau (1 + \overline{E}_x) F - \delta F_{xx} \\
\quad = -\dfrac{1}{R}\displaystyle\int_0^l F dx, \\
F(x,0) = 0, \\
F(0,t) = \displaystyle\int_0^t (J(s) - \overline{J}(s))\, e^{\frac{s-t}{\rho_0}} ds, \\
F(l,t) = \displaystyle\int_0^t (J(s) - \overline{J}(s))\, e^{\frac{s-t}{\rho_l}} ds, \\
J(s) - \overline{J}(s) = -\dfrac{1}{R}\displaystyle\int_0^l F(x,t) dx.
\end{cases} \tag{3.6}
$$

By estimate (2.7), we easily observe that $|F|$ can not attain its maximum value in the domain, but only on the boundary. So, from (3.2), we have

$$|J(t) - \overline{J}(t)| \leq \frac{1}{R} \int_0^l |F(x,t)| dx \leq C \int_0^t |J(\tau) - \overline{J}(\tau)| d\tau.$$

By Gronwall's inequality, we have

$$|J(t) - \overline{J}(t)| \leq 0.$$

That is,

$$J(t) \equiv \overline{J}(t).$$

By the uniqueness theorem of Corollary 2.13, we obtain that

$$E(x,t) \equiv \overline{E}(x,t).$$

Therefore, when $R > 0$, the solution of problem (1.7) is unique. □

If (A8) is satisfied, we can verify (A6) and use Theorems 2.2 and 2.3 to increase the regularity of the solution by a standard argument to obtain the following.

THEOREM 3.3. *For the case of $R > 0$, if $E_0 \in C^{2+\alpha}([0,l])$, $\Phi \in C^{1+\frac{\alpha}{2}}([0,T])$ and* (A4), (A7), *and* (A8), *are satisfied, then the solution E of problem* (1.7) *belongs to the space $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)$.*

Proof. We can easily verify (A6) if (A4) and (A8) are satisfied when $\sigma = 1$. For the fixed point $J(t)$, we already have the estimate $J(t) \in C^{\frac{1+\lambda}{2}}([0,T])$, which implies the boundary data belongs to $C^{1+\frac{1+\lambda}{2}}([0,T])$. Therefore, since $\Phi \in C^{1+\frac{\alpha}{2}}$, the solution E will belong to $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T)$ by Theorem 2.2. Hence, J will be in $C^{1+\frac{\alpha}{2}}([0,T])$. Repeating this argument yields the desired regularity. □

Using Theorem 2.3, and arguing as in §2 and in the proof of Theorem 3.3 where $k = 1$ up to ∞, we obtain the following.

THEOREM 3.4. *For the case of $R > 0$, if $E_0 \in C^\infty([0,l])$, $\Phi \in C^\infty([0,T])$, $V \in C^\infty(\mathbb{R})$, and* (A4), (A7), *and* (A8) *are satisfied, then the solution E of problem* (1.7) *belongs to $C^{2+\alpha,1+\frac{\alpha}{2}}(\overline{Q}_T) \cap C^\infty(\overline{Q}_T \backslash P_c)$, where P_c is defined in Theorem* 2.1.

Arguing as in §2, and using estimate (3.2), we obtain the following continuous dependence result.

THEOREM 3.5. *For the case of $R > 0$, if* (A4), (A7), *and* (A8) *are satisfied, then the solution of the problem* (1.7) *depends continuously on the initial function $E_0(x)$ and the given function $\Phi(t)$.*

3.2. The case of $R = 0$. In this section, we consider the case $R = 0$ in problem (1.7). We still suppose E_0 and Φ satisfy (A4) and one of the (A1)–(A3).

The method for the case of $R > 0$ in the last section cannot be used for the case of $R = 0$ directly because we cannot get the image $J(t)$ from the second equation in (1.7) but only (2.10) with $R = 0$. Therefore, for the case $R = 0$, we cannot prove compactness of a map similar to (3.4) by using the corresponding result in §2.2. Another difficulty is to obtain an a priori estimate when T is large. To overcome these difficulties, we introduce the following transformation from (1.7) when $R = 0$:

$$\int_0^t J(s) ds = L(t), \tag{3.7}$$

$$F = E - L(t). \tag{3.8}$$

Then F satisfies

$$\begin{cases} F_t + V(F + L(t))(1 + F_x) - \delta F_{xx} = 0, \\ \int_0^l F(x,t)dx = \Phi(t) - lL(t), \\ F(x,0) = E_0(x), \\ F_t(0,t) + \dfrac{1}{\rho_0}(F(0,t) + L(t)) = 0, \\ F_t(l,t) + \dfrac{1}{\rho_l}(F(l,t) + L(t)) = 0. \end{cases} \tag{3.9}$$

Remark. We cannot use this transformation to discuss the Dirichlet problem (2.1) in §2 because after transforming, the problem no longer has Dirichlet boundary conditions.

From the second equation in problem (3.9) we have

$$L(t) = \frac{1}{l}\left(\Phi - \int_0^l F(x,t)dx\right). \tag{3.10}$$

Substituting (3.10) into the first equation of problem (3.9), and change the boundary conditions into the form of integral equation, we obtain the following integrodifferential problem

$$\begin{cases} F_t + V\left(F + \dfrac{1}{l}\left(\Phi - \int_0^l F(x,t)dx\right)\right)(1 + F_x) - \delta F_{xx} = 0, \\ F(x,0) = E_0(x), \\ F(0,t) = E_0(0)\, e^{-\frac{t}{\rho_0}} - \dfrac{1}{\rho_0}\displaystyle\int_0^t L(s)\, e^{\frac{s-t}{\rho_0}} ds, \\ F(l,t) = E_0(l)\, e^{-\frac{t}{\rho_l}} - \dfrac{1}{\rho_l}\displaystyle\int_0^t L(s)\, e^{\frac{s-t}{\rho_l}} ds. \end{cases} \tag{3.11}$$

Now, consider the Banach space $\mathbf{B} = C^0([0,T])$ and the closed subset $\mathcal{M} = \{I \in \mathbf{B};\ I(0) = 0\}$ on which we define the map $\mathbf{T}_\sigma(I) = L$ for $\sigma \in [0,1]$ such that for any $I(t) \in \mathcal{M}$, $L(t)$ is the solution of the problem

$$\begin{cases} F_t + V\left(F + \dfrac{1}{l}\left(\sigma\Phi - \int_0^l F(x,t)dx\right)\right)(1 + F_x) - \delta F_{xx} = 0, \\ F(x,0) = \sigma E_0(x), \\ F(0,t) = \sigma E_0(0)\, e^{-\frac{t}{\rho_0}} - \sigma\dfrac{1}{\rho_0}\displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_0}} ds, \\ F(l,t) = \sigma E_0(l)\, e^{-\frac{t}{\rho_l}} - \sigma\dfrac{1}{\rho_l}\displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_l}} ds, \\ L(t) = \dfrac{1}{l}\left(\sigma\Phi - \displaystyle\int_0^l F(x,t)dx\right). \end{cases} \tag{3.12}$$

Now, this is the problem (2.2) discussed in §2.3 with

$$\begin{cases} E_1 = \sigma E_0(0)\, e^{-\frac{t}{\rho_0}} - \sigma \dfrac{1}{\rho_0} \displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_0}}\, ds, \\ E_2 = \sigma E_0(l)\, e^{-\frac{t}{\rho_l}} - \sigma \dfrac{1}{\rho_l} \displaystyle\int_0^t I(s)\, e^{\frac{s-t}{\rho_l}}\, ds. \end{cases}$$

We can easily verify (A1) and (A5), when $I \in \mathbf{B}$ and (3.7), (3.8) hold. Then by the result of Theorem 2.9, we know that problem (3.12) admits a solution $E \in W_\infty^{2,1}(Q_T)$, which is then in $C^0(\overline{Q}_T) \cap C^{1+\lambda, \frac{1+\lambda}{2}}(\overline{Q}_T \backslash P_c)$ for any $\lambda \in (0,1)$. From the last equation in problem (3.12), we have that the image of the map $L(t)$ exists and is in $\mathbf{B}$; also, $L(0) = 0$, so $L \in \mathcal{M}$; that is, the map is well defined. Moreover, since $L(t) \in C^{\frac{1+\lambda}{2}}([0,T])$, the map is compact. The continuity of the map can be verified by using the results in §2.4. It is easy to verify that $\mathbf{T}_\sigma(I) = 0$ when $\sigma = 0$.

We want to prove that each fixed point has a uniform estimate with respect to σ so that we can say the map has a fixed point in $\mathbf{B}$ when $\sigma = 1$ by the Leray–Schauder fixed point theorem. In fact, as we discussed in the last section, for any $t \in [0,T)$, if $|F|$ attains its maximum value in the interior of the domain or at the initial time, then $|F| \le C(\Phi, E_0)$. Therefore from the second equation of (3.9), we have that $|L| \le C(\Phi, E_0)$. For any $t \in [0,T)$, if $|F|$ attains its maximum value in the domain Q_T on the boundary, from the second equation of (3.9), we have

$$|L|(t) = \frac{1}{l}\left(\Phi - \int_0^l F(x,t)dx\right) \le C(1 + \|F\|_{L_\infty([0,l]\times[0,t])}) \le C\left(1 + \int_0^t |L|(s)ds\right).$$

By Gronwall's inequality, we have

$$\|L\|_{C^0([0,t])} \le C$$

for any $t \in [0,T)$, where C depends only on the given data.

That is, the mapping has at least one fixed point when $\sigma = 1$. Therefore, we have the following theorem.

THEOREM 3.6. *For the case of $R = 0$, if $E_0 \in C^2([0,l])$, $\Phi \in C^1([0,T])$, and* (A4) *and* (A7) *are satisfied, problem* (1.7) *admits at least one solution belonging to* $W_\infty^{2,1}(Q_T)$.

The uniqueness of the solution can be obtained by using an argument similar to the uniqueness argument done in the last section.

THEOREM 3.7. *For the case of $R = 0$, the solution of problem* (1.7) *is unique.*

If (A8) is satisfied, we can easily verify (A6). The regularity and continuous dependence of the solution can be proven as we have done in the last section. Beginning with (3.11), and using Theorems 2.10 and 2.11, we can increase the regularity step by step as we did in the last section to obtain the following results.

THEOREM 3.8. *For the case $R = 0$, if $E_0 \in C^{2+\alpha}([0,l])$, $\Phi \in C^{1+\frac{\alpha}{2}}([0,T])$ and* (A4), (A7), *and* (A8) *are satisfied, then the solution E of problem* (1.7) *exists and belongs to $C^{2+\alpha, 1+\frac{\alpha}{2}}(\overline{Q}_T)$.*

THEOREM 3.9. *For the case of $R = 0$, if $E_0 \in C^\infty([0,l])$, $\Phi \in C^\infty([0,T])$, $V \in C^\infty(\mathbb{R})$, and* (A4), (A7), *and* (A8) *are satisfied, then the solution E of problem* (1.7) *belongs to $C^{2+\alpha, 1+\frac{\alpha}{2}}(\overline{Q}_T) \cap C^\infty(\overline{Q}_T \backslash P_c)$, where P_c is defined in Theorem* 2.3.

THEOREM 3.10. *For the case of $R = 0$, if $E_0 \in C^2([0,l])$, $\Phi \in C^1([0,T])$, and* (A4), (A7), *and* (A8) *are satisfied, then the solution of problem* (1.7) *depends continuously on the initial function $E_0(x)$ and the given function $\Phi(t)$.*

In summary, combining all the results together, we have the following main theorem.

MAIN THEOREM. *If* $E_0 \in C^2([0,l])$, $\Phi \in C^1([0,T])$, *and* (A4), (A7), *and* (A8) *are satisfied, then problem* (1.7) *admits a unique solution and the solution is regular if the given functions are sufficiently smooth. Moreover, the solution of the problem depends continuously on the initial condition* $E_0(x)$ *and the given function* $\Phi(t)$.

Acknowledgment. The author would like to thank L.L. Bonilla, S. Bricher, B. Louro, and J.F. Rodrigues for many helpful discussions related to this work.

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic, New York, 1975.
[2] L. L. BONILLA, *Solitary waves in semiconductors with finite geometry and the Gunn effect*, SIAM J. Appl. Math., 51 (1991), pp. 727–747.
[3] L. L. BONILLA AND F. J. HIGUERA, *Gunn instability in finite samples of* GaAs. I. *Stationary states, stability and boundary conditions*, Phys. D, 52 (1991), pp. 458–476.
[4] ———, *Gunn instability in finite samples of* GaAs. II. *Oscillatory states in long samples*, Phys. D, 57 (1992), pp. 161–184.
[5] J. M. CHADAM AND H. M. YIN, *An iteration procedure for a class of integrodifferential equations of parabolic type*, J. Integral Equations Appl., 2 (1989), pp. 31–47.
[6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[7] J. B. GUNN, *Microwave oscillations in* III-IV *semiconductors*, Solid Stater Comm., 1 (1963), pp. 88–91.
[8] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Amer Math. Soc. Trans. 23, Providence, RI, 1968.
[9] P. A. MARKOWICH, C. RINGHOFER, AND C. SCHMEISER, *Semiconductor Equations*, Springer-Verlag, Wien, Austria, 1990.
[10] O. A. OLEINIK AND S. N. KRUZHKOV, *Quasilinear second-order parabolic equations with many independent variables*, Uspekhi Mat. Nauk, 16 (1961), pp. 105–146.
[11] K. SEEGER, *Semiconductor Physics*, 4th ed., Springer, Berlin, 1989.
[12] S. M. SZE, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981.
[13] M. P. SHAW, H. L. GRUBIN, AND P. R. SOLOMON, *The Gunn–Hilsum Effect*, Academic Press, New York, 1979.
[14] E. G. YANIK AND G. FAIRWEATHER, *Finite element methods for parabolic and hyperbolic partial integrodifferential equations*, Nonlin. Anal. Theory Meth. Appl., 12 (1988), pp. 785–809.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1393–1401, September 1994

008

THE STABILITY OF THE EQUILIBRIUM OF A NONLINEAR HILL'S EQUATION*

RAFAEL ORTEGA†

Abstract. Sufficient conditions for the stability of the trivial solution of a nonlinear Hill's equation are obtained. As a consequence, the classical Lyapunov's criterion for stability is extended to certain nonlinear differential equations.

The proofs are based on the computation of the corresponding Birkhoff normal forms together with an application of the twist theorem.

Key words. Lyapunov stability, Hill's equation, normal form, twist theorem

AMS subject classifications. 34D20, 58F10

1. Introduction.

The differential equation

$$(*) \qquad y'' + a(t)\, y + c(t)\, y^{2n-1} + \cdots = 0,$$

with $n \geq 2, a(t)$, and $c(t)$ T-periodic functions, can be seen as a nonlinear version of the classical Hill's equation. (Here it is understood that the remaining terms in the equation are also T-periodic in time and dominated by the power y^{2n} in a neighborhood of $y = 0$.) The purpose of this paper is the study of the stability in the Lyapunov sense of the trivial solution $y = 0$.

The linearization of $(*)$ around the origin leads to the classical Hill's equation

$$(**) \qquad y'' + a(t)\, y = 0.$$

It is well known that the stability of this linear equation is closely related to the position in the plane of the Floquet multipliers λ_1, λ_2. The elliptic case ($|\lambda_i| = 1, \lambda_i \neq \pm 1$) corresponds to the strong stability of $(**)$ in the sense of Krein (also called parametric stability), and the hyperbolic case ($|\lambda_i| \neq 1$) leads to instability and in the parabolic case ($\lambda_i = \pm 1$) both stability or instability can occur. For the equation $(**)$ many criterions for stability have been obtained since Lyapunov's times (see [4], [6], and [12]). However, one can construct examples where the linear equation is stable or even strongly stable but $y = 0$ is not a stable solution of $(*)$. The main result of this paper will show that if the coefficient $c(t)$ is either positive or negative then the stability of $(**)$ implies the stability of $y = 0$ in the nonlinear equation. As a consequence the classical stability criterions for the Hill's equation, such as Lyapunov criterion, can be extended to $(*)$ if one assumes that $c(t)$ does not change sign.

The original motivation for the present work was the study of the stability of the equilibrium position of the pendulum of variable length, sometimes called the swing. The corresponding equation can be written in the form

$$y'' + \alpha(t) \sin y = 0,$$

where $\alpha(t)$ is a positive and T-periodic function depending on the length. This equation can be included in the class $(*)$ with $n = 2, a(t) = \alpha(t), c(t) = -\alpha(t)/3! \leq 0$,

* Received by the editors October 21, 1992; accepted for publication June 18, 1993.

† Departamento de Matemática Aplicada, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain.

and the stability of $y = 0$ is implied by the stability of the linearized equation. This result for the swing only leaves open the non-semisimple parabolic case and extends previous results in [2], where α was a small perturbation of a certain constant, and in [9], where α had to satisfy an additional inequality.

The proof of the main theorem of this paper consists in a careful computation of certain Birkhoff normal forms together with an application of the twist theorem. This method of proof of the stability of a periodic solution gives additional information on the existence of subharmonic and quasi-periodic solutions (see [10, §§24 and 36]), but this fact will not be mentioned in the rest of the paper. The previous paper [9] used similar ideas but was based on the results on stability given in [10, §34]. This new paper uses some consequences of the twist theorem obtained more recently in [11] and [1], and improves the results in [9] on equation (*).

The rest of the paper is organized in the following way. Section 2 is devoted to state the main result and to discuss some of its consequences. Section 3 deals with the proof of the main result. Before this proof, some useful facts on the stability of fixed points of area-preserving maps are collected. Finally, in §4 there are examples of equations in the class (*) showing that the equilibrium may become unstable when the assumption of the main theorem is not satisfied.

In what follows, to say that a solution of a differential equation or a fixed point of a homeomorphism is stable will mean that it is Lyapunov stable in the future and also in the past. If it is not stable it will be called unstable.

2. Stability criteria. The equation under consideration is

$$y'' + a(t)\, y + c(t)\, y^{2n-1} + d(t, y) = 0, \qquad n \geq 2, \tag{2.1}$$

and the following general hypotheses are always assumed:

a, c: $\mathbb{R} \to \mathbb{R}$ are continuous and T-periodic functions and $\int_0^T |c(t)|\, dt \neq 0$.

d: $\mathbb{R} \times (-\varepsilon, \varepsilon) \to \mathbb{R}$, $(\varepsilon > 0)$, is a continuous function with continuous derivatives of all orders with respect to y, T-periodic with respect to t and such that

$$d(t, y) = O\left(|y|^{2n}\right), \qquad y \to 0, \text{ uniformly with respect to } t \in \mathbb{R}.$$

The solution $y = 0$ has the variational equation

$$y'' + a(t)\, y = 0, \tag{2.2}$$

and the corresponding Floquet multipliers will be denoted by $\lambda_i = \lambda_i[a], i = 1, 2$. If these multipliers do not lie on $\mathbf{S}^1$, the unit circle in $\mathbb{C}$, then $y = 0$ is not stable for (2.1).

THEOREM. *Assume the following.*

(i) *The equation* (2.2) *is stable*;

(ii) $c \geq 0$ *or* $c \leq 0$.

Then $y = 0$ *is a stable solution of* (2.1).

(The proof will be given in §3.)

Remark. This result extends Theorem 2 in [9]. In addition to (i) and (ii), the result in [9] assumes that $n = 2, a < (\pi/2T)^2$ and (2.2) is elliptic.

Following [12], a criterion of stability for the Hill's equation (2.2) is expressed as a finite number of inequalities of the form

$$F_i(a) < 0, \qquad i = 1, \ldots, N,$$

such that (2.2) is stable when they hold. Here $F_i = F_i(a)$ are functionals that are continuous with respect to the uniform topology. Examples of stability criteria for (2.2) are

$$\text{(Lyapunov)} \qquad F_1(a) = -\min_t a(t), \qquad F_2(a) = -4 + T\int_0^T a(t)\,dt,$$

(Zukovskii)

$$F_1(a) = \max_t \left[(p\pi/T)^2 - a(t)\right], F_2(a) = \max_t \left[a(t) - ((p+1)\pi/T)^2\right], \qquad (p \in \mathbb{N}).$$

(See [4], [6], and [12] for many other criteria.)

Every stability criterion for the Hill's equation together with the assumption (ii) produces a stability criterion for (2.1). An example is the following nonlinear version of Lyapunov's criterion.

COROLLARY. *Assume that*

(i′) $a > 0, T\int_0^T a(t)\,dt < 4$;

(ii) $c \geq 0$ *or* $c \leq 0$.

Then $y = 0$ is stable for (2.1).

3. Remarks on the stability of fixed points and proof of the main result. In this section we shall obtain a proof of the theorem of §2. This proof will be based on the theory of stability of fixed points of area-preserving maps in the plane. The basic result in this field is Moser's twist theorem and we shall present a consequence of it following the ideas and results in [10], [11], and [1].

3.1. Stability of fixed points of area-preserving maps. Let $F : \Omega \subset \mathbb{C} \to \mathbb{C}$ be an area-preserving map defined in an open neighborhood of the origin and such that $z = 0$ is a fixed point. It is assumed that F is smooth (C^∞ in the real sense) and, for convenience, the complex notation $F = F(z, \bar{z})$ is used.

The following lemma gives an expression for the first nonlinear jet of F.

LEMMA 3.1. *Assume that for some* $m \geq 3$,

$$F(z,\bar{z}) = \lambda z + O\left(|z|^{m-1}\right), \qquad z \to 0 \quad (\lambda \in \mathbf{S}^1).$$

Then there exists $H = H(z,\bar{z})$, *a real valued homogeneous polynomial of degree* m *such that*

$$F(z,\bar{z}) = \lambda\left[z + 2i\partial_{\bar{z}}H(z,\bar{z}) + O\left(|z|^m\right)\right], \qquad z \to 0. \tag{3.1}$$

Proof. Consider the expansion $\bar{\lambda}F(z,\bar{z}) = z + h(z,\bar{z}) + O(|z|^m)$, where h is a homogeneous polynomial of degree $m-1$. Since F is area-preserving,

$$1 = |\partial_z F|^2 - |\partial_{\bar{z}}F|^2 = 1 + \partial_z h + \overline{\partial_{\bar{z}}h} + O\left(|z|^{m-1}\right), \quad \text{implying } \partial_z h + \overline{\partial_{\bar{z}}h} = 0.$$

The Euler's theorem for homogeneous functions implies that $z\partial_z h + \bar{z}\partial_{\bar{z}}h = (m-1)h$. Define $H(z,\bar{z}) = \mathrm{Im}[\bar{z}h(z,\bar{z})/m]$. The previous identities show that $2i\partial_{\bar{z}}H = h$, concluding the proof.

Remark. The same argument appears in the proof of Theorem 1 in [1].

Assume now that F is in the conditions of Lemma 3.1 with $m = 2n, n \geq 2$. The polynomial H given by the lemma can be expressed in the form

$$H(z,\bar{z}) = \beta |z|^{2n} + \sum_{k=0}^{n-1} \{\alpha_k z^k \bar{z}^{2n-k} + \bar{\alpha}_k \bar{z}^k z^{2n-k}\}, \tag{3.2}$$

where $\beta \in \mathbb{R}; \alpha_0, \ldots, \alpha_{n-1} \in \mathbb{C}$.

PROPOSITION 3.2. *Assume that $\lambda \in \mathbf{S}^1, F$ satisfies* (3.1) *with $m = 2n$ for some $n \geq 2, H$ is given by* (3.2) *and one of the following conditions hold:*

(C1) $\lambda^{2p} \neq 1$ *for each $p = 1, \ldots, n$ and $\beta \neq 0$;*

(C2) $\lambda^{2p} = 1$ *for some $p = 1, \ldots, n$ and $H^{\#}(z,\bar{z}) \neq 0$, for each $z \in \mathbb{C} - \{0\}$, where*

$$H^{\#}(z,\bar{z}) = (1/2p) \sum_{r=0}^{2p-1} H\left(\lambda^r z, \bar{\lambda}^r \bar{z}\right).$$

Then $z = 0$ is stable with respect to F.

Proof. Assume first that (C1) holds. Then it follows from the theory of normal forms for symplectic maps (see [10, §23]) that there exists a symplectic diffeomorphism ψ defined in a neighborhood of the origin, with $\psi(0) = 0$ and such that the conjugate map $G = \psi^{-1} \circ F \circ \psi$ has the expansion

$$G(z,\bar{z}) = \lambda \left[z + i\gamma |z|^{2n-2} z + O\left(|z|^{2n}\right)\right], \qquad z \to 0,$$

where $\gamma = 2n\beta$. In consequence $\gamma \neq 0$ and the conclusion follows from [10, §34]. It must be remarked that [10] assumes that F is real analytic, however the same arguments are valid for F sufficiently smooth applying the corresponding version of the twist theorem in [8].

If (C2) holds we compute the Taylor expansion of the iterated F^{2p} and obtain

$$F^{2p}(z,\bar{z}) = z + 2i \sum_{r=0}^{2p-1} \bar{\lambda}^r \left(\partial_{\bar{z}} H\right)\left(\lambda^r z, \bar{\lambda}^r \bar{z}\right) + O\left(|z|^{2n}\right), \qquad z \to 0.$$

In consequence,

$$F^{2p}(z,\bar{z}) = z + 4pi\partial_{\bar{z}} H^{\#}(z,\bar{z}) + \cdots$$

and we think of $z = 0$ as a parabolic fixed point of F^{2p} to apply Theorem 1 in [1]. The assumptions (b) and (c) of that theorem are verified since F^{2p} is area-preserving while (a) is equivalent in our setting to $H^{\#}(z,\bar{z}) \neq 0, \forall z \in \mathbb{C} - \{0\}$. In consequence $z = 0$ is surrounded by F^{2p}-invariant curves and is therefore stable. As remarked in [1] the conclusion also follows from the results in [11].

3.2. Proof of the main result. It will proceed in two steps. First, it is assumed that the linearized equation satisfies an additional condition and later it is shown that it can be removed. The condition is as follows: Let $\Psi(t)$ be the solution of (2.2) with initial conditions $\Psi(0) = 1, \Psi'(0) = i$, then

$$\Psi(t+T) = \bar{\lambda}\Psi(t) \quad \forall t \in \mathbb{R} \quad (\lambda = \lambda_i[a], i = 1 \text{ or } 2). \tag{3.3}$$

Step 1. It is assumed that (3.3) holds.

Let $y(t;z,\bar{z})$ denote the solution of (2.1) satisfying $y(0;z,\bar{z}) = \operatorname{Re} z, y'(0;z,\bar{z}) = \operatorname{Im} z$, and consider the Poincaré map

$$P(z,\bar{z}) = y(T;z,\bar{z}) + iy'(T;z,\bar{z}).$$

This map is defined and smooth in a neighborhood of the origin and is area-preserving. The stability of the trivial solution of (2.1) is equivalent to the stability of $z = 0$ as a fixed point of P. We shall apply Proposition 3.2 with $F = P$ and for this purpose we need to compute the corresponding polynomial H. First we shall compute an expansion of P using the following fact: "Let $y = y(t)$ be the solution of

$$y'' + a(t)y + f(t) = 0, \quad y(0) = \operatorname{Re} z, \quad y'(0) = \operatorname{Im} z \quad (z \in \mathbb{C}, f \in C[0,T]).$$

Then $y(T) + iy'(T) = \lambda\{z - i\int_0^T f(t)\Psi(t)\,dt\}$." (It is a consequence of the formula of variation of constants together with (3.3).) Applying the previous statement to (2.1) one obtains

$$P(z,\bar{z}) = \lambda\left\{z - i\int_0^T \left[c(t)\,y(t;z,\bar{z})^{2n-1} + d(t,y(t;z,\bar{z}))\right]\Psi(t)\,dt\right\}.$$

On the other hand, the theorem of differentiability with respect to the initial conditions implies that, uniformly in $t \in [0,T]$,

$$y(t;z,\bar{z}) = [\bar{\Psi}(t)z + \Psi(t)\bar{z}]/2 + O(|z|^2), \qquad y'(t;z,\bar{z}) = [\bar{\Psi}'(t)z + \Psi'(t)\bar{z}]/2 + O\left(|z|^2\right).$$

A combination of both formulas produces the expansion

(3.4)
$$P(z,\bar{z}) = \lambda\left\{z - (i/2^{2n-1})\int_0^T c(t)\left[\bar{\Psi}(t)z + \Psi(t)\bar{z}\right]^{2n-1}\Psi(t)\,dt\right\} + O\left(|z|^{2n}\right).$$

Then P satisfies (3.1) with H given by

(3.5)
$$\begin{aligned} H(z,\bar{z}) &= -(1/2^{2n})\int_0^T c(t)[\bar{\Psi}(t)z + \Psi(t)\bar{z}]^{2n}/2n\,dt \\ &= -\int_0^T c(t)[\operatorname{Re}(\bar{\Psi}(t)z)]^{2n}/2n\,dt. \end{aligned}$$

The coefficient β is given in this case by

$$\beta = -(1/2^{2n+1}n)\binom{2n}{n}\int_0^T c(s)\,|\Psi(s)|^{2n}\,ds.$$

For each $z \neq 0$, the function $t \to \operatorname{Re}(\bar{\Psi}(t)z)$ is a nontrivial real-valued solution of (2.2) and it can only vanish on a discrete set. In particular, $\operatorname{Re}\Psi(t)$, $\operatorname{Im}\Psi(t)$ are linearly independent solutions, implying $|\Psi(t)| \neq 0\ \forall t \in \mathbb{R}$. Assume for instance that $c \geq 0$. Then $\beta < 0$ and $H(z,\bar{z}) < 0\ \forall z \in \mathbb{C} - \{0\}$. When $\lambda^{2p} \neq 1$ for each $p = 1,\ldots,n$, (C1)

holds. If $\lambda^{2p} = 1$ for some $p = 1, \dots, n$ the definition of $H^{\#}$ and the negativity of H imply that $H^{\#}(z, \bar{z}) < 0 \ \forall z \in \mathbb{C} - \{0\}$, so that (C2) holds.

Step 2 (the general case). For each $t_0 \in \mathbb{R}$ let $\Phi(t, t_0)$ denote the matrix solution of $Y' = A(t)Y, Y(t_0) = I$, where

$$A(t) = \begin{pmatrix} 0 & 1 \\ -a(t) & 0 \end{pmatrix}.$$

The matrix $\Phi(t_0 + T, t_0)$ is a monodromy matrix of (2.2) and the condition (3.3) is equivalent to $\Phi(T, 0) = R[\theta]$, where $\lambda = e^{i\theta}$ and $R[\theta]$ is the rotation of angle θ given by

$$R[\theta] = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

If $\lambda = \pm 1$, (3.3) always holds. In fact, since (2.2) is stable, the monodromy matrix must be $\pm I$. If $\lambda \neq \pm 1$ it follows from Proposition 7 in [9] that there exists t_0 and $\alpha > 0$ such that

$$\Phi(t_0 + T, t_0) = D_\alpha R[\theta] D_\alpha^{-1}, \tag{3.6}$$

where $D_\alpha = \operatorname{diag}(\alpha, \alpha^{-1})$ and $e^{i\theta}$ is one of the Floquet multipliers of (2.2). The change of the independent variable $\tau = (t - t_0)/\alpha^2$ reduces equations (2.2) and (2.1) to the corresponding equations

$$(d^2/d\tau^2)\, y + a^*(\tau)\, y = 0 \quad \text{and} \quad (d^2/d\tau^2)\, y + a^*(\tau)\, y + c^*(\tau)\, y^{2n-1} + d^*(\tau, y) = 0,$$

with $a^*(\tau) = \alpha^4 a(\alpha^2\tau + t_0), c^*(\tau) = \alpha^4 c(\alpha^2\tau + t_0), d^*(\tau, y) = \alpha^4 d(\alpha^2\tau + t_0, y)$. These equations are periodic with period $T^* = T/\alpha^2$ and (3.6) implies that the monodromy matrix corresponding to the transformed equation satisfies $\Phi^*(T^*, 0) = R[\theta]$, so that (3.3) holds for the transformed equation and Step 1 can now be applied.

4. Examples of instability. In this section we present an example showing that the equilibrium can be unstable if the hypothesis (ii) of the main theorem in §2 does not hold. The construction of this example will follow after stating a result on unstable fixed points of area-preserving maps.

4.1. Roots of the unity and unstable fixed points. As in §3.1, we denote by $F : \Omega \subset \mathbb{C} \to \mathbb{C}$ a smooth area-preserving map such that $z = 0$ is a fixed point. It is assumed that F has an expansion of the kind

$$F(z, \bar{z}) = \lambda\left[z + 2i\partial_{\bar{z}} H(z, \bar{z}) + O(|z|^m)\right], \qquad z \to 0,$$

where H is a real-valued homogeneous polynomial of degree $m, m \geq 3$, and λ is a root of the unity; $\lambda^n = 1$ for some $n \geq 1$.

PROPOSITION 4.1. *In the previous setting define the polynomial*

$$H^{\#}(z, \bar{z}) = (1/n) \sum_{r=1}^{n} H\left(\lambda^r z, \bar{\lambda}^r \bar{z}\right)$$

and assume that $H^{\#}$ changes the sign; i.e., there exist $z_1, z_2 \in \mathbb{C}$ such that $H^{\#}(z_1, \bar{z}_1) < 0, H^{\#}(z_2, \bar{z}_2) > 0$. Then $z = 0$ is not Lyapunov stable.

Remark 1. The application of this result to the case $n = m = 3$ leads to the classical criterion of instability of Levi-Civita at the third root of the unity (see [10, §31]).

Remark 2. It is enough to prove the result in the parabolic case ($\lambda = 1$). The general case is reduced to it using the iterated F^n. This reduction repeats some of the arguments in the proof of Proposition 3.2. In the parabolic case there are several results on instability in [5], [7], and [11]. It is not difficult to verify that Proposition 4.1 follows from the result in [11] in the analytic case. However the proof of instability in [11] has not many details and we shall give an independent proof valid also for the nonanalytic case.

The proof will use the following result.

LEMMA 4.2. *Let $F : \Omega \subset \mathbb{C} \to F(\Omega) \subset \mathbb{C}$ be an area-preserving homeomorphism with $F(0) = 0$. Assume that there exist positive constants α, δ with $\alpha < \pi$ such that*

$$F\left(\bar{K}_\alpha \cap D_\delta\right) \subset K_\alpha, \tag{4.1}$$

where $K_\alpha = \{z \in \mathbb{C}/z = re^{i\theta}, r \geq 0, |\theta| < \alpha\}, D_\delta = \{z \in \mathbb{C}/|z| < \delta\}$. Then $z = 0$ is not stable.

Proof. By a contradiction argument assume that $z = 0$ is stable. It follows from [10, §25] that there exists an invariant open set G with $0 \in G, \bar{G} \subset D_\delta$. The Poincaré's recurrence theorem (see, for instance, [10, §37]) implies that almost every point in G is recurrent. The assumption (4.1) and the invariance of G imply that $F(\bar{K}_\alpha \cap \bar{G}) \subset K_\alpha \cap \bar{G}$. It is now easy to deduce that the set $\Delta = [(K_\alpha - \{0\}) \cap G] - F(\bar{K}_\alpha \cap \bar{G})$ is open and nonempty, in particular it has positive measure. In consequence the points in Δ cannot be recurrent, a contradiction with the recurrence theorem.

Proof of the Proposition 4.1. It will be assumed that $\lambda = 1$. The function $\varphi(\theta) = H(e^{i\theta}, e^{-i\theta})$ is real analytic, 2π-periodic and changes the sign. In consequence there exist $\theta^* \in [0, 2\pi]$ and $\alpha > 0$ such that $\varphi(\theta^*) = 0$ and $\varphi'(\theta) < 0$ if $0 < |\theta - \theta^*| \leq \alpha$. In what follows it will be assumed that $\theta^* = 0$ (otherwise F can be replaced by the conjugate map $F_1(z) = e^{-i\theta^*} F(e^{i\theta^*} z)$).

We now consider the hamiltonian system $z' = 2i\partial_{\bar{z}} H(z, \bar{z})$ and denote by $Z_t(z, \bar{z})$ the solution satisfying $Z_0(z, \bar{z}) = z$. It is well known that the corresponding 1-time map approximates well the map F (see for instance [3, p. 314]). Actually,

$$|F(z, \bar{z}) - Z_1(z, \bar{z})| = O\left(|z|^m\right), \qquad z \to 0. \tag{4.2}$$

The canonical change of variables $z = (2\rho)^{1/2} e^{i\theta}$ transforms the hamiltonian system into

$$\theta' = (m/2)(2\rho)^{m/2-1} \varphi(\theta), \qquad \rho' = -(2\rho)^{m/2} \varphi'(\theta), \quad (\theta, \rho) \in \mathbb{R}/2\pi\mathbb{Z} \times (0, \infty).$$

It follows from these equations that K_α is positively invariant with respect to the hamiltonian flow. To be more precise, consider the metric of the group $\mathbb{R}/2\pi\mathbb{Z}$ given by

$$\|\theta\| = \inf\left\{|\tilde{\theta} + 2\pi p|; p \in \mathbb{Z}\right\}, \quad \theta = \tilde{\theta} + 2\pi\mathbb{Z}, \quad \tilde{\theta} \in \mathbb{R},$$

and denote by Arg: $\mathbb{C} - \{0\} \to \mathbb{R}/2\pi\mathbb{Z}$ the argument function. Then

(4.3)

$$\|\text{Arg } Z_1(z, \bar{z})\| \leq \max\left[\alpha/2, \|\text{Arg } z\| - (mk/2)|z|^{m-2}\right], z \in \bar{K}_\alpha \text{ sufficiently small.}$$

Here $k = \min[|\varphi(\theta)|; \alpha/2 \le |\theta| \le \alpha]$. We remark that there exist $L > 0$ such that

$$\|\mathrm{Arg}\,(z_1) - \mathrm{Arg}\,(z_2)\| \le L|z_1|^{-1}|z_1 - z_2| \quad \forall z_1, z_2 \in \mathbb{C} - \{0\}.$$

Combining this inequality with (4.2) and (4.3) we obtain

$$\begin{aligned}\|\mathrm{Arg}\ F(z,\bar z)\| &\le \|\mathrm{Arg}\ Z_1(z,\bar z)\| + \|\mathrm{Arg}\ F(z,\bar z) - \mathrm{Arg}\ Z_1(z,\bar z)\| \\ &\le \max\left[\alpha/2, \|\mathrm{Arg}\, z\| - (mk/2)\,|z|^{m-2}\right] \\ &\quad + O\left(|z|^{m-1}\right) < \alpha \quad \text{if } z \in \bar K_\alpha \text{ sufficiently small.}\end{aligned}$$

This estimate shows that (4.1) holds for some small δ and the proof finishes with the application of Lemma 4.2.

4.2. An example of unstable equilibrium. Consider the equation

$$y'' + y + c_n(t)\, y^{2n-1} = 0, \qquad n \ge 2 \tag{4.4}$$

and assume that c_n is continuous and periodic with period $T = \pi/n$. In addition

$$\int_0^{\pi/n} c_n(t)\, e^{ipt}\, dt = 0, \quad p = 0, 1, \dots, 2n-1, \qquad \int_0^{\pi/n} c_n(t)\, e^{i2nt}\, dt \ne 0. \tag{4.5}$$

We shall prove that $y = 0$ is unstable. Remark that the condition (i) of the theorem holds, but c_n has to change the sign, so that (ii) fails.

The linearized equation is $y'' + y = 0$ and, for the period π/n, the Floquet multipliers are $e^{\pm i\pi/n}$. Following the notation of §3 we have that $\Psi(t) = e^{it}$ and (3.3) holds with $\lambda = e^{-i\pi/n}$. The Poincaré map has an expansion of the type (3.1) with $m = 2n$ and, according to (3.5) and (4.5),

$$H(z,\bar z) = -(1/2^{2n}n)\,\mathrm{Re}\,[\Gamma z^{2n}] \quad \text{with } \Gamma = \int_0^{\pi/n} c_n(t)\, e^{-i2nt}\, dt.$$

Since λ is a root of the unity of order $2n$, $H^{\#} = H$ and Proposition 4.1 can be applied.

Acknowledgment. I thank Professor Martínez-Amores for several conversations on the presentation of this paper.

REFERENCES

[1] D. Aharonov and U. Elias, *Invariant curves around a parabolic fixed point at infinity*, Ergodic Theory Dynamical Systems, 10 (1990), pp. 209–229.

[2] V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics*, Addison-Wesley, Redwood City, CA, 1989.

[3] D. K. Arrowsmith and C. M. Place, *An Introduction to Dynamical Systems*, Cambridge Univ. Press, Cambridge, UK, 1990.

[4] L. Cesari, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Springer-Verlag, New York, Berlin, 1971.

[5] T. Levi-Civita, *Sopra alcuni criteri di instabilità*, Ann. Mat. Pura Appl., 5 (1901), pp. 221–307.

[6] W. Magnus and S. Winkler, *Hill's Equation*, Dover, New York, 1979.

[7] R. McGehee, *A stable manifold theorem for degenerate fixed points with applications to celestial mechanics*, J. Differential Equations, 14 (1973), pp. 70–88.

[8] J. MOSER, *On invariant curves of area-preserving mappings of an annulus*, Nachr. Akad. Wiss. Göttingen, Math.-Phys., K1 (1962), pp. 1–10.

[9] R. ORTEGA, *The twist coefficient of periodic solutions of a time-dependent Newton's equation*, J. Dynamics Differential Equations, 4 (1992), pp. 651–665.

[10] C. L. SIEGEL AND J. K. MOSER, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, Berlin, 1971.

[11] C. SIMO, *Stability of degenerate fixed points of analytic area preserving mappings*, Astérisque, 98–99 (1982), pp. 184–194.

[12] V. M. STARZINSKII, *Survey of Works on Conditions of Stability of the Trivial Solution of a System of Linear Differential Equations with Periodic Coefficients*, in Amer. Math. Soc. Transl. Ser. 2, Vol. 1., Providence, RI, 1955.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1402–1411, September 1994

009

MATCHED EXPANSION SOLUTIONS OF THE FIRST-ORDER TURNING POINT PROBLEM *

L. A. SKINNER†

Abstract. Uniformly valid asymptotic solutions for second-order linear differential equations with first-order turning points are obtained by the method of matched asymptotic expansions. A key feature of the analysis is that the high frequency and strong exponential behavior of these solutions is factored out of the matching processes. This produces a set of essentially elementary boundary layer problems. A variation of the Langer expansion theorem for first-order turning points is independently established and used in verifying the formal calculations. The results emphasize the basic WKB structure of turning point asymptotics. A new property of Airy functions is also involved.

Key words. matched asymptotic expansions, turning points

AMS subject classification. 34E20

1. Introduction.

This paper is concerned with uniformly valid asymptotic expansions as $\varepsilon \to 0^+$ of solutions to the differential equation

$$\varepsilon^2 y'' + [xa(x) + \varepsilon b(x) + \varepsilon^2 c(x,\varepsilon)]\, y = 0, \tag{1.1}$$

where $a(x), b(x) \in C^\infty[-1,1], c(x,\varepsilon) \in C^\infty([-1,1] \times [0,1])$ and $a(x) > 0$. Without loss of generality we also assume $a(0) = 1$. This is the same differential equation that Langer treated in his 1949 paper [2], which was the culmination of his fundamental work on first-order turning point theory. It is more general than the differential equation subsequently taken up by Olver [4], who had $b(x) = 0$ and $c(x,\varepsilon)$ depending only on x, for his extension of Langer's results to the complex plane. Olver has since suggested an extension of his work to deal with the full equation (1.1), [5, pp. 426–429], but the theory for this has not been completely worked out.

The purpose of the present paper is to establish some new results for (1.1) based on the method of matched asymptotic expansions. Part of the motivation for this work is that the results of Langer and Olver tend to obscure the basic WKB structure of asymptotic solutions to (1.1), and also their basic boundary layer structure. A more fundamental point is that we should expect a wider range of applicability to the uniform solution of turning point problems in general from the method of matched asymptotic expansions. Although much has been done on matching methods for turning point problems in recent years, and much of it is summarized in the monograph by Wasow [7], no theory of uniformly valid composite expansions for these problems has been developed.

In the case of (1.1), we know from WKB theory [5], [7] that for $x > 0$ there is an oscillatory solution of the form

$$y(x,\varepsilon) = \mu(\varepsilon)\, e^{i\nu s(x)} [p(x) + O(\varepsilon)] \tag{1.2}$$

where $\nu = \varepsilon^{-1}$ and, from substituting into (1.1),

$$s(x) = \int_0^x [|t|\, a(t)]^{1/2}\, dt. \tag{1.3}$$

* Received by the editors November 9, 1992; accepted for publication July 7, 1993.

† Department of Mathematical Sciences, University of Wisconsin, Milwaukee, Wisconsin 53201.

Substitution also yields $p(x) = h(x)\exp[ir(x)]$ where $h(x) = [|x|a(x)]^{-1/4}$ and

$$r(x) = \frac{1}{2}\int_0^x b(t)\,h^2(t)\,dt. \tag{1.4}$$

The reason for absolute value signs will be seen later. In addition to (1.2), upon introducing $X = \nu^{2/3}x$ in (1.1) it is apparent (formally) that $y(\varepsilon^{2/3}X, \varepsilon) = P(X) + O(\varepsilon^{1/3})$, where $P(-X)$ satisfies Airy's equation. The solution that matches with (1.2) is

$$P(X) = \pi^{1/2}e^{i\pi/4}\left[\mathrm{Ai}(-X) - i\,\mathrm{Bi}(-X)\right], \tag{1.5}$$

provided $\mu(\varepsilon) = \varepsilon^{1/6}$. Indeed, if $x = \varepsilon^k$ and $\frac{2}{5} < k < \frac{2}{3}$, then both $P(\nu^{2/3}x)$ and

$$\varepsilon^{1/6}p(x)\,e^{i\nu s(x)} = \varepsilon^{1/6}x^{-1/4}e^{i\nu t}\left[1 + o(1)\right] \tag{1.6}$$

as $\varepsilon \to 0^+$, where $t = (2/3)x^{3/2}$. Thus, as indicated in [1, pp. 163–168], the composite function

$$q(x,\varepsilon) = \varepsilon^{1/6}p(x)\,e^{i\nu s(x)} + P\left(\nu^{2/3}x\right) - \varepsilon^{1/6}x^{-1/4}e^{i\nu t} \tag{1.7}$$

should be a uniform approximation to $y(x,\varepsilon)$. A reasonable conjecture is

$$y(x,\varepsilon) = q(x,\varepsilon) + O\left(\varepsilon^{1/3}\right) \tag{1.8}$$

for $0 \le x \le 1$; however, this has never been proved.

Rather than pursue (1.8), the idea for dealing with (1.1) in this paper is to set

$$y(x,\varepsilon) = e^{i\nu s(x)}z(x,\varepsilon) \tag{1.9}$$

and solve instead for $z(x,\varepsilon)$. It turns out that $z(x,\varepsilon)$, unlike $y(x,\varepsilon)$, is an essentially elementary, although complex valued, boundary layer function. The differential equation for $z(x,\varepsilon)$ is

$$\varepsilon z'' + 2is'(x)\,z' + \left[b(x) + is''(x) + \varepsilon c(x,\varepsilon)\right]z = 0. \tag{1.10}$$

Several terms of a formal matched asymptotic expansion solution for $z(x,\varepsilon)$ are derived in the next section. In §3 we prove that the resulting formal composite expansion is a uniformly valid solution of (1.10) for $0 \le x \le 1$. A second solution of (1.1) for $0 \le x \le 1$ and analogous exponential solutions for $-1 \le x \le 0$ are given in §4, together with connection formulas.

2. Formal solution. It is appropriate, in view of our preliminary calculations, to write the N-term outer expansion for $z(x,\varepsilon)$ as

$$O_N z(x,\varepsilon) = \varepsilon^{1/6}\sum_{n=0}^{N-1}\varepsilon^{n/3}z_n(x). \tag{2.1}$$

That is, we assume (temporarily) a solution of (1.10) with an asymptotic expansion of the form (2.1) for $x > 0$. From (1.2), $z_0(x) = p(x)$. Also, (1.10) indicates $z_n(x) = 0$ unless $n/3$ is an integer. Thus $z_1(x) = z_2(x) = 0$.

The inner expansion for $z(x,\varepsilon)$ is the expansion of $z(\varepsilon^{2/3}X,\varepsilon)$ in powers of $\varepsilon^{1/3}$. Denoting the first N terms of this expansion by

$$I_N z(x,\varepsilon) = \sum_{n=0}^{N-1} \varepsilon^{n/3} Z_n(X), \tag{2.2}$$

and, putting $X = \nu^{2/3}x$ in (1.10), we find

$$Z_0'' + 2iX^{1/2}Z_0' + \tfrac{1}{2}iX^{-1/2}Z_0 = 0. \tag{2.3}$$

If we write this as $\mathcal{L}Z_0 = 0$, then in addition

$$\mathcal{L}Z_1 = -b_0 Z_0(X), \tag{2.4}$$

$$\mathcal{L}Z_2 = -ia_1 X^{1/2}\left[\tfrac{3}{4}Z_0(X) + XZ_0'(X)\right] - b_0 Z_1(X), \tag{2.5}$$

where $b_0 = b(0), a_1 = a'(0)$.

In terms of $g(X) = P(X)\exp[-(2i/3)X^{3/2}]$, where $P(X)$ is given by (1.5), the general solution of (2.3) is

$$Z_0(X) = c_{01}g(X) + c_{02}\bar{g}(X)\exp\left[-(4i/3)X^{3/2}\right], \tag{2.6}$$

where $\bar{g}(X)$ is the complex conjugate of $g(X)$. Also, from well-known expansions for the Airy functions,

$$g(X) \sim X^{1/4}\sum_{m=0}^{\infty} g_m X^{-3m/2} \tag{2.7}$$

as $X \to \infty$, where $g_0 = 1$ and $g_1 = -5i/48$. Thus for $I_1 z(x,\varepsilon) = Z_0(\nu^{2/3}x)$ to match with $O_1 z(x,\varepsilon) = \varepsilon^{1/6}p(x)$, it must be that $c_{02} = 0$. Indeed then,

$$O_1 I_1 z(x,\varepsilon) = I_1 O_1 z(x,\varepsilon) = \varepsilon^{1/6}x^{-1/4}, \tag{2.8}$$

provided $c_{01} = 1$.

Let $C_N = O_N + I_N - O_N I_N$. We prove in §3 that $z(x,\varepsilon) = C_N z(x,\varepsilon) + O(\varepsilon^{N/3})$ uniformly for $0 \le x \le 1$. Presently, with $N = 1$, this gives us, in place of (1.8), the two term expansion

$$z(x,\varepsilon) = g\left(\nu^{2/3}x\right) + \varepsilon^{1/6}\left[p(x) - x^{-1/4}\right] + O\left(\varepsilon^{1/3}\right). \tag{2.9}$$

To determine higher-order terms of (2.2) it is helpful to note that

$$\mathcal{L}X^m g(X) = mX^m\left[(m-1)X^{-2} + 2iX^{-1/2}\right]g(X) + 2mX^{m-1}g'(X), \tag{2.10}$$

$$\begin{aligned}\mathcal{L}X^m g'(X) &= \tfrac{1}{4}i(1-4m)X^{m-3/2}g(X)\\ &\quad - X^m\left[i(1-4m)X^{-1/2} - m(m-1)X^{-2}\right]g'(X).\end{aligned} \tag{2.11}$$

Thus we find

$$Z_1(X) = c_{11}g(X) + c_{12}\bar{g}(X)\exp\left[-(4i/3)X^{3/2}\right] + b_0\left[iX^{1/2}g(X) + g'(X)\right], \tag{2.12}$$

and as with $Z_0(X)$, we clearly need $c_{12} = 0$ in order to match with (2.1). It follows that

$$O_2I_2z(x,\varepsilon) = \varepsilon^{1/6}\left(x^{-1/4} + ib_0x^{1/4}\right) + c_{11}\varepsilon^{1/2}x^{-1/4}. \tag{2.13}$$

On the other hand,

$$p(x) = x^{-1/4}\left[1 + ib_0x^{1/2} - k_1x + O\left(x^{3/2}\right)\right], \tag{2.14}$$

where $k_1 = (a_1 + 2b_0^2)/4$, and therefore

$$I_2O_2z(x,\varepsilon) = X^{-1/4} + ib_0\varepsilon^{1/3}X^{1/4}. \tag{2.15}$$

Hence $c_{11} = 0$ too, and with $X = \nu^{2/3}x$,

$$C_2z(x,\varepsilon) = C_1z(x,\varepsilon) + b_0\varepsilon^{1/3}\left[iX^{1/2}g(X) + g'(X) - iX^{1/4}\right]. \tag{2.16}$$

For $Z_2(X)$ we now have

$$\mathcal{L}Z_2 = -ia_1\left[\tfrac{3}{4}X^{1/2}g(X) + X^{3/4}g'(X)\right] - b_0^2\left[iX^{1/2}g(X) + g'(X)\right]. \tag{2.17}$$

As above, a particular solution of (2.17) is readily found and matching again eliminates the homogeneous solution terms. The result is

$$Z_2(X) = \tfrac{1}{5}a_1\left[-Xg(X) + X^2g'(X)\right] - \tfrac{1}{2}b_0^2Xg(X). \tag{2.18}$$

Thus

$$C_3z(x,\varepsilon) = C_2z(x,\varepsilon) + \varepsilon^{2/3}\left[Z_2(X) + k_1X^{3/4}\right]. \tag{2.19}$$

3. Confirmation. As in [6], we shall write $f(x) \in C^\infty[0,\infty]$ if both $f(x)$ and $f(1/x)$ are in $C^\infty[0,1]$. Thus, if $f(x,X) \in C^\infty([0,1]\times[1,\infty])$, then $f(x,X)$ has an asymptotic (at least) power series expansion about $(x,X) = (0,\infty)$. The coefficient of x^mX^{-n} in this expansion is $f^{[m,-n]}(0,\infty)$, where

$$f^{[m,-n]}(x,X) = \frac{1}{(m!)(n!)}\left(\frac{\partial}{\partial x}\right)^m\left(-X^2\frac{\partial}{\partial X}\right)^n f(x,X). \tag{3.1}$$

Our objective in this section is to verify the formal calculations of §1 by proving the following theorem.

THEOREM 1. *Let $a(x), b(x) \in C^\infty[-1,1]$, $c(x,\varepsilon) \in C^\infty([-1,1]\times[0,1])$ and assume $a(x) > 0$. Then* (1.1) *has a solution of the form* (1.9) *such that for $N \geq 1$,*

$$z(x,\varepsilon) = \sum_{n=0}^{N-1}\varepsilon^{n/3}\left[V_n\left(\nu^{2/3}x\right) + \varepsilon^{1/6}U_n(x)\right] + O\left(\varepsilon^{N/3}\right) \tag{3.2}$$

uniformly as $\varepsilon \to 0^+$ *for* $0 \le x \le 1$, *where* $X^{1/2}V_n(X^2) \in C^\infty[1,\infty]$ *and* $V_n(X^4)$, $x^{-1/2}U_n(x^2) \in C^\infty[0,1]$.

It follows from this theorem that $z(x,\varepsilon)$ has an outer expansion of the form (2.1) and an inner expansion of the form (2.2). It follows also that $I_nO_nz(x,\varepsilon) = O_nI_nz(x,\varepsilon)$ for any $n \ge 1$. Indeed the nth term of (3.2) is $C_{n+1}z(x,\varepsilon) - C_nz(x,\varepsilon)$. From here (existence) it is a routine matter to justify the computation of these expansions by substitution into the differential equation (1.10). Thus the first term of (3.2) is given by (2.9), the second by (2.16) and the third by (2.19).

To prove Theorem 1 we require two other results. The first one, from [6], confirms the validity of the method of matched asymptotic expansions for functions that have a certain structure. An elementary example is $f(t,T) = (1+t+T)^{-\pi}$.

THEOREM 2. *Let* $f(t,T) = T^{\alpha-1}\phi(t,T)$ *where* $0 < \alpha \le 1$. *If* $f(t,T) \in C^\infty([0,b] \times [0,1])$ *and* $\phi(t,T) \in C^\infty([0,b] \times [1,\infty])$, *then*

$$f(t,\mu t) = \sum_{n=0}^{N-1} \mu^{-n}\left[v_n(\mu t) + \mu^{\alpha-1}u_n(t)\right] + O(\mu^{-N}) \tag{3.3}$$

uniformly for $0 \le t \le b$ *as* $\mu \to \infty$, *where*

$$u_n(t) = t^{\alpha-n-1}\left[\phi^{[0,-n]}(t,\infty) - \sum_{m=0}^{n} \phi^{[m,-n]}(0,\infty)\, t^m\right], \tag{3.4}$$

$$v_m(T) = T^{\alpha+m-1}\left[\phi^{[m,0]}(0,T) - \sum_{n=0}^{m-1} \phi^{[m,-n]}(0,\infty)\, T^{-n}\right]. \tag{3.5}$$

Note that $t^{-\alpha}u_n(t) \in C^\infty[0,b]$ and $T^{1-\alpha}v_m(T) \in C^\infty[1,\infty]$. The plan is to use Theorem 2 to show that (3.2) is implied by the following result, which is a variation of one of the results mentioned earlier due to Langer [2].

THEOREM 3. *Under the same hypotheses as in Theorem* 1, *there exists a solution to* (1.1) *of the form* (1.9) *such that*

$$z(x,\varepsilon) = \sum_{n=0}^{N-1} \varepsilon^n\left[C_n(x)\, g\left(\nu^{2/3}\sigma(x)\right) + \varepsilon^{1/3}B_n(x)\, g'\left(\nu^{2/3}\sigma(x)\right)\right] + O(\varepsilon^N) \tag{3.6}$$

uniformly for $0 \le x \le 1$, *where* $\sigma(x) = [(3/2)s(x)]^{2/3}$, $C_n(x) = A_n(x) + i[\sigma(x)]^{1/2}B_n(x)$, *and* $A_n(x), B_n(x) \in C^\infty[0,1]$.

Integral formulas for the coefficients in (3.6) can be obtained by substituting into (1.1). They are given in [3]. Thus, $C_0(x) = [\sigma(x)]^{1/4}p(x)$, and $B_0(x) = [\sigma(x)]^{-1/4}h(x)\sin r(x)$. We do not get explicit formulas for higher order terms, however, although major simplifications occur if $b(x) = 0$, as in [4]. A simple proof of Theorem 3 is presented below.

Proof of Theorem 1. We will prove Theorem 1 by showing that the individual terms of (3.6) have asymptotic expansions of the same form as the one being claimed for $z(x,\varepsilon)$ itself. Pick m and let

$$f(t,T) = C_m(t^2)\, g\left(T^2\rho(t^2)\right), \tag{3.7}$$

where $\rho(t) = t^{-1}\sigma(t)$. Note that $\rho(t) > 0$ on $[0, 1]$. We also have $C_m(t^2) \in C^\infty[0,1]$ and, from (2.7), $T^{1/2}g(T^2) \in C^\infty[1,\infty]$. Hence this $f(t,T)$ satisfies the hypotheses of Theorem 2 with $\alpha = \frac{1}{2}$. It follows that

$$f\left(x^{1/2}, \nu^{1/3}x^{1/2}\right) = \sum_{n=0}^{N-1} \nu^{-n/3}\left[v_n\left(\nu^{1/3}x^{1/2}\right) + \nu^{-1/6}u_n\left(x^{1/2}\right)\right] + O\left(\nu^{-N/3}\right) \tag{3.8}$$

for $0 \le x \le 1$, where $v_n(X), u_n(x)$ are given by (3.4), (3.5). In other words

$$C_m(x)\, g\left(\nu^{2/3}\sigma(x)\right) = \sum_{n=0}^{N-1} \nu^{-n/3}\left[\hat{v}_n\left(\nu^{2/3}x\right) + \nu^{-1/6}\hat{u}_n(x)\right] + O\left(\nu^{-N/3}\right), \tag{3.9}$$

where $\hat{v}_n(X) = v_n(X^{1/2}), \hat{u}_n(x) = u_n(x^{1/2})$. Consequently $\hat{v}_n(X^4), x^{-1/2}\hat{u}_n(x^2) \in C^\infty[0,1]$ and $X^{1/2}v_n(X^2) \in C^\infty[1,\infty]$. The same argument is applicable to $B_m(x)g'(\nu^{2/3}\sigma(x))$. This completes the proof of Theorem 1.

Proof of Theorem 3. Following the previously mentioned discussion by Oliver [5, pp. 426–429], to prove Theorem 3 we note first that the Liouville transformation

$$w(\zeta,\varepsilon) = \left[\sigma'(x) + \varepsilon\tau'(x)\right]^{1/2} y(x,\varepsilon), \qquad \zeta = \sigma(x) + \varepsilon\tau(x), \tag{3.10}$$

where $\tau(x) = [\sigma(x)]^{-1/2}r(x)$, converts (1.1) into

$$\varepsilon^2 w'' + \left[\zeta + \varepsilon^2\gamma(\zeta,\varepsilon)\right] w = 0, \tag{3.11}$$

where $\gamma(\zeta,\varepsilon) \in C^\infty([0,\zeta_0] \times [0,\varepsilon_0])$ for some $\zeta_0, \varepsilon_0 > 0$. Since $\sigma'(x) > 0$ for $0 \le x \le 1$, we can choose ε_0 so that $\sigma'(x) + \varepsilon\tau'(x) > 0$ on $[0,1] \times [0,\varepsilon_0]$. The image of this rectangle is a trapezoid. Thus we let $\zeta_0 = \sigma(1)$ if $\tau(1) > 0$. Otherwise, $\zeta_0 = \sigma(1) + \varepsilon_0\tau(1)$. It is a straightforward matter, following [4] now, but taking into account that $\mathrm{Ai}(-Z), \mathrm{Bi}(-Z)$ are bounded for $Z \ge 0$, to show (3.11) has a solution $w(\zeta,\varepsilon)$ such that for $0 \le \zeta \le \zeta_0$,

$$w(\zeta,\varepsilon) = \sum_{n=0}^{N-1} \varepsilon^{2n}\left[\alpha_n(\zeta)\, P\left(\nu^{2/3}\zeta\right) + \varepsilon^{4/3}\beta_n(\zeta)P'\left(\nu^{2/3}\zeta\right)\right] + O\left(\varepsilon^{2N}\right). \tag{3.12}$$

The coefficients in this expansion, $\alpha_n(\zeta)$ and $\beta_n(\zeta)$, are in $C^\infty[0,\zeta_0]$. Therefore

$$\alpha_n(\sigma(x) + \varepsilon\tau(x)) = \sum_{m=0}^{M-1} \varepsilon^m\left[\tau^m(x)\,\alpha_n^{[m]}(\sigma(x))\right] + O\left(\varepsilon^M\right) \tag{3.13}$$

uniformly for $0 \le x \le x_0$, where $x_0 > 0$ is a lower bound for the inverse image of ζ_0, and the same is true with β_n in place of α_n. Thus if we could expand $P(\nu^{2/3}\sigma(x) + \nu^{-1/3}\tau(x))$ in terms of $P(\nu^{2/3}\sigma(x))$, and its derivative, (3.6) would be proved for $0 \le x \le x_0$. This possibility is covered by Theorem 4, which follows. The validity of (3.6) for $x_0 \le x \le 1$ is trivial, since (3.6) is asymptotically equivalent to the outer expansion (2.1) for $x_0 \le x \le 1$. Indeed, if (2.1) holds for $x_0 \le x < \infty$, then (3.6) holds for $0 \le x < \infty$. This completes our proof of Theorem 3.

THEOREM 4. *Let $V(z)$ be any solution of Airy's equation, $V''(z) = zV(z)$. For $\delta \ne 0$,*

$$V\left(\delta^{-2}\xi + \delta\eta\right) = f(\delta,\xi,\eta)\, V\left(\delta^{-2}\xi\right) + \delta g(\delta,\xi,\eta)\, V'\left(\delta^{-2}\xi\right), \tag{3.14}$$

where $f(\delta,\xi,\eta), g(\delta,\xi,\eta)$ are entire functions.

Proof. Any solution of Airy's equation is an entire function. Hence the Taylor expansion

$$V(z+w)=\sum_{n=0}^{\infty} w^n V^{[n]}(z) \tag{3.15}$$

converges for all $(z,w)\in\mathbb{C}\times\mathbb{C}$. Airy's equation also implies

$$V^{[n]}(z)=p_n(z)V(z)+q_n(z)V'(z) \tag{3.16}$$

where $p_n(z), q_n(z)$ are polynomials. In particular, $p_0(z)=1, q_0(z)=0$, and by differentiating (3.16) we find

$$(n+1)p_{n+1}(z)=p_n'(z)+zq_n(z), \qquad (n+1)q_{n+1}(z)=p_n(z)+q_n'(z). \tag{3.17}$$

Simple induction shows $\deg p_n(z)\le n/2$, $\deg q_n(z)\le (n-1)/2$. Thus we can write

$$p_n(z)=\sum_{k=0}^{\mathcal{K}(n)} p_{nk}z^k, \qquad q_n(z)=\sum_{k=0}^{\mathcal{K}(n-1)} q_{nk}z^k, \tag{3.18}$$

where $\mathcal{K}(n)=[n/2]$, and $p_{nk}, q_{nk}\ge 0$ for all n,k. Taking (temporarily) $V=\mathrm{Bi}$, we have

$$\mathrm{Bi}(z+w)=\sum_{n=0}^{\infty} w^n\left[p_n(z)\mathrm{Bi}(z)+q_n(z)\mathrm{Bi}(z)\right]. \tag{3.19}$$

Also, $\mathrm{Bi}(z), \mathrm{Bi}'(z)>0$ if $z\in\mathbb{R}^+$. Therefore,

$$\phi(z,w)=\sum_{n=0}^{\infty} w^n p_n(z), \qquad \psi(z,w)=\sum_{n=0}^{\infty} w^n q_n(z), \tag{3.20}$$

converge for all $(z,w)\in\mathbb{R}^+\times\mathbb{R}^+$, and thus they converge (absolutely) for all $(z,w)\in\mathbb{C}\times\mathbb{C}$. Furthermore, upon differentiating (3.20) with respect to w, it is clear, in view of (3.17), that differentiation with respect to z also is legitimate. Thus $\phi(z,w), \psi(z,w)$ are entire functions, and

$$V(z+w)=\phi(z,w)V(z)+\psi(z,w)V'(z). \tag{3.21}$$

Next we note that

$$\sum_{n=0}^{N} p_{2n,n}z^n\le\sum_{n=0}^{N} p_{2n}(z)\le\sum_{n=0}^{\infty} p_n(z)<\infty \tag{3.22}$$

for all $z\in\mathbb{R}^+$, and a comparable statement is true with $q_{2n+1,n}$ in place of $p_{2n,n}$. Hence

$$\mathcal{P}(z)=\sum_{n=0}^{\infty} p_{2n,n}z^n, \qquad \mathcal{Q}(z)=\sum_{n=0}^{\infty} q_{2n+1,n}z^n \tag{3.23}$$

are entire functions, too. Finally, it follows that

$$\lim_{\delta\to 0}\sum_{n=0}^{\infty}(\delta\eta)^n\sum_{k=0}^{\mathcal{K}(n)}p_{nk}(\delta^{-2}\xi)^k=\mathcal{P}(\xi\eta^2), \tag{3.24}$$

$$\lim_{\delta\to 0}\delta^{-1}\sum_{n=0}^{\infty}(\delta\eta)^n\sum_{k=0}^{\mathcal{K}(n-1)}q_{nk}(\delta^{-2}\xi)^k=\mathcal{Q}(\xi\eta^2). \tag{3.25}$$

Therefore

$$f(\delta,\xi,\eta)=\phi(\delta^{-2}\xi,\delta\eta),\qquad g(\delta,\xi,\eta)=\delta^{-1}\psi(\delta^{-2}\xi,\delta\eta), \tag{3.26}$$

with the above limits for $f(0,\xi,\eta), g(0,\xi,\eta)$ are entire functions. This completes the proof of Theorem 4.

4. Other solutions. In addition to the solution of (1.1) given by (1.9), which we now rename

$$y^{(1)}(x,\varepsilon)=e^{i\nu s(x)}z^{(1)}(x,\varepsilon), \tag{4.1}$$

there is a companion, linearly independent solution

$$y^{(3)}(x,\varepsilon)=e^{-i\nu s(x)}z^{(3)}(x,\varepsilon), \tag{4.2}$$

where $z^{(3)}(x,\varepsilon)$ has a uniform expansion of the same form as $z^{(1)}(x,\varepsilon)$. Thus

$$z^{(m)}(x,\varepsilon)=\sum_{n=0}^{N-1}\varepsilon^{n/3}\left[V_n^{(m)}\left(\nu^{2/3}x\right)+\varepsilon^{1/6}U_n^{(m)}(x)\right]+O\left(\varepsilon^{N/3}\right) \tag{4.3}$$

uniformly for $0\le x\le 1$, where $X^{1/2}V_n^{(m)}(X^2)\in C^\infty[1,\infty]$ and $V_n^{(m)}(X^2)$, $x^{-1/2}U_n^{(m)}(x^2)\in C^\infty[0,1]$. If we let $g^{(1)}(X)=g(X), g^{(3)}(X)=\bar{g}(X)$ and put

$$p^{(m)}(x)=h\left((-1)^{m+1}x\right)\exp\left[i^m r\left((-1)^{m+1}x\right)\right], \tag{4.4}$$

$$G^{(m)}(X)=X^{1/2}g^{(m)}(X)-X^{1/4}, \tag{4.5}$$

$$K^{(m)}(X)=Xg^{(m)}(X)-X^2g^{(m)'}(X)+\tfrac{5}{4}X^{3/4}, \tag{4.6}$$

then for $m=1$, from §2,

$$V_0^{(m)}(X)=g^{(m)}(X),\qquad U_0^{(m)}(x)=p^{(m)}(x)-x^{-1/4}, \tag{4.7}$$

$$V_1^{(m)}(X)=b_0\left[i^mG^{(m)}(X)-(-1)^m g^{(m)'}(X)\right],\qquad U_1^{(m)}(x)=0, \tag{4.8}$$

$$V_2^{(m)}(X)=(-1)^m\left[\tfrac{1}{5}a_1K^{(m)}(X)+\tfrac{1}{2}b_0^2X^{1/2}G^{(m)}(X)\right],\qquad U_2^{(m)}(x)=0. \tag{4.9}$$

The corresponding terms of (4.3) for $m = 3$ are also given by (4.7)–(4.9). Indeed, if all the coefficient functions in (1.1) are real, then $y^{(3)}(x, \varepsilon)$ is the complex conjugate of $y^{(1)}(x, \varepsilon)$.

For $-1 \leq x \leq 0$, the analogous solutions of (1.1) behave exponentially. They have the form

$$y^{(m)}(x, \varepsilon) = z^{(m)}(-x, \varepsilon) \exp[i^m \nu s(x)], \qquad m = 2, 4. \tag{4.10}$$

In fact, if we let

$$g^{(2)}(X) = \pi^{1/2} \mathrm{Bi}(X) \exp\left[-(2/3) X^{3/2}\right], \tag{4.11}$$

$$g^{(4)}(X) = 2\pi^{1/2} \mathrm{Ai}(X) \exp\left[(2/3) X^{3/2}\right], \tag{4.12}$$

then (4.3), with the definitions (4.4)–(4.9), also holds for $m = 2, 4$. This can be seen formally by paralleling the calculations of §2 or by substituting $-x$ for x and $i\varepsilon$ for ε in the results for $m = 1, 3$. The proof of (4.3) for $m = 2, 4$ is essentially the same as for $m = 1$. We have, after substituting $-x$ for x in (1.1), to deal with $\exp[(2\nu/3)\zeta^{3/2}]$ in place of the bounded function $\exp[(2i\nu/3)\zeta^{3/2}]$. But $\exp[(2\nu/3)\zeta^{3/2}] = \mathrm{Bi}(\nu^{2/3}\zeta)/g^{(2)}(\nu^{2/3}\zeta)$, and therefore

$$\exp\left\{(2\nu/3)\left[\sigma(x) + \nu^{-1/3}\tau(x)\right]^{3/2}\right\} = k(x, \nu)\, e^{\nu s(x)}, \tag{4.13}$$

where $k(x, \nu) = 0(1)$ for $0 \leq x \leq 1$, by Theorem 4.

Altogether now we have four solutions of (1.1). Any one of them is a linear combination of any two others. In particular,

$$y^{(m)}(x, \varepsilon) = c_{m1} y^{(1)}(x, \varepsilon) + c_{m3} y^{(3)}(x, \varepsilon), \qquad m = 2, 4. \tag{4.14}$$

To determine the coefficients in (4.14) we can use our inner expansion results. Indeed, from

$$y^{(m)}(0, \varepsilon) = g^{(m)}(0) - (-1)^m b_0 \varepsilon^{1/3} g^{(m)'}(0) + O\left(\varepsilon^{4/3}\right), \tag{4.15}$$

$$y^{(m)'}(0, \varepsilon) = (-1)^{m+1} \varepsilon^{-2/3} g^{(m)'}(0) - k_2 g^{(m)}(0) + O(\varepsilon), \tag{4.16}$$

where $k_2 = (2a_1 + b_0^2)/10$, a short calculation reveals

$$c_{mn} = \frac{m}{4} \exp[-i^{m+n}\pi/4] + O(\varepsilon). \tag{4.17}$$

Hence, for example, taking $N = 1$ in (4.3), in addition to

$$y^{(2)}(x, \varepsilon)\, e^{\nu s(x)} = g^{(2)}\left(-\nu^{2/3}x\right) + \varepsilon^{1/6}\left[h(x)\, e^{-r(x)} - (-x)^{-1/4}\right] + O\left(\varepsilon^{1/3}\right) \tag{4.18}$$

for $-1 \leq x \leq 0$, we also have

$$y^{(2)}(x, \varepsilon) = \alpha(x, \varepsilon) \cos[\nu s(x) + \pi/4] + \beta(x, \varepsilon) \sin[\nu s(x) + \pi/4] \tag{4.19}$$

for $0 \leq x \leq 1$, where

$$\alpha(x, \varepsilon) = A\left(\nu^{2/3}x\right) + \varepsilon^{1/6}\left[h(x) \cos r(x) - x^{-1/4}\right] + O\left(\varepsilon^{1/3}\right), \tag{4.20}$$

$$\beta(x,\varepsilon) = B\left(\nu^{2/3}x\right) + \varepsilon^{1/6}h(x)\sin r(x) + O\left(\varepsilon^{1/3}\right). \tag{4.21}$$

The leading terms here are

$$A(X) = \pi^{1/2}\left[\mathrm{Bi}(-X)\cos W + \mathrm{Ai}(-X)\sin W\right], \tag{4.22}$$

$$B(X) = \pi^{1/2}\left[\mathrm{Bi}(-X)\sin W - \mathrm{Ai}(-X)\cos W\right], \tag{4.23}$$

where $W = (2/3)X^{3/2} + \pi/4$. It is readily checked that $A(X) = X^{-1/4}[1 + O(X^{-3})]$ and $B(X) = O(X^{-7/4})$ as $X \to \infty$. These functions also appear to be monotone.

Acknowledgment. The author is grateful to the referees of this paper for a number of helpful comments.

REFERENCES

[1] J. KEVORKIAN AND J. D. COLE, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1981.

[2] R. E. LANGER, *The asymptotic solutions of ordinary linear differential equations of the second order, with special reference to a turning point*, Trans Amer. Math. Soc., 67 (1949), pp. 461–490.

[3] R. Y. S. LYNN AND J. B. KELLER, *Uniform asymptotic solutions of second order linear differential equations with turning points*, Comm. Pure. Appl. Math., 23 (1970), pp. 379–408.

[4] F. W. J. OLVER, *The asymptotic solution of linear differential equations of the second order for large values of a parameter*, Philos. Trans. Roy. Soc. London, Ser. A, 247 (1954), pp. 307–327.

[5] ———, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[6] L. A. SKINNER, *Uniformly valid composite expansions for Laplace integrals*, SIAM J. Math. Anal., 19 (1988), pp. 918–925.

[7] W. WASOW, *Linear Turning Point Theory*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1412–1432, September 1994

010

DISCRETE SPLINE FILTERS FOR MULTIRESOLUTIONS AND WAVELETS OF l_2*

AKRAM ALDROUBI†, MURRAY EDEN†, AND MICHAEL UNSER†

Abstract. The authors consider the problem of approximation by B-spline functions, using a norm compatible with the discrete sequence-space l_2 instead of the usual norm L_2. This setting is natural for digital signal/image processing and for numerical analysis. To this end, sampled B-splines are used to define a family of approximation spaces $\mathbf{S}_m^n \subset l_2$. For n odd, $\mathbf{S}_m^n$ is partitioned into sets of multiresolution and wavelet spaces of l_2. It is shown that the least squares approximation in $\mathbf{S}_m^n$ of a sequence $s \in l_2$ is obtained using translation-invariant filters. The authors study the asymptotic properties of these filters and provide the link with Shannon's sampling procedure. Two pyramidal representations of signals are derived and compared: the l_2-optimal and the stepwise l_2-optimal pyramids, the advantage of the latter being that it can be computed by the repetitive application of a single procedure. Finally, a step by step discrete wavelet transform of l_2 is derived that is based on the stepwise optimal representation. As an application, these representations are implemented and compared with the Gaussian/Laplacian pyramids that are widely used in computer vision.

Key words. multiresolution, wavelets, splines, sampling, ideal filter, pyramid

AMS subject classifications. 42C15, 41A15, 94A11, 94A12, 94A15

1. Introduction.

Images, signals, and numerical data are usually available to us as a sequence of real or complex numbers. The sequence space l_2 is therefore natural to consider. However, for the purpose of deriving numerical algorithms, it is sometimes desirable to represent an l_2 sequence by an analog function. This is often realized by interpolation techniques [20], [21], [37], [44], [45]. The computations are usually performed numerically on digital computers, and the results are sequences of numbers. Image magnification, reduction, signal coding, and reconstruction are examples [8], [21], [22], [37], [38]. In order to allow for such dual discrete/analog representations, we develop the theory of polynomial spline approximation for discrete sequences. For this purpose, we consider the problem of least squares approximation of discrete functions in the discrete spline spaces:

$$\mathbf{S}_m^n := \left\{ v \in l_2 : v(k) = \sum_{i \in Z} c(i) b_m^n(k - mi), \quad c \in l_2 \right\}, \tag{1}$$

where b_m^n is the sampled B-spline functions of order n (cf. §2.2). It should be noted that any sequence $v \in \mathbf{S}_m^n$ can be obtained by sampling a polynomial spline function of order n (for an extensive treatment of polynomial splines, see [9], [15], [32], [34]). Thus, the approximation of a sequence $s(k)$ in $\mathbf{S}_m^n$ is equivalent to fitting $s(k)$ with a uniformly spaced analog polynomial spline function that minimizes the discrete l_2-norm of the error (cf. Remark 1 in §5.1). For nonuniformly spaced knot points, the latter problem is usually solved by standard matrix techniques [14]. In our case, we treat the uniformly spaced knot points. We show that in this case, the approximation in $\mathbf{S}_m^n$ can be obtained by discrete translation-invariant filtering, as illustrated in Fig. 1. Therefore, this theory is particularly well adapted to signal and image processing. We study the properties of the approximation filters and discuss the theory in light of Shannon's sampling procedure. We then use the results to construct multiresolution

* Received by the editors July 7, 1992; accepted for publication (in revised form) April 30, 1993.
† Biomedical Engineering and Instrumentation Programming, National Institutes of Health, Bethesda, Maryland 20892.

and wavelet spaces for l_2 instead of the usual space L_2 [10], [26], [29], [43]. Other approaches to constructing multiresolution and wavelet spaces of l_2 can be found in [33] (cf. Remark 2 in §5.2).

This paper is organized as follows: In §2, we use discretized B-splines of order n to construct and analyze a family of discrete sequence spaces $\mathbf{S}_m^n$, where n indexes a smoothness constraint and where m is a scale index measuring the coarseness of the space. In §3, we solve the problem of finding the best l_2 approximation to a signal in $\mathbf{S}_m^n$ and show that it can be obtained by a prefiltering followed by a down-sampling, an up-sampling, and an interpolation, as shown in Fig. 1. Both the prefiltering and the interpolation can be carried out by translation-invariant filtering using fast algorithms [41]. In §4, we provide the link between the approximation problem in $\mathbf{S}_m^n$ and the classical Shannon sampling procedure [3], [4], [22], [39]. More specifically, we prove that the frequency response of the prefilters $\mathring{H}_m^n(f)$ tend to the ideal discrete lowpass filter with periodic support in $\bigcup_{j\in\mathcal{Z}}[j-1/2m, j+1/2m]$, and that the discrete spline interpolators $H_m^n(f)$ tend to the ideal lowpass filter with periodic support in $\bigcup_{j\in\mathcal{Z}}[j-1/2m, j+1/2m]$ and gain m. Related convergence results for the analog case can be found in [4], [15], [25], [28], [35]. In §5, we use our results to construct and discuss two multiresolution representations of signals: the optimal spline pyramid (OP) and the stepwise optimal spline pyramid (SOP). Based on the SOP and some techniques similar to those developed by Daubechies, Mallat, and Vetterli [13], [27], [42], we derive a stepwise discrete wavelet transform of l_2 (the stepwise optimal wavelet pyramid SWP). Finally, we use an example to compare the OP and the SWP representations with the Gaussian/Laplacian pyramids that are widely used in computer vision [6].

2. Notation and preliminaries.

2.1. Definitions and notation. The signals considered here are discrete functions with "finite energy." The collection of all such signals constitutes the space of square summable sequences l_2.

The symbol "$*$" will be used for three slightly different binary operations that are defined below: the convolution, the mixed convolution, and the discrete convolution. The ambiguity should be easily resolved from the context.

For two functions f and g defined on $\mathcal{R}$, $*$ denotes the usual convolution:

$$(f*g)(x)=\int_{-\infty}^{+\infty} f(\xi)g(x-\xi)d\xi, \quad x\in\mathcal{R}. \tag{2}$$

The mixed convolution between a sequence $\{b(k)\}_{k\in\mathcal{Z}}$ and a function f defined on $\mathcal{R}$ is the function $b*f$ defined on $\mathcal{R}$, given by

$$(b*f)(x)=\sum_{k=-\infty}^{k=+\infty} b(k)f(x-k), \quad x\in\mathcal{R}. \tag{3}$$

The discrete convolution between two sequences a and b is the sequence $a*b$:

$$(a*b)(l)=\sum_{k=-\infty}^{k=+\infty} a(k)b(l-k), \quad l\in\mathcal{Z}. \tag{4}$$

Whenever it exits, the convolution inverse $(b)^{-1}$ of a sequence b is defined by

$$\big((b)^{-1}*b\big)(k)=\delta_0(k), \tag{5}$$

where δ_i is the unit impulse located at i ; i.e., $\delta_i(i) = 1$ and $\delta_i(k) = 0$ for $k \neq i$.

We will use the term *Fourier transform* to describe both the usual Fourier transform for functions defined on $\mathcal{R}$:

$$\hat{g}(f) = \int_{\mathcal{R}} g(x) e^{-i2\pi fx} dx, \tag{6}$$

and the usual Fourier transform for sequences,

$$\hat{b}(f) = \sum_{i \in \mathcal{Z}} b(k) e^{-i2\pi fk}. \tag{7}$$

A continuous filter $\hat{\lambda}(f)$ is the Fourier transform of a function λ on $\mathcal{R}$ (the impulse response) that defines a bounded convolution operator on L_2:

$$\lambda : \ g \in L_2 \rightarrow \lambda * g \in L_2. \tag{8}$$

Since the convolution product $\lambda * g$ becomes a multiplication product $\hat{\lambda}\hat{g}$ in Fourier space, the filter $\hat{\lambda}$ selectively alters the frequency components of $\hat{g}$.

A discrete filter $\hat{h}(f)$ is the Fourier transform of a function h on $\mathcal{Z}$ (the impulse response) that defines a bounded convolution operator on l_2:

$$h : \ u \in l_2 \rightarrow h * u \in l_2. \tag{9}$$

The reflection of a sequence b is the function $b^\vee$, given by

$$b^\vee(k) = b(-k) \quad \forall k \in \mathcal{Z}. \tag{10}$$

The modulation $\tilde{b}(k)$ of a sequence b is obtained by changing the signs of the odd components of b:

$$\tilde{b}(k) = (-1)^k b(k). \tag{11}$$

The operator $\downarrow_m$ of down-sampling by the integer factor m assigns to a sequence b the sequence $\downarrow_m [b]$, given by

$$(\downarrow_m [b])(k) = b(mk) \quad \forall k \in \mathcal{Z}. \tag{12}$$

The operator $\uparrow_m$ of up-sampling by the integer factor m takes a discrete signal b and expands it by adding $m - 1$ zeros between consecutive samples:

$$(\uparrow_m [b])(k) = \begin{cases} b(k'), & k = mk', \\ 0, & \text{elsewhere.} \end{cases} \tag{13}$$

2.2. The discrete spline spaces $\mathbf{S}_m^n$. We begin by defining the discrete B-spline $b_m^n(k)$ of order n and integer coarseness $m \geq 1$:

$$b_m^n(k) = \beta^n(k/m) \quad \forall k \in \mathcal{Z}, \tag{14}$$

where $\beta^n(x)$ are the continuous symmetrical B-splines of order n. These are obtained by the n-fold convolution of the B-spline of order zero:

$$\beta^n(x) = (\beta^0 * \beta^0 * \cdots * \beta^0)(x) \qquad (n \text{ convolution}), \tag{15}$$

where $\beta^0(x)$ is the characteristic function in the interval $[-1/2, 1/2)$ (i.e., $\beta^0(x) = 1$ in $[-1/2, 1/2)$ and $\beta^0(x) = 0$ elsewhere). The bell-shaped functions $\beta^n(x)$ have compact support. They were introduced by Schoenberg, who used them to construct a simple basis for the polynomial splines spaces of order n [34].

Using the sequences b_m^n in (14), we define the subspaces $\mathbf{S}_m^n$ of l_2 to be

$$(16)\quad S_m^n := \left\{ v \in l_2 : v(k) = \sum_{i \in Z} c(i) b_m^n(k - mi) = (b_m^n * \uparrow_m [c])(k), \quad c \in l_2 \right\},$$

where n and m are positive integers and where the operator $\uparrow_m$ is defined by (13). As shown in §3, for n odd (which we will assume throughout) the vector spaces $\mathbf{S}_m^n$ are closed subspaces of l_2, and $\mathbf{S}_1^n = l_2$. The discrete functions in $\mathbf{S}_m^n$ are smooth in the sense that they are samples of polynomial spline functions of class C^{n-1}. In this way, the index n is the description of a smoothness constraint. In §3, it is shown that if $m_2 = km_1$ (m_1, m_2, k are positive integers), then $\mathbf{S}_{m_2}^n \subset \mathbf{S}_{m_1}^n$. Thus, in some sense, the index m is related to the coarseness of the spaces $\mathbf{S}_m^n$.

2.3. Review of some results on the continuous fundamental spline filters. The fundamental spline function of order n, $\eta^n(x)$ (also known as cardinal, or interpolating spline) has the value 1 at $x = 0$, and is zero at all the other knot points (the only knot points we consider here are the integers). Thus, it is used to interpolate between data points producing a continuous spline function of order n [4], [28], [34]. Given a discrete signal $s(k)$, its spline interpolation σ^n is given by

$$(17)\qquad \sigma^n(x) = (s * \eta^n)(x) = \sum_{i \in \mathcal{Z}} s(i) \eta^n(x - i).$$

Equation (17) states that the polynomial spline interpolant $\sigma^n(x)$ is obtained by filtering the tempered distribution $\sum_{i \in \mathcal{Z}} s(i)\delta(x - i)$ with a filter whose impulse response is $\eta^n(x)$. Using Poisson's formula, the Fourier transform of $\eta^n(x)$ is given by

$$(18)\qquad H^n(f) = \frac{(\operatorname{sinc}(f))^{n+1}}{\sum_{i \in \mathcal{Z}} (\operatorname{sinc}(f - i))^{n+1}} = \begin{cases} 0, & f \in \mathcal{Z} \setminus 0, \\ 1, & f = 0, \\ (1 + U^n(f))^{-1}, & \text{elsewhere}, \end{cases}$$

where $\operatorname{sinc}(x) = \sin(\pi x)/\pi x$ and where $U^n(f)$ is given by

$$(19)\qquad U^n(f) = \begin{cases} \sum_{i=1}^{i=\infty} (i/f + 1)^{-n-1} + (i/f - 1)^{-n-1}, & n \text{ odd}, \\ \sum_{i=1}^{i=\infty} (-1)^i \left((i/f + 1)^{-n-1} - (i/f - 1)^{-n-1} \right), & n \text{ even}. \end{cases}$$

An important feature is that the Fourier transform $H^n(f)$ of $\eta^n(x)$ converges to the ideal lowpass filter; a well-known property [4], [15], [28], [35] stated in the following theorem.

THEOREM 1. *The Fourier transforms of the fundamental spline interpolators $H^n(f)$ converge to the ideal lowpass filter as n goes to infinity pointwise almost everywhere and in $L_p(-\infty, +\infty)$ for all $p \in [1, \infty)$:*

$$(20)\qquad L_p - \lim_{n \to \infty} H^n(f) = \operatorname{rect}(f) = \begin{cases} 1, & |f| < 1/2, \\ 1/2, & |f| = 1/2, \\ 0, & |f| > 1/2. \end{cases}$$

3. Least squares approximation in the spaces $\mathbf{S}_m^n$. Since one of our goals is to find least squares approximations in the spaces $\mathbf{S}_m^n$, we start by studying the properties of $\mathbf{S}_m^n$.

3.1. Properties of $\mathbf{S}_m^n$. First, we prove that the spaces $\mathbf{S}_m^n$ in (16) are well-defined subspaces of l_2 by showing that $b_m^n * \uparrow_m [c] \in l_2$, for all $c \in l_2$. To see this, we note that since β^n has compact support, the sequence b_m^n defined by (14) has finitely many nonzero values. Thus, b_m^n is absolutely summable. This implies that b_m^n defines a bounded convolution operator from l_2 into itself. From this and the fact that the up-sampling operator is an isometry, we get

$$\|b_m^n * (\uparrow_m [c])\|_{l_2} \leq \|b_m^n\|_{l_1} \|c\|_{l_2} . \tag{21}$$

From the last inequality, it immediately follows that $\mathbf{S}_m^n$, given by (16), are well-defined subspaces of l_2.

There are embedding relations between the spaces $\mathbf{S}_m^n$. These embeddings follow from the well-known embedding properties of the continuous polynomial splines of order n [25], [27], [30].

PROPOSITION 2. *If n is odd, then*

$$\mathbf{S}_{lm}^n \subset \mathbf{S}_m^n \quad \forall l \in \mathcal{Z}^+. \tag{22}$$

Proof. For n odd, the B-spline $\beta^n(x/lm)$ (where l is a positive integer) is also a polynomial spline with knot points on $m\mathcal{Z}$. Thus, it can be written, in terms of $\beta^n(x/m)$ and a sequence $u \in l_2$, as

$$\beta^n(x/lm) = \sum_{i \in \mathcal{Z}} u(i) \beta^n \left(\frac{x}{m} - i \right). \tag{23}$$

Both this equality and the definition of b_m^n given by (14) imply that

$$b_{lm}^n = b_m^n * (\uparrow_m [u]) . \tag{24}$$

We use (24), together with the operator identity

$$\uparrow_{ml} = \uparrow_m \uparrow_l \tag{25}$$

and the equality

$$\uparrow_m [v_1] * \uparrow_m [v_2] = \uparrow_m [v_1 * v_2] \tag{26}$$

to get

$$(\uparrow_{lm} [c]) * b_{lm}^n = (\uparrow_{lm} [c]) * (\uparrow_m [u]) * b_m^n = (\uparrow_m [(\uparrow_l [c]) * u] * b_m^n \quad \forall c \in l_2. \tag{27}$$

The Fourier transform of u,

$$U(f) = l \mathrm{sinc}^{n+1}(lf) / \mathrm{sinc}^{n+1}(f), \tag{28}$$

is continuous and bounded above by a constant. Thus, we have

$$\|\uparrow_l [c] * u\|_{l_2} \leq \mathrm{Const} \|c\|_{l_2} . \tag{29}$$

The proof of the proposition then follows from (27) and (29) and from the definition of $\mathbf{S}_m^n$ given by (16).

Since our aim is to find least squares approximations in $\mathbf{S}_m^n$, we need to show that $\mathbf{S}_m^n$ are closed subspaces of l_2—a result that we state in the following theorem.

THEOREM 3. *If n is odd, then $\mathbf{S}_1^n = l_2$ and $\mathbf{S}_m^n$ are closed subspaces of l_2.*

The proof of the theorem relies on the following simple lemma.

LEMMA 4. *Let $B_m^n(f)$ denote the Fourier transform of $b_m^n(k)$. If n is odd, then there exist two positive constants α_1 and α_2 such that*

$$\alpha_1 \le B_m^n(f) \quad \forall f \in [-1/2m, 1/2m], \tag{30}$$

$$B_m^n(f) \le \alpha_2 \quad \forall f \in \mathcal{R}. \tag{31}$$

Proof. We first note that the Fourier transform of $\beta^0(x)$ is the function $\operatorname{sinc}(f)$. From this fact, Poisson's formula, and the definition of b_m^n (equation (14)), $B_m^n(f)$ can be expressed as

$$B_m^n(f) = m \sum_{i \in \mathcal{Z}} (-1)^{(n+1)mi} \left(\frac{\sin(m\pi f)}{m\pi(f-i)} \right)^{n+1} \tag{32}$$

Clearly, the function $B_m^n(f)$ is both symmetrical ($B_m^n(f) = B_m^n(-f)$) and periodic with period 1. Since the terms of the series in (32) are continuous and of the order of $|i|^{-n-1}$, it follows that the series in (32) converges uniformly for all $n > 0$ in the interval $f \in [0, 1]$. Thus, $B_m^n(f)$ is continuous. Since $B_m^n(f)$ is continuous and periodic, it is bounded above by some constant α_2.

For n odd and for $f \in [0, 1/2m]$, all the terms of the series (32) are nonnegative, and the term for $i = 0$ is strictly positive. Hence, $B_m^n(f)$ is bounded below by a strictly positive constant α_1.

Proof of Theorem 3. To prove that $\mathbf{S}_m^n$ are closed, we show that the operator $b_m^n * \uparrow_m : c \in l_2 \to b_m^n * \uparrow_m [c] \in l_2$ is coercive (i.e., $\|b_m^n * \uparrow_m [c]\|_{l_2} \ge \alpha \|c\|_{l_2}$ for all $c \in l_2$ for some $\alpha > 0$). Taking the Fourier transform of $b_m^n * \uparrow_m [c]$ and using Plancherel's theorem, we get

$$\|b_m^n * \uparrow_m [c]\|_{l_2}^2 = \int_0^1 |B_m^n(f)\hat{c}(mf)|^2 \, df = m^{-1} \int_0^m |B_m^n(f/m)\hat{c}(f)|^2 \, df, \tag{33}$$

where $B_m^n(f)$ is the Fourier transform of b_m^n. By integrating over intervals of length 1 and using the fact that $\hat{c}(f)$ and $B_m^n(f)$ are periodic with period 1, we rewrite the term following the last equality in (33) to obtain

(34)

$$\begin{aligned} m^{-1} \int_0^m |B_m^n(f/m)\hat{c}(f)|^2 \, df &= m^{-1} \int_0^1 |\hat{c}(f)|^2 \sum_{j=0}^{m-1} |B_m^n(f/m - j/m)|^2 df \\ &\ge m^{-1} \operatorname*{essinf}_{f \in I=[0,1]} \left(\sum_{j=0}^{m-1} |B_m^n(f/m - j/m)|^2 \right) \|\hat{c}\|_{l_2}^2 \\ &\ge m^{-1} \operatorname*{essinf}_{f \in I=[0,1/2]} \left(|B_m^n(f/m)|^2 \right) \|\hat{c}\|_{l_2}^2 . \end{aligned}$$

Therefore, the coercivity of the operator $b_m^n * \uparrow_m$ follows directly from (33), (34), and Lemma 4.

3.2. Approximations in $\mathbf{S}_m^n$. Since by Theorem 3 $\mathbf{S}_m^n$ is a closed subspace of l_2, the least squares approximation s_a of s is given by the orthogonal projection on $\mathbf{S}_m^n$. Hence, the error $s - s_a$ is orthogonal to $\mathbf{S}_m^n$. In particular, because of definition (16), the error is orthogonal to b_m^n and to all of its shifted versions at integer multiples of m:

$$\langle (s - s_a)(k), b_m^n(k - lm) \rangle_{l_2} = 0 \quad \forall l \in \mathcal{Z}. \tag{35}$$

Using the expression of $s_a \in \mathbf{S}_m^n$ given by

$$s_a(k) = \sum_{i \in \mathcal{Z}} c_a(i) b_m^n(k - mi) = (\uparrow_m [c_a] * b_m^n)(k), \tag{36}$$

we rewrite (35) to get

$$\langle s(k), b_m^n(k - lm) \rangle_{l_2} = \sum_{i \in \mathcal{Z}} c_a(i) \langle b_m^n(k - im), b_m^n(k - lm) \rangle_{l_2} \quad \forall l \in \mathcal{Z}. \tag{37}$$

Using the fact that b_m^n is symmetric, we can express (37) as the convolution equation

$$\downarrow_m [b_m^n * s] = c_a * \downarrow_m [b_m^n * b_m^n]. \tag{38}$$

This equation can be solved to obtain the unknown sequence c_a. A filtering interpretation of this process is given in [40]. The facts that this procedure is well defined, that the filters are stable, and that equation (37) can be solved follow from the following theorem.

THEOREM 5. *The Fourier transform $T_m^n(f)$ of $t_m^n(l) := \downarrow_m [b_m^n * b_m^n](l)$ is strictly positive. Moreover, $t_m^n \in l_1$, and it has a convolution inverse $(t_m^n)^{-1} \in l_1$ with Fourier transform $(T_m^n(f))^{-1}$ that is also strictly positive.*

Proof. The sequence $t_m^n(l) := \downarrow_m [b_m^n * b_m^n](l)$ has finitely many nonzero values. Thus, $t_m^n \in l_1$, and it defines a bounded convolution operator from l_2 into itself (e.g., $||t_m^n * c||_{l_2} \leq \text{Const}||c||_{l_2}$). Using the relation between the Fourier transform of a discrete signal $b(k)$ (cf. (7)) and its down-sampled version $\downarrow_m [b]$

$$(\downarrow_m [b])\hat{}(f) = m^{-1} \sum_{j=0}^{m-1} \hat{b}(f/m - j/m), \tag{39}$$

we obtain the Fourier transform $T_m^n(f)$ of t_m^n:

$$T_m^n(f) = m^{-1} \sum_{j=0}^{m-1} |B_m^n(f/m - j/m)|^2. \tag{40}$$

The function $T_m^n(f)$ is precisely the sum that appears in the right-hand side of the first inequality in (34). Therefore, Lemma 4 implies that $T_m^n(f)$ is strictly positive and is bounded above by a constant. It follows that $(t_m^n)^{-1} \in l_2$ exists. In fact, since t_m^n has only finitely many nonzero values, $T_m^n(f)$ is a strictly positive trigonometric polynomial. Therefore, $(t_m^n)^{-1}$ decays exponentially fast as $|l| \to \infty$. Hence, $(t_m^n)^{-1}(l)$ is also absolutely summable.

From Theorem 5, t_m^n and $(t_m^n)^{-1}$ define bounded convolution operators on l_2 that are the inverses of each other (cf. (9)). Thus, they are the impulse responses

of the filters $T_m^n(f)$ and $(T_m^n(f))^{-1}$. We use $(t_m^n)^{-1}$ to solve (38) and obtain the approximation s_a:

$$\begin{aligned} s_a &= b_m^n * \uparrow_m [c_a] \\ &= b_m^n * \uparrow_m \left[(t_m^n)^{-1} * \downarrow_m [b_m^n * s]\right] \\ &= b_m^n * \uparrow_m \left[\downarrow_m \left[\uparrow_m [(t_m^n)^{-1}] * b_m^n * s\right]\right], \end{aligned} \tag{41}$$

where in the last equality of (41), we have used the functional equality

$$a * \downarrow_m [b] = \downarrow_m [\uparrow_m [a] * b] \quad \forall a, b \in l_2. \tag{42}$$

3.3. Fundamental discrete spline filters. The space $\mathbf{S}_m^n$ (n, m fixed) can be generated by bases other than $\{b_m^n(k - mi)\}_{i \in \mathcal{Z}}$. A complete characterization of all unconditional bases of a given separable Hilbert space as well as a simple way to obtain any particular basis from any other can be found in [1]. In particular, all Riesz bases for $\mathbf{S}_m^n$ can be characterized in term of $\{b_m^n(k - mi)\}_{i \in \mathcal{Z}}$ by appropriate "linear combinations." For instance, it is not difficult to show that if we sample the Battle/Lemarié spline scaling function $\phi^n(x/m)$ on $\mathcal{Z}$, we obtain the sequence $l_m^n = \uparrow_m [(b_1^{2n+1})^{-1/2}] * b_m^n$. The set $\{l_m^n(k - mi)\}_{i \in \mathcal{Z}}$ is also a basis for $\mathbf{S}_m^n$. However, it does not form an orthogonal basis of $\mathbf{S}_m^n$. The orthogonal basis is given by $\{o_m^n(k - mi)\}_{i \in \mathcal{Z}}$ generated by $o_m^n = \uparrow_m [(t_m^n)^{-1/2}] * b_m^n$.

A particular basis of interest is the fundamental basis $\{h_m^n(k - mi)\}_{i \in \mathcal{Z}}$, in which the representation of any sequence $s(k) \in \mathbf{S}_m^n$ is directly obtained from the sequence values $\{s(mi)\}_{i \in \mathcal{Z}}$:

$$s(k) = \sum_{i \in \mathcal{Z}} s(mi) h_m^n(k - mi) = \uparrow_m \downarrow_m [s] * h_m^n. \tag{43}$$

The fundamental sequence $h_m^n(k)$ is obtained by sampling the continuous fundamental spline filter $\eta^n(x)$, defined at the beginning of §2.3:

$$h_m^n(k) = \eta^n(k/m) \quad \forall k \in \mathcal{Z}. \tag{44}$$

The sequence $h_m^n(k)$ is the linear combination of $b_m^n(k)$ given by [38]:

$$h_m^n = \uparrow_m [(b_1^n)^{-1}] * b_m^n. \tag{45}$$

The existence of $(b_1^n)^{-1}$ follows from Lemma 4. In fact, since b_1^n has finitely many nonzero values, it follows that $(b_1^n)^{-1}$ decays exponentially fast. Thus, both b_1^n and $(b_1^n)^{-1}$ are in $\in l_1$. The fact that $\{h_m^n(k - mi)\}_{i \in \mathcal{Z}}$ is a basis of $\mathbf{S}_m^n$ follows immediately from the definition of $\mathbf{S}_m^n$, Lemma 4, and equations (26) and (45).

Using identity (26), we manipulate (41) so as to exhibit h_m^n. We get

$$s_a = h_m^n * \uparrow_m \left[\downarrow_m \left[\mathring{h}_m^n * s\right]\right], \tag{46}$$

where

$$\mathring{h}_m^n = \uparrow_m [(t_m^n)^{-1} * b_1^n] * b_m^n. \tag{47}$$

From (46), it is not difficult to see that h_m^n and $\mathring{h}_m^n$ are biorthogonal [11]:

$$\downarrow_m \left[\mathring{h}_m^n * h_m^n\right](k) = \delta_0(k) \quad \forall k \in \mathcal{Z}. \tag{48}$$

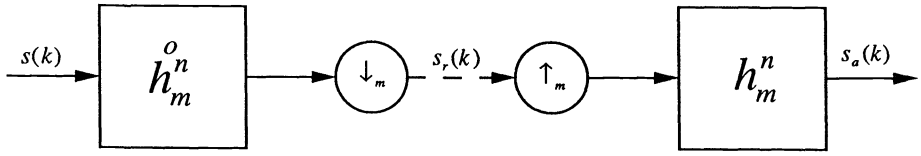


FIG. 1. *Schematic representation of the least squares approximation in* $\mathbf{S}_m^n$.

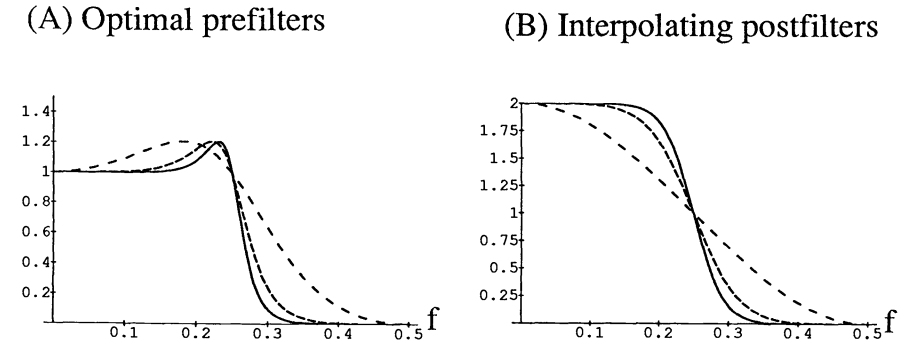


FIG. 2. *Least squares spline filters.* (A) *Prefilters* $\mathring{H}_2^1(f)$ (- - - -), $\mathring{H}_2^3(f)$ (- - - -), *and* $\mathring{H}_2^5(f)$ (*continuous line*).

Similar to Hummel [22], we can interpret $\mathring{h}_m^n$ to be the optimal prefilter needed before the interpolator h_m^n, given by (45). The process that can be used to determine the best approximation of a signal in $\mathbf{S}_m^n$ is illustrated in Fig. 1. In effect, this procedure is equivalent to prefiltering the data with $\mathring{h}_m^n$, down-sampling by a factor m, then up-sampling by m and applying a discrete fundamental spline interpolation. The frequency responses of the filters ($n = 1, 3, 5$ and $m = 2$) used in our procedure are shown in Fig. 2. The graphs show that the lowpass characteristics of these filters improve with n. This behavior will be analyzed in more detail in the next section.

4. Asymptotic properties of the filters. We will show that for m fixed, the prefilters $\mathring{H}_m^n(f)$ and the interpolating filters $H_m^n(f)$ tend pointwise and in $L_2(-1/2, 1/2)$ to an ideal discrete lowpass filter with periodic support in $\bigcup_{j\in\mathbb{Z}}[j - 1/2m, j + 1/2m]$.

These convergence properties are described in the following theorem.

THEOREM 6. *For* n *odd, the prefilter* $\mathring{H}_m^n(f)$ *converges in* $L_2(-1/2, 1/2)$ *and pointwise almost everywhere to an ideal discrete lowpass filter* Prect_m (*the periodic rectangular pulse*) *as* n *tends to infinity*:

$$\lim_{n\to\infty} \mathring{H}_m^n(f) = \begin{cases} 1, & |f| < 1/2m, \\ 1/2, & |f| = 1/2m, \\ 0, & 1/2m < |f| < 1. \end{cases} \tag{49}$$

Similarly, for n *odd, the interpolating filter* $H_m^n(f)$ *converges in* $L_2(-1/2, 1/2)$ *and pointwise almost everywhere to an ideal discrete lowpass filter with gain* m *as* n *tends to infinity*:

$$\lim_{n\to\infty} H_m^n(f) = m\,\mathrm{Prect}_m(f). \tag{50}$$

Using Plancherel's Theorem, we immediately obtain the following corollary.

COROLLARY 7. *For n odd, the impulse responses $\mathring{h}_m^n(k)$ converge in l_2 to the ideal discrete interpolator, $D\mathrm{sinc}_m(k) = \mathrm{sinc}(k/m)$ with $k \in \mathcal{Z}$, as n tends to infinity.*

Similarly, for n odd, the interpolators $h_m^n(k)$ converge in l_2 to the ideal discrete interpolators with gain m, $mD\mathrm{sinc}_m(k)$, as n tends to infinity.

These results are conceptually interesting because they provide the link with Shannon's sampling theory [2], [4], [5], [7], [17], [18], [20], [24], [31], [39], [46]. In particular, for the case of uniform sampling, Shannon's sampling paradigm for non-bandlimited signals states that a signal must first be prefiltered by an ideal filter before sampling and that the signal "reconstruction" is obtained by an ideal post-filtering. The approximation in $\mathbf{S}_m^n$ gives rise to the same structure as illustrated by Fig. 1. It consists of a prefiltering with $\mathring{H}_m^n$, followed by down-sampling by a factor m. This first step gives an m fold reduction in the data. The approximation is then obtained by up-sampling and postfiltering with H_m^n. Moreover, in the limit, all the filters converge to discrete ideal filters. Similar results for the analog polynomial spline case can be found in [4], [39]. The general case for analog functions is described in [2].

The above asymptotic results also explain the appearance of Gibbs oscillations which occur when sequences are approximated by elements in spaces $\mathbf{S}_m^n$ with sufficiently high smoothing order n.

Proof of Theorem 6. Symmetry allows us to restrict our attention to the frequency interval $f \in [0, 1/2)$. The Fourier transform of b_m^n is given by

$$B_m^n(f) = m \sum_{i \in \mathcal{Z}} \mathrm{sinc}^{n+1}(m(f-i)). \tag{51}$$

We use (39) and (40) in conjunction with the fact that the Fourier transform of a discrete signal is periodic with period 1. We also use $(\uparrow_m [b])\hat{}(f) = \hat{b}(mf)$ to express the Fourier transform of $\uparrow_m [(t_m^n)^{-1}]$ and $\uparrow_m [b_1^n] = \uparrow_m [\downarrow_m [b_m^n]]$ in terms of $B_m^n(f)$ as

$$(\uparrow_m [b_1^n])\hat{}(f) = m^{-1} \sum_{j=L}^{L+m-1} B_m^n(f - j/m), \tag{52}$$

$$(\uparrow_m [(t_m^n)^{-1}])\hat{}(f) = \frac{m}{\sum_{j=L}^{L+m-1} |B_m^n(f-j/m)|^2}, \tag{53}$$

where L is an arbitrary integer. Using (51)–(53), we obtain the Fourier transform $\mathring{H}_m^n(f)$ of $\mathring{h}_m^n$:

$$\mathring{H}_m^n(f) = \frac{B_m^n(f) \sum_{j=L}^{L+m-1} B_m^n(f-j/m)}{\sum_{j=L}^{L+m-1} |B_m^n(f-j/m)|^2}. \tag{54}$$

Using (51), (18) and straightforward trigonometric identities, (54) can be written as

$$\mathring{H}_m^n(f) = (H^n(f))^{-1} \frac{\sum_{j=L}^{L+m-1} (-1)^j (1-j/mf)^{-n-1} (H^n(f-j/m))^{-1}}{\sum_{j=L}^{L+m-1} (1-j/mf)^{-2n-2} (H^n(f-j/m))^{-2}}, \tag{55}$$

where $H^n(f)$ is the Fourier transform of the continuous fundamental spline function of order n given by (18). We choose L in (55) to be $L = -(m-1)/2$ if m is odd, and $L = -(m/2-1)$ if m is even. With this choice, we take limits in (55) to get

$$\lim_{n\to\infty} \mathring{H}_m^n(f) = \lim_{n\to\infty} \frac{\sum_{j=L}^{L+m-1} (-1)^j (1-j/mf)^{-n-1}}{\sum_{j=L}^{L+m-1} (1-j/mf)^{-2n-2}} = 1 \quad \forall f \in (0, 1/2m). \tag{56}$$

We use Schwarz's inequality on (55) to get the estimate

$$\left|\mathring{H}_m^n(f)\right| \le m^{1/2} |H^n(f)|^{-1} \left(\sum_{j=L}^{L+m-1} (1-j/mf)^{-2n-2} \left(H^n(f-j/m)\right)^{-2} \right)^{-1/2}. \tag{57}$$

For $f \in (1/2m, 1/2)$, it follows from (57) and Theorem 1 that $\mathring{H}_m^n(f)$ converges pointwise to 0 as n tends to infinity. Moreover, (57) yields the upper bound

$$\left|\mathring{H}_m^n(f)\right| \le m^{1/2}. \tag{58}$$

Because of Lebesgue's dominated convergence theorem, equations (56)–(58) imply that $\mathring{H}_m^n(f)$ tends to $P\mathrm{rect}_m(f)$ in $L_2(-1/2, 1/2)$.

To prove the second part of the theorem, we first note that $H_m^n(f)$ is given by

$$H_m^n(f) = \frac{B_m^n(f)}{(\uparrow_m [b_1^n])\hat{}(f)}. \tag{59}$$

From (51) and for n odd, it can be seen that

$$B_m^n(f) = m^{-n} B_1^n(f) \left(\frac{\sin(m\pi f)}{\sin(\pi f)} \right)^{n+1} \tag{60}$$

Using the fact that $(\uparrow_m [b])\hat{}(f) = \hat{b}(mf)$, and using expressions (18) and (60), we simplify (59) to obtain

$$H_m^n(f) = \frac{B_m^n(f)}{B_1^n(mf)} = m^{-n} \left(\frac{\sin(m\pi f)}{\sin(\pi f)} \right)^{n+1} \frac{B_1^n(f)}{B_1^n(mf)} = m \frac{H^n(mf)}{H^n(f)}. \tag{61}$$

The last equality in (61) and Theorem 1 together yield the pointwise convergence. For n odd, a simple estimate derived for $H^n(f)$ (cf. (18) and (19)) yields that for $f \in (-1/2, 1/2)$, $1/2 < H^n(f) < 1$. Hence, from (61) we get

$$|H_m^n(f)| \le 2m, \tag{62}$$

which implies the $L_2(-1/2, 1/2)$ convergence of $H_m^n(f)$ to the ideal discrete filter $P\mathrm{rect}_m(f)$ with gain m and periodic support in $\bigcup_{j\in\mathbb{Z}} [j - 1/2m, j + 1/2m]$.

5. Multiresolution pyramids and step-by-step discrete wavelet transform.

5.1. The optimal and stepwise optimal discrete spline pyramids. A multiresolution pyramid representation of a discrete signal consists of several versions of the signal at different resolution levels. The name pyramid derives from the fact that the low-resolution levels are described by fewer samples than their high-resolution counterparts. In applied mathematics and image processing, multiscale representations have been used to find efficient algorithms that start computations at coarse levels and subsequently refine them at finer levels [19], [36].

A multiresolution representation of a signal is commonly obtained by the repeated application of a filtering and a down-sampling to produce the pyramid layers. The Gaussian pyramid for images [6] is an example in which each pyramid level is obtained from the previous one by applying a Gaussian filter and down-sampling each row and column of the image by a factor of 2. A shortcoming of this method is that it does not attempt to minimize the loss of information that occurs when one signal is approximated by another at a coarser resolution. Using (46), we can circumvent this limitation, and produce a multiscale representation that optimizes the fine-to-coarse conversion error. For m fixed, we interpret (46) as representing a signal at a lower resolution: the signal s is prefiltered by $\mathring{h}^n_m$, and only one sample out of m is then retained. This sequence is then up-sampled (cf. (13)) and filtered with h^n_m to obtain the best approximation s_a in $\mathbf{S}^n_m$. In effect, the signal $s_r = \downarrow_m [\mathring{h}^n_m * s]$ contains in a compressed form (factor of compression equal to m) all the information needed to reconstruct the approximation s_a. Hence, by selecting a sequence of integers $\{m = p^j\}_{j=1,\dots,N.}$, we can use equation (46) to obtain a multiresolution pyramid $\{s_{r(j)}\}_{j=1,\dots,N.}$:

$$\text{(63)} \qquad \begin{cases} s_{r(j)} = \downarrow_{p^j} \left[\mathring{h}^n_{p^j} * s\right], & j = 1, \dots, N; \\ s_{r(0)} = s. \end{cases}$$

We have used the notation $s_{r(j)}$ to represent level j ($m = p^j$) of the pyramid in (63), which is obtained by filtering the signal s with $\mathring{h}^n_{p^j}$ and then decimating with a factor equal to p^j. More importantly, since $\mathbf{S}^n_{p^{j+1}} \subset \mathbf{S}^n_{p^j}$ (cf. Proposition 2), the filter used to produce the signal $s_{r(j+1)}$ from $s_{r(j)}$ at the previous resolution level can be obtained by using (46) and the fact that for any sequence b, we have that

$$\text{(64)} \qquad \downarrow_{p^{j+1}} [b] = \downarrow_p \left[\downarrow_{p^j} [b]\right].$$

The signal $s_{r(j+1)}$ is given by the following.

The optimal pyramid (OP):

$$\text{(65)} \qquad \begin{aligned} s_{r(j+1)} &= \downarrow_p \left[\mathring{h}^n_p * x_{r(j)}\right], \\ x_{r(j)} &= k^n_{p^j} * s_{r(j)}, \end{aligned}$$

where the operators $k^n_{p^j}$ is given by

$$\text{(66)} \qquad k^n_{p^j} = \uparrow_p \left[\left(t^n_{p^{j+1}}\right)^{-1} * t^n_p\right] * \downarrow_{p^j} \left[h^n_{p^j} * h^n_{p^j}\right].$$

Remark 1. Given a regular function $\sigma(x) \in L_2$ with sufficient decay, it can be approximated by an analog spline $\sigma^n_{p^j}(x)$ with knot points on $p^j\mathcal{Z}$. The analog spline approximation $\sigma^n_{p^j}(x)$ that minimizes the l_2-error $\sigma(k) - \sigma^n_{p^j}(k)$ computed at the integer points $\mathcal{Z}$ can be obtained from (63) or from algorithm OP: $\sigma^n_{p^j}(x) = \sum_{k\in\mathcal{Z}} \sigma_{r(j)}(k)\eta^n(x/p^j - k)$, where $\eta^n(x)$ is the interpolating spline as in §2.3. Thus, the approximation problem in $\mathbf{S}^n_{p^j}$ consists of finding a coarse polynomial spline approximation that minimizes the discrete l_2-norm of the error at the integers instead of the usual minimization of the L_2-norm on $\mathcal{R}$.

The error $e_{(j)} = s - s_a = s - h^n_{p^j} *\uparrow_{p^j} [s_{r(j)}]$ between the original signal and its approximation is the smallest error in l_2 that can be obtained for approximations of s in $\mathbf{S}^n_{p^j}$. However, a drawback of this representation is that the filter $k^n_{p^j}$ in (65) depends upon the resolution level j. On the other hand, the first equation of (65) is independent of the resolution level, and is precisely the first pyramid level for the representation of the signal $x_{r(j)}$. This observation suggests an alternative algorithm for a multiresolution representation of a signal based on the first equation of (65) only.

The stepwise optimal pyramid (SOP):

$$\begin{aligned} \breve{s}_{r(j+1)} &= \downarrow_p \left[\mathring{h}^n_p * \breve{s}_{r(j)}\right] \\ \breve{s}_{r(0)} &= s. \end{aligned} \tag{67}$$

If (67) is used instead of (65) for the pyramidal representation of s, then the error $\breve{e}_{(j)} = s - h^n_{p^j} *\uparrow_{p^j} [\breve{s}_{r(j)}]$ is always larger than or equal to the error $e_{(j)} = s - s_a$. The question of how the two algorithms (65) and (67) compare is partially answered by the following theorem.

THEOREM 8. *For n odd, the filter $K^n_{p^j}(f)$ corresponding to $k^n_{p^j}$ converges in $L_2(-1/2, +1/2)$ and pointwise almost everywhere to a discrete allpass filter as n tends to infinity:*

$$\lim_{n\to\infty} K^n_{p^j}(f) = 1 \quad \forall f \in \mathcal{R}. \tag{68}$$

The proof of this theorem will be omitted, since, except for the use of the identity

$$[b^n_p] = \downarrow_{p^j} [b^n_{p^{j+1}}], \tag{69}$$

it is not very different from the proof of Theorem 6.

Heuristically, the above result states that for sufficiently large n the optimal multiresolution algorithms (65) can be replaced by the simpler and more practical algorithm (67), with only minor differences in the outcome. The advantage of the stepwise optimal algorithm is that the passage from one level to the next always uses the same algorithm, and can therefore be implemented using a fast recursive filtering similar to the one described in [41].

5.2. A stepwise discrete wavelet representation. The pyramids discussed in the previous sections are redundant. For instance, the stepwise optimal pyramid OP is redundant because it consists of the signal itself ($s = s_{r(0)}$), to which N copies are added that are of increasingly coarser resolution: $P = \{s_{r(0)}, s_{r(1)}, \ldots, s_{r(N)}\}$. The redundant information coincides with the data $s_{r(1)}, \ldots, s_{r(N)}$ and, for $m = 2$, the number of additional samples is approximately equal to the size of $s = s_{r(0)}$. In the case $m = 2$, which is a case of practical interest, we will derive a nonredundant representation equivalent to the SOP pyramid. The main idea is to find a suitable

representation of $(\mathbf{S}_2^n)^\perp$, the orthogonal complement of $\mathbf{S}_2^n$, analogous to (16). To do this, we use techniques similar to the ones developed by Daubechies, Mallat, and Vetterli [13], [27], [42]. We start by defining the discrete function w_2^n and the corresponding space $\mathbf{O}_2^n$ associated with it:

$$w_2^n(k) = (-1)^{k+1} b_2^n(k+1). \tag{70}$$

$$\mathbf{O}_2^n := \left\{ v \in l_2 : v(k) = \sum_{i \in \mathcal{Z}} c(i) w_2^n(k-2i) = (w_2^n *\uparrow_2 [c])(k), \quad c \in l_2 \right\}. \tag{71}$$

We have the following result.

THEOREM 9. *The space $\mathbf{O}_2^n$ is the orthogonal complement of $\mathbf{S}_2^n$ in l_2: $\mathbf{O}_2^n = (\mathbf{S}_2^n)^\perp$.*

Before proving this theorem, we first note that the function w_2^n is the discrete equivalent of a continuous wavelet, as defined in [13], [27]; however, in this case w_2^n is not orthogonal to a shifted version of itself. An important point is that the error signal $d_{a(1)} = s - s_a$, resulting from approximating s by $s_a \in \mathbf{S}_2^n$, can be obtained by filtering, as in §3.2:

$$d_a = w_2^n *\uparrow_2 \left[(t_2^n)^{-1} *\downarrow_2 [(w_2^n)^\vee * s]\right] = \delta_1 * \tilde{h}_2^n *\uparrow_2 \left[\downarrow_2 [\delta_{-1} * \overset{\sim}{\circ} h_2^n * s]\right], \tag{72}$$

where the reflection operator " $^\vee$ " and the modulation operator " $\tilde{}$ " are defined by (10) and (11), respectively, in §2.

Proof. First, we show that $w_2^n(k-2i)$ is orthogonal to $b_2^n(k)$ by showing that $\downarrow_2 [(w_2^n)^\vee * b_2^n] = 0$ (where $(w_2^n)^\vee(k) = w_2^n(-k)$). Using (39) and the properties of the Fourier transform, we obtain

$$\begin{aligned}
(\downarrow_2 [(w_2^n)^\vee * b_2^n])\widehat{\ }(f) &= \frac{1}{2}\left(B_2^n(f/2)\overline{W_2^n(f/2)} + B_2^n(f/2-1/2)\overline{W_2^n(f/2-1/2)}\right) \\
&= \frac{1}{2} e^{-i\pi f} (B_2^n(f/2)B_2^n(f/2-1/2) \\
&\qquad - B_2^n(f/2-1/2)B_2^n(f/2-1)) \\
&= \frac{1}{2} e^{-i\pi f} \left(B_2^n(f/2)B_2^n(f/2-1/2) - B_2^n(f/2-1/2)B_2^n(f/2)\right) \\
&= 0,
\end{aligned} \tag{73}$$

where $\overline{W_2^n(f)}$ denotes the complex conjugate $W_2^n(f)$, which is the Fourier transform of $w_2^n(k)$. It only remains to show that any element $s \in l_2$ can be written as a sum of its least squares approximations in $\mathbf{S}_2^n$ and in $\mathbf{O}_2^n$. We sum the Fourier transforms of d_a and s_a, which are the approximations of s in $\mathbf{O}_2^n$ and $\mathbf{S}_2^n$, respectively; we then use (39), the second equation in (72), periodicity, and Lemma 4 to obtain

$$\begin{aligned}
\hat{d}_a(f) + \hat{s}_a(f) &= B_2^n(f-1/2)\frac{B_2^n(f-1/2)S(f) - B_2^n(f-1)S(f-1/2)}{|B_2^n(f-1/2)|^2 + |B_2^n(f-1)|^2} \\
&\quad + B_2^n(f)\frac{B_2^n(f)S(f) + B_2^n(f-1/2)S(f-1/2)}{|B_2^n(f-1/2)|^2 + |B_2^n(f)|^2} \\
&= \hat{s}(f),
\end{aligned} \tag{74}$$

from which the proof follows.

Using (72), we derive the difference or detail representation $\{d_{r(j)}\}_{j=1,\ldots,N}$:

The stepwise wavelet pyramid (SWP).

$$(75)\qquad \begin{cases} \breve{d}_{r(j+1)} = \downarrow_2 \left[\delta_{-1} * \tilde{\mathring{h}}_2^n * \breve{s}_{r(j)}\right] \\ \breve{s}_{r(j+1)} = \downarrow_2 \left[\mathring{h}_2^n * \breve{s}_{r(j)}\right] \\ \breve{s}_{r(0)} = s. \end{cases}$$

From Theorem 9 and (46), (72), and (75), it can be seen that the SOP representation is obtained from the SWP pyramid by the iterative algorithm:

The stepwise wavelet decomposition.

$$(76)\quad \breve{s}_{r(N-j)} = h_2^n * \uparrow_2 \left[\breve{s}_{r(N-j+1)}\right] + \delta_1 * \tilde{h}_2^n * \uparrow_2 \left[\breve{d}_{r(N-j+1)}\right], \qquad j = 1, \ldots, N.$$

Remark 2.

(i) The algorithms (75) and (76) constitute a biorthogonal, perfect reconstruction filter bank [33], [43].

(ii) There are two corresponding analog scaling functions $\varphi_{h_2^n}$ and $\varphi_{\mathring{h}_2^n}$, and two analog biorthogonal wavelets $\psi_{\tilde{h}_2^n}$ and $\psi_{\tilde{\mathring{h}}_2^n}$, for which the associated L_2 analysis of analog functions defined on $\mathcal{R}$ via nonorthogonal projections is exactly obtained by (75) and (76) (cf. [11]). Obviously, $\varphi_{\mathring{h}_2^n}$ and $\psi_{\tilde{\mathring{h}}_2^n}$ are not polynomial splines.

(iii) There are infinitely many basis functions for $\mathbf{S}_m^n$ (cf. §3.3). For each basis, it is possible to obtain a step-by-step wavelet decomposition (or a perfect reconstruction, biorthogonal filter banks) similar to (75) and (76). However, they will not be a good approximation to OP in general.

(iv) Two basis functions generating the same space $\mathbf{S}_m^n$ do not correspond to analog scaling functions that generate the same space; e.g., the scaling functions $\varphi_{h_2^n}$ and $\varphi_{b_2^n}$ associated with h_2^n and b_2^n do not generate the same multiresolution space V_0, even though b_2^n and h_2^n generate the same space $\mathbf{S}_2^n$.

(v) If we choose the biorthogonal filter bank decomposition using the orthogonal basis o_2^n in §3.3 instead of h_2^n, then the corresponding stepwise wavelet algorithms are precisely the Mallat wavelet decomposition and reconstruction algorithms for the analog scaling function $\phi(x)$ associated with the QMF o_2^n [27]. In this case, there exists an underlying discrete multiresolution $E_{2^j}^n \neq \mathbf{S}_{2^j}^n$, for which these algorithms give the best l_2 approximation of a sequence $s(k)$ in $E_{2^j}^n$ [33]. These are also the analysis/synthesis algorithms for the L_2 multiresolution wavelet $V_j(\phi)/W(\phi)$, corresponding to the function ϕ associated with o_2^n. However, ϕ is not a spline function. Moreover, this algorithm does not correspond to the same analog multiresolution $V(\varphi_{h_2^n})$ (see the previous remark, (iv)). For the interpretation of (75) and (76) we refer to §3, Remark 1 in §5.1, and Theorem 8.

6. Experiments. Although the filters used in (65), (67), and (75) have an infinite impulse response, they can still be implemented exactly using the recursive algorithm described in [41]. An alternative approach is to use a standard finite impulse response (FIR) implementation with truncated filters. In the latter case, the computation is approximate, but the error is easily controlled by choosing an appropriate number of coefficients. Table 1 gives those filter coefficients for the cases $n = 1$

TABLE 1
Filters' coefficients for $n = 1$ and $n = 3$.

	$\mathring{h}_2^1$	h_2^1	$\mathring{h}_2^3$	h_2^3
$k = 0$	0.707107	1	0.596797	1
$1, -1$	0.292893	0.5	0.313287	0.600481
$2, -2$	−0.12132	0	−0.082769	0
$3, -3$	−0.0502525	0	−0.0921993	−0.127405
$4, -4$	0.0208153	0	0.0540288	0
$5, -5$	0.00862197	0	0.0436996	0.034138
$6, -6$	−0.00357134	0	−0.0302508	0
$7, -7$	−0.0014793	0	−0.0225552	−0.00914725
$8, -8$	0.000612745	0	0.0162251	0
$9, -9$		0	0.0118738	0.002451
$10, -10$		0	−0.00861788	0
$11, -11$		0	−0.00627964	−0.000656743
$12, -12$		0	0.00456713	
$13, -13$		0	0.00332464	
$14, -14$		0	−0.00241916	
$15, -15$		0	−0.00176059	
$16, -16$		0	0.00128128	
$17, -17$		0	0.000932349	
$18, -18$		0	−0.000678643	

and $n = 3$. In our experiments, we used the first of these approaches. To avoid border effects and discontinuities, we have used the common practice of extending the signals/images at the boundaries by taking their mirror images.

We have performed three experiments on a test image, the MRI image. First we compared different approximations of the image by varying the parameters n and m in the approximation spaces $\mathbf{S}_m^n$. To assess the appropriateness of the approximation we used the signal-to-noise ratio [23] associated with the approximation s_a, as defined by

$$SNR = 20 \log \left(\frac{|\sup(s) - \inf(s)|}{\|s - s_a\|_{l_2}} \right). \tag{77}$$

Table 2 gives the measurements of the SNR for values of $m = 2, \ldots, 8$ and $n = 1, 3$. These measurements show that for fixed value of m, the SNR for $n = 3$ is higher than for $n = 1$. The improvement seems to saturate quickly, and we anticipate no significant gain for values of n larger than 5. This conclusion is consistent with the convergence results given in Theorem 6, which indicate that the approximation process tends to an ideal filtering process for increasing values of n. As a consequence, higher orders of n will, in general, improve the SNR for images with a predominance of lower frequency components.

Our second experiment was a comparison of the two multiresolution representations OP and SOP. Table 3 gives the values of the SNR for a multiresolution representation of the MRI obtained by the optimal algorithm (67) for the case $p = 2$, $j = 1, 2, 3$, and $n = 1, 3$. Obviously, the two algorithms are equivalent for the determination of level 1. As predicted, the SNR measured for the approximations obtained using the optimal algorithm OP are higher than those obtained form the stepwise optimal algorithm SOP. However, as predicted by Theorem 8, the difference between the two SNRs are small, particularly for the largest value of $n = 3$. Indeed, these differences (which are of the order of 0.01 dB) are very small if compared to the degradation of

TABLE 2
Approximation error (in dB) *evaluated for* $m = 1, 2, \ldots, 8$ *and* $n = 1, 3$ *in terms of the signal-to-noise ratio for the* MRI *image.*

	$n = 1$	$n = 3$
$m = 2$	32.65	35.39
$m = 3$	27.84	29.08
$m = 4$	24.95	25.58
$m = 5$	23.22	23.73
$m = 6$	22.30	22.68
$m = 7$	21.22	21.62
$m = 8$	20.54	20.85

TABLE 3
Comparison between the optimal pyramid and the stepwise optimal pyramid representations in term of the signal-to-noise ratio (in dB) *for the* MRI *image.*

	$n = 1$		$n = 3$	
	OP	SOP	OP	SOP
Level-1	32.65	32.65	35.39	35.39
Level-2	24.95	24.93	25.58	25.57
Level-3	20.54	20.50	20.85	20.83

the SNR between two successive levels, which is of the order of 5dB-10 dB. In fact, the SNR differences in our experiment are negligible, and the results are much better than we had expected.

We compared our multiresolution representation given by the SOP algorithm (67) to the Laplacian pyramid LP which was developed for compact image coding [6]. Each level in the difference-image pyramid consists of the difference between the image at one level and its interpolated version at the next lower level. In other words, each layer of such a pyramid represents the loss of information between a level and its approximation at the coarser level. For this experiment we chose the value $n = 3$, $p = 2$, and $j = 1, 2, 3$ in the SOP algorithm. Fig. 3 shows the difference images for the two representations, with the same intensity scaling to facilitate comparison. For the initial LP, there is significant information at each level, and the initial image is still easily recognizable. In the case of the SOP, the energy in the difference is reduced drastically, and only very high-frequency details are visible. This improvement can be applied advantageously to progressive image transmission. For lossless image coding, the number of bits per pixel (bit-rate) necessary to transmit the bottom of the pyramid up to level j is approximately

$$B_j = \sum_{i=j}^{N} H_i 4^{i-1}, \tag{78}$$

where H_i denotes the entropy at the ith level of the Laplacian pyramid, and N is the depth of the pyramid. The corresponding rate-distortion curves for our test image are given in Fig. 4. The customary measure of distortion that is used for this type of experiment is the relative mean square error in percent of the total signal energy, as measured on the finer scale. Clearly, the SOP achieves the best performance at all resolution levels. Thus, for the comparable compression factor, we can gain image quality when the SOP is used instead of the LP representation.

Finally, Fig. 5 displays an equivalent SWP representation of the same MRI image. This decomposition was obtained by successive processing along the rows and columns

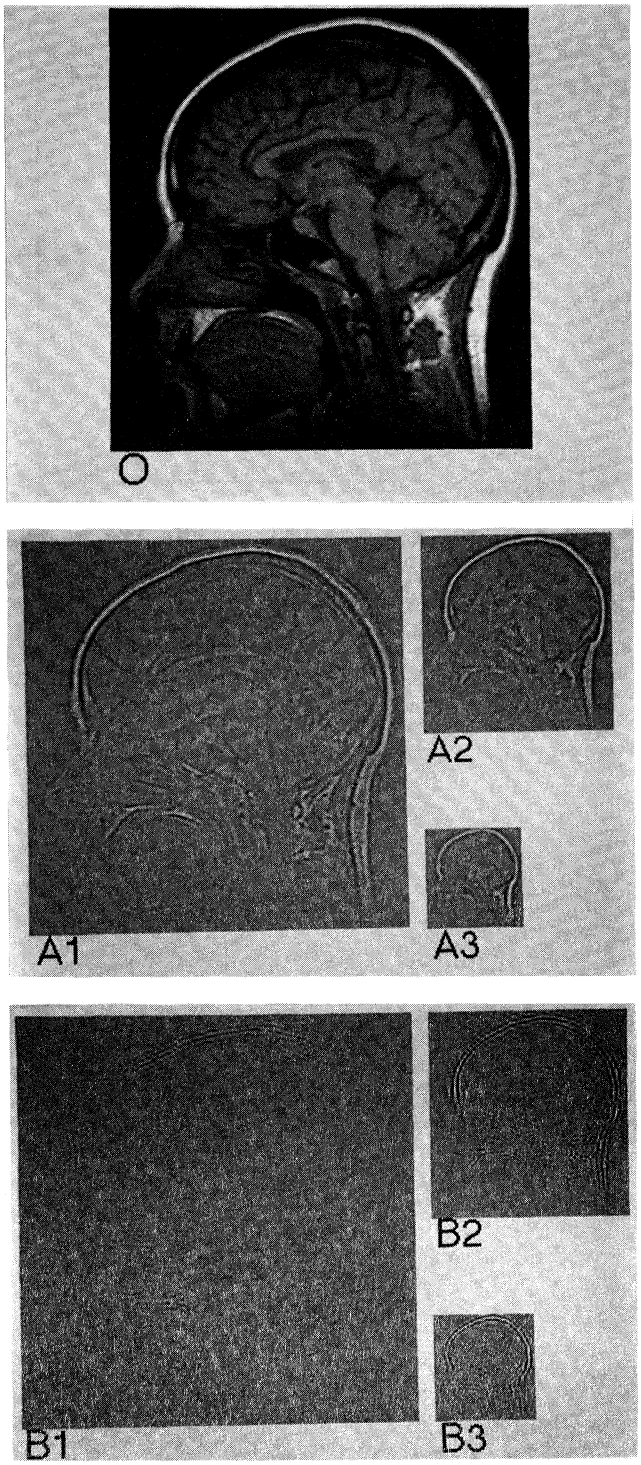


FIG. 3. *Error images between two consecutive levels of the* SOP *pyramid and the Laplacian pyramid* LP *for the* MRI *image O.* (A1-A3) *error/difference images of the Laplacian pyramid.* (B1-B3) *error/difference images of the* SOP *pyramid* ($n = 3$).

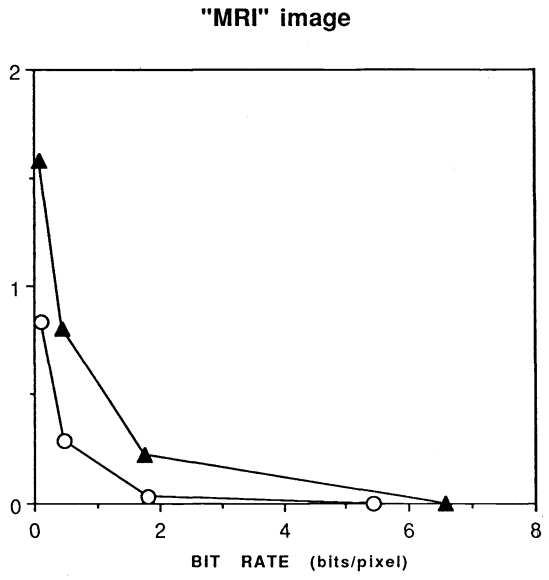


FIG. 4. *Rate distortion* (MSE) *as a function of the number of bits per pixel needed for lossless transmission up to level i*: SOP (*circle*) *and* LP (*triangle*).

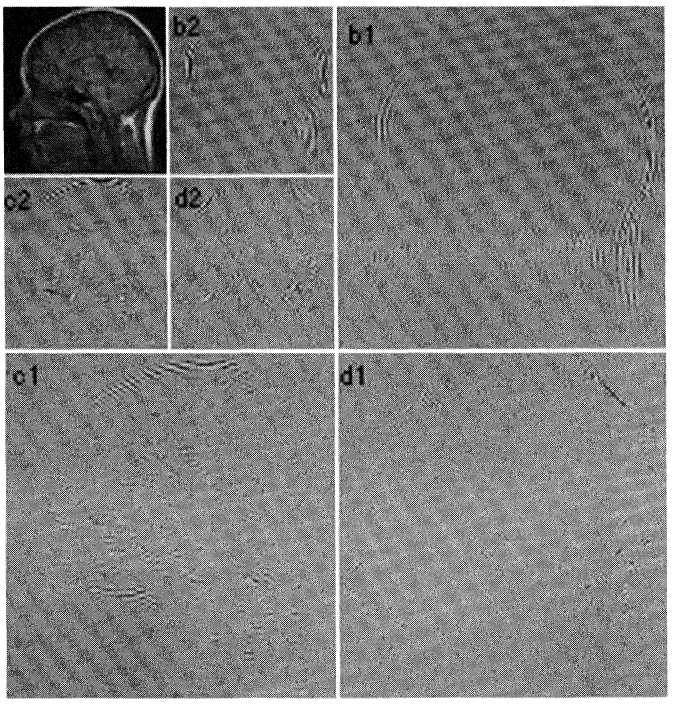


FIG. 5. *The stepwise wavelet representation of the* MRI *image with a level depth* 2 (*i.e.*, $m = 2^j$, $j = 1, 2$) *with* $n = 3$.

of the data, following the separable technique first described by Vetterli in [42]. We note that the three quadrants b_1, c_1, and d_1 provide a compressed representation of the difference at level 1 in the SOP (image A1). Likewise, the difference at level 2 (image A2) is represented by the wavelet components b_2, c_2, and d_2. The component in a_2 is precisely the SOP approximation after two iterations (level 2). The decomposition is clearly nonredundant; and, as expected, we have experimentally tested that the original image can be fully recovered from the stepwise pyramid without error. This wavelet decomposition can be used for both image compression and for coding, as described in [12], [16], and [42].

Acknowledgment. We thank two anonymous reviewers for their helpful comments. We also thank Mr. Barry Bowman for his editorial assistance.

REFERENCES

[1] A. ALDROUBI AND M. UNSER, *Families of multiresolution and wavelet spaces with optimal properties*, Numer. Funct. Anal. Optim., 14 (1993), pp. 417–446.

[2] ———, *Sampling procedures in function spaces and asymptotic equivalence with Shannon's sampling theory*, Numer. Funct. Anal. Optim., 15 (1994), pp. 1–21.

[3] ———, *Families of Wavelet Transforms in Connection with Shannon's Sampling Theory and the Gabor Transform*, Wavelets- A Tutorial in Theory and Applications, 2 (1992), pp. 509–528.

[4] A. ALDROUBI, M. UNSER, AND M. EDEN, *Cardinal spline filters: stability and convergence to the ideal sinc interpolator*, Signal Process., 28 (1992), pp. 127–138.

[5] J. J. BENEDETTO AND W. HELLER, *Irregular sampling and the theory of frames*, preprint.

[6] P. J. BURT AND E. H. ADELSON, *The Laplacian pyramid as a compact code*, IEEE Trans. Comm., COM-31 (1983), pp. 337–345.

[7] P. L. BUTZER, *A survey of the Whittaker-Shannon sampling therorem and some of its extensions*, J. Math. Res. Exposition, 3 (1983), pp. 185–212.

[8] P. L. BUTZER, W. ENGELS, S. RIES, AND R. L. STENS, *The Shannon sampling series and the reconstruction of signals in terms of linear quadratic and cubic splines*, SIAM J. Appl. Math., 46 (1986), pp. 299–323.

[9] C. K. CHUI, *Multivariate Splines*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988.

[10] C. K. CHUI AND J. Z. WANG, *A cardinal spline approach to wavelets*, Proc. Amer. Math. Soc., 113 (1991), pp. 785–793.

[11] A. COHEN, I. DAUBECHIES, AND J. C. FEAUVEAU, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[12] R. R. COIFMAN, Y. MEYER, S. QUAKE, AND M. V. WICKERHAUSER, *Signal processing and compression with wave packets*, Proceedings of the Conference on Wavelets, 1989.

[13] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[14] C. DE BOOR, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[15] C. DE BOOR, K. HÖLLIG, AND S. RIEMENSCHNEIDER, *Bivariate cardinal interpolation by splines on a three-direction mesh*, Illinois J. Math., 29 (1985), pp. 533–566.

[16] R. A. DEVORE, B. JAWERTH, AND B. J. LUCIER, *Image compression through wavelet transform coding*, IEEE Trans. Inform. Theory, 38 (1992), pp. 719–746.

[17] H. G. FEICHTINGER AND K. GRÖCHENIG, *Multidimensional irregular sampling of band-limited functions in L_p spaces*, ISNM 90 Oberwolfach, (1989), pp. 135–142.

[18] G. G. GILBERT, *A sampling theorem for wavelet subspaces*, IEEE Trans. Information Theory, 38 (1992), pp. 881–884.

[19] W. HACKBUSH, *Multi-Grid Methods and Applications*, Springer-Verlag, New York, 1985.

[20] J. R. HIGGINS, *Five short stories about the cardinal series*, Bull. Amer. Math. Soc., 121 (1985), pp. 45–89.

[21] H. S. HOU AND H. C. ANDREWS, *Cubic splines for image interpolation and digital filtering*, IEEE Trans. Acoust. Speech Signal Process., ASSP-26 (1978), pp. 508–517.

[22] R. HUMMEL, *Sampling for spline reconstruction*, SIAM J. Appl. Math., 43 (1983), pp. 278–288.

[23] A. K. JAIN, *Fundamentals of digital, image processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[24] A. J. JERRI, *The shannon sampling theorem-its various extensions and applications: A tutorial review*, Proc. IEEE, 65 (1977), pp. 1565–1596.
[25] P. G. LEMARIÉ, *Ondelettes Φ localisation exponentielles*, J. Math. Pures Appl., 67 (1988), pp. 227–236.
[26] S. G. MALLAT, *Multiresolution approximations and wavelet orthogonal bases of $L_2(\mathcal{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.
[27] ———, *A theory of multiresolution signal decomposition: the wavelet representation*, IEEE Trans. Pattern Anal. Machine Intell., PAMI-11 (1989), pp. 674–693.
[28] M. J. MARSDEN, F. B. RICHARDS, AND S. D. RIEMENSCHNEIDER, *Cardinal spline interpolation operators on l_p data*, Indiana Univ. Math. J., 24 (1975), pp. 677–689.
[29] Y. MEYER, *Ondelettes, functions splines, et analyses graduees*, Univ. of Toronto, 1986.
[30] ———, *Ondelettes*, Hermann, Editeurs des Sciences et des Arts, Paris, 1990.
[31] M. Z. NASHED AND G. G. WALTER, *General sampling theorems for functions in reproducing kernel Hilbert spaces*, Math. Control Signals Systems, 4 (1991), pp. 373–412.
[32] P. M. PRENTER, *Splines and variational methods*, John Wiley, New York, 1975.
[33] O. RIOUL, *A discrete-time multiresolution theory*, IEEE Trans. Signal Processing, 41 (1993), pp. 2591–2606.
[34] I. J. SCHOENBERG, *Contribution to the problem of approximation of equidistant data by analytic functions*, Quart. Appl. Math., 4 (1946), pp. 45–99, 112–141.
[35] ———, *Notes on spline functions* III: *on the convergence of the interpolating cardinal splines as their degree tends to infinity*, Israel J. Math., 16 (1973), pp. 87–92.
[36] D. TERZOPOULOS, *Image analysis using multigrid relaxation methods*, IEEE Trans. Pattern Anal. Mach. Intell., 8 (1986), pp. 129–139.
[37] K. TORAICHI, S. YANG, M. KAMADA, AND R. MORI, *Two-dimensional spline interpolation for image reconstruction*, Pattern Recognition, 21 (1988), pp. 275–284.
[38] M. UNSER, A. ALDROUBI, AND M. EDEN, *Fast B-spline transforms for continuous image representation and interpolation*, IEEE Trans. Pattern Anal. Machine Intell., 13 (1991), pp. 277–285.
[39] ———, *Polynomial spline signal approximations: filter design and asymptotic equivalence with Shannon's sampling theorem*, IEEE Trans. Inform. Theory, 38 (1991), pp. 95–103.
[40] ———, *B-spline signal processing. Part* I: *Theory*, IEEE Trans. Signal Processing, 41 (1993), pp. 821–834.
[41] ———, *B-spline signal processing. Part* II: *Efficient design and applications*, IEEE Trans. Signal Processing, 41 (1993), pp. 834–848.
[42] M. VETTERLI, *Multi-dimensional sub-band coding: some theory and algorithms*, Signal Process., 6 (1984), pp. 97–112.
[43] M. VETTERLI AND C. HERLEY, *Wavelets and filter banks*, IEEE Trans. Signal Processing, 40 (1992), pp. 2207–2231.
[44] J. M. WHITTAKER, *Interpolation Function Theory*, Cambridge Tracts in Mathematics and Mathematical Physics, Cambridge University Press, Cambridge, UK, 1935.
[45] A. ZAYED, G. HINSEN, AND P. BUTZER, *On Lagrange interpolation and Kramer-type sampling theorems associated with Sturm-Liouville problems*, SIAM J. Appl. Math., 50 (1990), pp. 893–909.
[46] A. I. ZAYED, *On Kramer's Sampling Theorem associated with general Sturm-Liouville problems and Lagrange interpolation*, SIAM J. Appl. Math., 51 (1991), pp. 575-604.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, pp. 1433–1460, September 1994

011

WAVELET ANALYSIS OF REFINEMENT EQUATIONS*

LARS F. VILLEMOES†

Abstract. The Besov regularity of a compactly supported refinement equation solution is determined by the spectral radius of a linear operator acting on $\ell^p(\mathbb{Z})$. The proof of this is obtained by using a wavelet basis. Exact criteria for Hölder and Sobolev regularity follow immediately. Continuity, differentiability, and integrability can also be characterized.

The results are applied to examples from the theory of orthonormal and biorthogonal wavelets and subdivision schemes for curve design.

Key words. refinement equations, wavelets, Besov spaces

AMS subject classifications. 39A10, 42C15, 46E35

1. Introduction.

The purpose of this paper is to analyze the smoothness of a compactly supported solution g to the refinement equation

$$g(x) = \sum_{k\in\mathbb{Z}} 2c_k\, g(2x-k), \tag{1.1}$$

where only a finite number of the complex coefficients c_k are different from zero and $\sum_k c_k = 1$.

The dependence of the regularity of g on the choice of coefficients c_k has been studied recently by many authors, both with varying motivations and definitions of regularity. In the construction of compactly supported orthonormal and biorthogonal wavelet bases for $L^2(\mathbb{R})$, the equation (1.1) is satisfied by the scaling function of the underlying multiresolution analysis [Ma], [M], [D], [C3]. The wavelet is then a finite linear combination of translates of g. When studying the convergence properties of subdivision schemes for curve design, (1.1) also arises in a natural way [CDM], [DGL], [DDD].

Conditions are known for g to be continuous, to have continuous derivatives or to be in a certain Hölder class $C^s(\mathbb{R})$. The conditions are either sufficient [DL2], [CDM], or necessary and sufficient, possibly with minor restrictions on the coefficients [CH1], [CH2], [R], [DDD], [DGL], [MP]. The results are often formulated in terms of joint spectral properties of two matrices defined from the coefficients c_k, and are obtained by a direct study of (1.1).

By methods based on the Fourier transform, conditions that are either sufficient or necessary for Hölder regularity have been found [D], [DL1], [CC]. Optimality holds only in special situations (positive Fourier transform) [DD2], [R], [CD]. However, in terms of Sobolev spaces $H^s(\mathbb{R})$, precise results can be obtained [E], [V], [G], [H].

We show in §2 that an exact condition for g to be in the Besov space $B_p^{s,q}(\mathbb{R})$ can be obtained by using a wavelet decomposition of g, (Theorem 2.5). The interplay between the dyadic scaling in (1.1) and the dyadic scalings of the wavelet basis makes this analysis quite simple. We reformulate the condition in terms of the spectral radius of a linear operator $\ell^p(\mathbb{Z}) \to \ell^p(\mathbb{Z})$ in §3.

* Received by the editors March 23, 1992; accepted for publication (in revised form) April 17, 1993.

† Mathematical Institute, Technical University of Denmark, DK 2800 Lyngby, Denmark. Present address, Department of Mathematics, Royal Institute of Technology, S-10044 Stockholm, Sweden (larsv@ math.kth.se).

Estimates of this spectral radius are given for the cases $p = \infty, 2, 1$ in §§4, 5, and 6, respectively. We comment upon the immediate applications to Hölder and Sobolev regularity, and then describe how to get criteria for continuity, differentiability and integrability in §7.

The solution g can be obtained as the limit of the iterative scheme $g_{n+1}(x) = \sum_k 2c_k g_n(2x-k)$ staring from a suitable function g_0. Convergence results for this iteration are given in §8, where we also briefly discuss the convergence of a corresponding discrete iterative scheme.

Finally, in §9, we illustrate the obtained results by three examples. An orthonormal scaling function from [D], a symmetric 4-coefficient example studied in both [dR] and [VH], and a family of coefficients corresponding to a 4-point dyadic interpolating scheme considered in [DD1] and [DLG].

We consider only global regularity. For results about local regularity and the connections with fractals, see, e.g., [DL2] and [W]. When the dilation factor 2 is replaced by any integer $N \geq 3$ in (1.1), all results of this paper extend in a straightforward way. For multidimensional generalizations of (1.1) as considered in [CDM], [DDD], and [CD], the methods presented here do not apply directly. Nor is this the case if we allow $c_k \neq 0$ for infinitely many k. However, we believe that the idea of using an adapted wavelet basis for the study of g could be fruitful in these situations as well.

The compactly supported solution. Let $\hat{f}$ denote the Fourier transform of f. (Using the normalization $\hat{f}(\xi) = \int f(x)e^{-i\xi x}\,dx$ for integrable f.) If we define the *symbol* of (1.1) to be the trigonometric polynomial

$$m(\xi) = \sum_k c_k e^{-ik\xi}, \tag{1.2}$$

we get the equivalent form of (1.1):

$$\hat{g}(2\xi) = m(\xi)\hat{g}(\xi). \tag{1.3}$$

If g is supposed to be a compactly supported distribution, then $\hat{g}$ can be extended to an entire function of exponential type. This is the easy part of the the Paley–Wiener theorem for distributions. The existence and uniqueness of the solution to (1.1) which we will consider here is guaranteed by the following Theorem 1.1, easily obtained by combining results from [DL1] and [DDD]. Note that the uniqueness part is easy, and that the condition $\sum_k c_k = 1 \Leftrightarrow m(0) = 1$ is necessary if $\hat{g}(0) \neq 0$. In particular if g is integrable with nonzero integral.

THEOREM 1.1. *If $\sum_k c_k = 1$ there is a unique compactly supported distribution g satisfying* (1.1) *and the normalization $\hat{g}(0) = 1$. This solution is given by*

$$\hat{g}(\xi) = \prod_{j=1}^{\infty} m(2^{-j}\xi). \tag{1.4}$$

The infinite product converges uniformly on every compact subset of $\mathbb{C}$ to an entire function of exponential type, as well as in the sense of tempered distributions when restricting to real ξ.

It follows directly from (1.1) that the support of g must be included in the closed convex hull of the support of $c = (c_k)$ regarded as function $\mathbb{R} \supset \mathbb{Z} \to \mathbb{C}$.

Wavelets and Besov spaces. The main tool for analyzing the regularity of g will be to decompose g with respect to an orthonormal basis of wavelets. It will be sufficient to consider the specific basis due to Meyer [M, p. 74]. The *Meyer wavelet* ψ

and *scaling function* φ have both smooth and compactly supported Fourier transforms and are therefore in the Schwartz class $\mathcal{S}(\mathbb{R})$ of rapidly decaying functions. From their explicit construction we will retain the features that

$$\text{supp}\,\hat{\varphi} = [-\tfrac{4\pi}{3}, \tfrac{4\pi}{3}], \qquad \text{supp}\,\hat{\psi} = [-\tfrac{8\pi}{3}, -\tfrac{2\pi}{3}] \cup [\tfrac{2\pi}{3}, \tfrac{8\pi}{3}], \tag{1.5}$$

and $\hat{\varphi} = 1$ in a neighborhood of the origin. Let τ_a be the translation operator ($\tau_a f(x) = f(x-a)$) and define

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k).$$

Then the system $\{\tau_k\varphi, \psi_{jk} \mid j = 0, 1, 2, \dots, k \in \mathbb{Z}\}$ is an orthonormal basis for the Hilbert space $L^2(\mathbb{R})$ of square integrable functions. Moreover, every tempered distribution $f \in \mathcal{S}'(\mathbb{R})$ has a unique decomposition

$$f = \sum_{k\in\mathbb{Z}} \beta(k)\tau_k\varphi + \sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}} \alpha_j(k)\psi_{jk}, \tag{1.6}$$

$$\text{where} \quad \begin{cases} \beta(k) &= \langle f, \tau_k\varphi\rangle = f(\tau_k\overline{\varphi}), \\ \alpha_j(k) &= \langle f, \psi_{jk}\rangle = f(\overline{\psi}_{jk}). \end{cases}$$

The smoothness of f is measured very precisely by the decay of the coefficients in (1.6). In terms of Besov spaces $B_p^{s,q}(\mathbb{R})$ we have from [M].

THEOREM 1.2. *Let $s \in \mathbb{R}$ and $1 \le p, q \le \infty$. The tempered distribution f given by* (1.6) *is in the Besov space $B_p^{s,q}(\mathbb{R})$ if and only if*

(1) $\beta \in \ell^p(\mathbb{Z})$ *and*

(2) $\{2^{j((1/2)-(1/p)+s)}\|\alpha_j\|_p\}_{j=0}^{\infty} \in \ell^q(\{0, 1, 2, \dots\})$.

Here $\|\cdot\|_p$ denotes the usual norm of $\ell^p(\mathbb{Z})$. The sum of $\|\beta\|_p$ and the ℓ^q-norm of the sequence in (2) *defines an equivalent norm of $B_p^{s,q}(\mathbb{R})$.*

We have $B_2^{s,2}(\mathbb{R}) = H^s(\mathbb{R})$ for all $s \in \mathbb{R}$ where $H^s(\mathbb{R})$ is the Sobolev space $\{f \in \mathcal{S}'(\mathbb{R}) \mid (1+\xi^2)^{s/2}\hat{f} \in L^2(\mathbb{R})\}$. When $s > 0$ is not an integer $B_\infty^{s,\infty}(\mathbb{R}) = C^s(\mathbb{R})$ is the Hölder space of $n = \lfloor s \rfloor$ times continuously differentiable functions with nth derivative having uniform modulus of continuity $O(h^{s-n})$. If $W^{n,p}(\mathbb{R})$ is the Sobolev space of $L^p(\mathbb{R})$-functions with distributional derivatives up to order n also in $L^p(\mathbb{R})$, then the inclusions

$$B_p^{n,1}(\mathbb{R}) \subset W^{n,p}(\mathbb{R}) \subset B_p^{n,\infty}(\mathbb{R}) \tag{1.7}$$

hold. All this can be found in [P], or even from the wavelet characterization of the mentioned function spaces given in [M].

Roughly, for $f \in B_p^{s,q}(\mathbb{R})$ one can think of s as the fractional order of differentiability of f, p as the power to which the corresponding derivative is integrable and q as a subtle refinement parameter. However, the Besov spaces only coincide with Sobolev or potential spaces in the case $p = q = 2$.

2. Characterization of Besov regularity. Our first task will be to compute the coefficients of g with respect to the wavelet basis. Let the operator T_m acting on complex sequences and associated to the symbol m of (1.2) be determined by

$$T_m y(k) = \sum_l 2c_{k-2l}y(l) = 2c * D_2 y(k), \qquad k \in \mathbb{Z}. \tag{2.1}$$

We define the discrete dilation D_n by putting $D_n y(k) = y(k/n)$ for $k \in n\mathbb{Z}$ and zero elsewhere. Also, we denote by ε_n the Dirac sequence $\varepsilon_n(k) = 1$ for $k = n$ and zero for

$k \neq n$. (In the terminology of [CDM], T_m is a *subdivision operator*.) By iteration in (1.1) we get

$$g(2^{-j}x) = \sum_l (T_m^j \varepsilon_0)(l) g(x-l). \tag{2.2}$$

This gives formally the wavelet coefficients for $j = 0, 1, 2, \dots$ and $k \in \mathbb{Z}$,

$$\begin{aligned}\alpha_j(k) = \langle g, \psi_{jk}\rangle &= \int g(x) 2^{j/2} \overline{\psi}(2^j x - k)\, dx \\ &= 2^{-j/2} \int g(2^{-j}x) \overline{\psi}(x-k)\, dx \\ &= 2^{-j/2} \sum_l (T_m^j \varepsilon_0)(l) \langle g, \psi_{0,k-l}\rangle.\end{aligned}$$

Since g is a compactly supported distribution, the definition of α_j has meaning whenever $\psi \in \mathcal{S}(\mathbb{R})$.

LEMMA 2.1. *Let $\psi \in \mathcal{S}(\mathbb{R})$, define $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ and $\alpha_j(k) = \langle g, \psi_{jk}\rangle$. Then for every $j = 0, 1, \dots,$*

$$\alpha_j = 2^{-j/2} \alpha_0 * T_m^j \varepsilon_0.$$

As an almost immediate consequence of Theorem 1.2 and Lemma 2.1, we obtain the following.

COROLLARY 2.2. *Let $s \in \mathbb{R}$ and $1 \leq p, q \leq \infty$. Define $d(j) = \|\alpha_0 * T_m^j \varepsilon_0\|_p 2^{j(s-1/p)}$ with α_0 as in Lemma 2.1. Consider the two statements:*

(1) $g \in B_p^{s,q}(\mathbb{R})$.

(2) $d \in \ell^q(\{0, 1, 2, \dots\})$.

If $\hat{\psi}$ is of class C^∞ with compact support disjoint from the origin, then (1) *implies* (2), *and if ψ is the Meyer wavelet* (1) *and* (2) *are equivalent.*

Proof. We prove the last statement first. Let φ and ψ be as in Theorem 1.2. Note that $\hat{g}\overline{\hat{\varphi}}$ is a C^∞ function with compact support and therefore the Fourier transform of a function f in the Schwartz class $\mathcal{S}(\mathbb{R})$. Hence $\beta(k) = \langle g, \tau_k \varphi\rangle = f(k)$ is a rapidly decreasing sequence and we do not have to consider the global condition (1) of Theorem 1.2.

The equivalence of (1) and (2) now follows from Lemma 2.1 since

$$\|\alpha_j\|_p 2^{j(\frac{1}{2} - \frac{1}{p} + s)} = \|\alpha_0 * T_m^j \varepsilon_0\|_p 2^{j(s - \frac{1}{p})}. \tag{2.3}$$

Assume now that ψ has a smooth Fourier transform with compact support disjoint from the origin, and that (1) is satisfied. Denote by $\widetilde{\varphi}$ and $\widetilde{\psi}$ the Meyer scaling function and wavelet respectively. We can decompose g in the wavelet series:

$$g = \sum_{k \in \mathbb{Z}} \widetilde{\beta}(k) \widetilde{\tau}_k \varphi + \sum_{j=0}^{\infty} \sum_{k \in \mathbb{Z}} \widetilde{\alpha}_j(k) \widetilde{\psi}_{jk}.$$

This series converge in the sense of tempered distributions and the coefficients have the decay properties described by Theorem 1.2. For sufficiently large j' we have $\langle \tau_k \widetilde{\varphi}, \psi_{j'k'}\rangle = 0$, and there is a positive integer J such that $\langle \widetilde{\psi}_{jk}, \psi_{j'k'}\rangle = 0$ for $|j - j'| > J$ and all $k, k' \in \mathbb{Z}$. This follows from the properties of the supports of the Fourier transforms of $\widetilde{\varphi}$, $\widetilde{\psi}$ and ψ, (see (1.5)). Hence, after inserting the wavelet series

for g we get

$$\alpha_{j'}(k') = \langle g, \psi_{j'k'} \rangle = \sum_{|j-j'|\le J} \sum_{k\in\mathbb{Z}} \widetilde{\alpha}_j(k) \langle \widetilde{\psi}_{jk}, \psi_{j'k'} \rangle,$$

for sufficiently large j'. Because of the rapid decay of $\widetilde{\psi}$ and ψ, the operator $y(k') \mapsto z(k') = \sum_k y(k)\langle \widetilde{\psi}_{jk}, \psi_{j'k'} \rangle$ is continuous $\ell^p(\mathbb{Z}) \to \ell^p(\mathbb{Z})$ for all $p \in [1,\infty]$ and it is easily verified that its kernel depends only on $j - j'$. Letting C_p denote the largest operator norm $\ell^p(\mathbb{Z}) \to \ell^p(\mathbb{Z})$ for $|j - j'| \le J$ we obtain therefore the estimate

$$\|\alpha_{j'}\|_p \le C_p \sum_{|j-j'|\le J} \|\widetilde{\alpha}_j\|_p.$$

The statement (2) now follows from Theorem 1.2 and (2.3) with α replaced by $\widetilde{\alpha}$. □

The next step will be to eliminate the apparent dependence on the choice of ψ in Corollary 2.2. To formulate this result we need *Cohen's criterion*:

DEFINITION 2.3. *We say the symbol m satisfies* Cohen's criterion *if there is a compact neighborhood K of $0 \in \mathbb{R}$ such that*

(1) *For almost every $\xi \in [-\pi,\pi]$, there is a unique $\eta \in K$ with $\eta - \xi \in 2\pi\mathbb{Z}$.*

(2) *$m(\xi) \ne 0$ for all $\xi \in \bigcup_{j=1}^{\infty} 2^{-j}K$.*

If $g \in L^2(\mathbb{R})$, the integer translates of g generates a Riesz basis for the closure of their linear span if and only if m satisfies Cohen's criterion [C2]. Note that $\hat{g}$ does not vanish on K exactly when m satisfies (2). The following alternative formulations of Cohen's criterion will be very useful and can essentially be found in [L]. For convenience, we give a proof in the appendix.

PROPOSITION 2.4. *The following four statements are all equivalent with the statement that Cohen's criterion is not satisfied by m:*

(1) *There is a $\xi \in [-\pi,\pi]$ such that $\hat{g}(\xi + 2\pi k) = 0$ for all $k \in \mathbb{Z}$.*

(2) *There is a $\xi \in [-\pi,\pi]$ such that $\sum_{k\in\mathbb{Z}} e^{ik\xi}\tau_k g = 0$.*

(3) *Either we have $m(\xi) = m(\xi+\pi) = 0$ for some $\xi \in [-\pi,\pi]$, or there exists a nonempty finite subset $\mathcal{N}$ of the unit circle in the complex plane such that $\mathcal{N}^2 = \mathcal{N}$, $1 \notin \mathcal{N}$ and $m(\xi+\pi) = 0$ whenever $e^{i\xi} \in \mathcal{N}$.*

(4) *There are trigonometric polynomials Q and m_0 with $Q(0) = m_0(0) = 1$ such that m_0 satisfies Cohen's criterion and $Q(\xi)m(\xi) = Q(2\xi)m_0(\xi)$ for all ξ. Moreover, $Q(\xi) = 0$ for some $\xi \in [-\pi,\pi]$ and $g = \sum_k q(k)\tau_k g_0$ where $q(k)$ are the coefficients of Q and g_0 has symbol m_0.*

By (4) of Proposition 2.4 we can always express g as a finite linear combination of the integer translates of a two-scale difference equation solution g_0 whose symbol satisfies Cohen's criterion. Then it is an easy exercise to show that g and g_0 have exactly the same regularity, using the fact that g_0 has compact support. Therefore, in the following results about the regularity of g, the assumption that Cohen's criterion holds is essentially no restriction. We know how to deal with the situations where the criterion fails.

It is clear from (3) of Proposition 2.4 that if $m(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^M \widetilde{m}(\xi)$, then m satisfies Cohen's criterion if and only if $\widetilde{m}$ does.

THEOREM 2.5. *Assume $m(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^M \widetilde{m}(\xi)$ where $\widetilde{m}(\pi) \ne 0$ and $\widetilde{m}$ is a trigonometric polynomial satisfying Cohen's criterion. Put*

$$r(j) = \|T_{\widetilde{m}}^j \varepsilon_0\|_p 2^{j(s-M-\frac{1}{p})}$$

for $j = 0,1,2,\ldots$, where $T_{\widetilde{m}}$ is defined from $\widetilde{m}$ in analogy with (2.1). If g is the

solution to (1.1) *described by Theorem* 1.1, *then*

$$g \in B_p^{s,q}(\mathbb{R}) \Leftrightarrow r \in \ell^q(\{0,1,2,\dots\})$$

for all $s \in \mathbb{R}$, $1 \le p,q \le \infty$. *If* $\widetilde{m}$ *fail to satisfy Cohen's criterion the implication* "$\Leftarrow$" *still holds.*

This result is a trivial consequence of Corollary 2.2 and the following lemma.

LEMMA 2.6. *If* $m(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^M \widetilde{m}(\xi)$ *and* α_0 *is defined from* ψ *as in Corollary* 2.2, *where* $\hat{\psi}$ *has compact support disjoint from the origin, then*

$$\|\alpha_0 * T_m^j \varepsilon_0\|_p \le C \|T_{\widetilde{m}}^j \varepsilon_0\|_p \, 2^{-jM}$$

for some constant C *and all* $j = 0,1,2,\dots$. *If furthermore* $\widetilde{m}(\pi) \neq 0$ *and* $\widetilde{m}$ *satisfies Cohen's criterion, a* ψ *as above can be chosen such that*

$$\|\alpha_0 * T_m^j \varepsilon_0\|_p \ge C \|T_{\widetilde{m}}^j \varepsilon_0\|_p \, 2^{-jM}$$

for some $C > 0$ *and all* $j = 0,1,2,\dots$.

Proof. If $\widetilde{g}$ is defined from $\widetilde{m}$ in the sense of Theorem 1.1, the factorization of m leads to

$$\hat{g}(\xi) = \left(\frac{1-e^{-i\xi}}{i\xi}\right)^M \hat{\widetilde{g}}(\xi),$$

as it is easily seen from the identity $(1-z^2) = (1-z)(1+z)$ with $z = e^{-i\xi}$. Note that $\widetilde{m}(0) = 1$ guarantees the existence of this $\widetilde{g}$. Multiplying with $\overline{\hat{\psi}}$ we obtain $\hat{g}(\xi)\overline{\hat{\psi}}(\xi) = (1-e^{-i\xi})^M \hat{h}(\xi)$ where $h \in \mathcal{S}(\mathbb{R})$ is defined from $\hat{h}(\xi) = (i\xi)^{-M} \hat{\widetilde{g}}(\xi)\overline{\hat{\psi}}(\xi)$. With $\gamma(k) = h(k)$ for $k \in \mathbb{Z}$ we have $\gamma \in \ell^1(\mathbb{Z})$ and

$$\alpha_0 = (\varepsilon_0 - \varepsilon_1)^{*M} * \gamma.$$

Here, and in the following, $*$ denotes convolution and $y^{*M} = y * y * \cdots * y$ (M factors).

Writing $Y(\xi) = \sum_k y(k) e^{-ik\xi}$, the action of T_m can be described by $Y(\xi) \mapsto 2m(\xi)Y(2\xi)$. Using this, together with the identity $(1-z)\prod_{l=0}^{j-1}(1+z^{2^l}) = (1-z^{2^j})$ for $z = e^{-i\xi}$, one finds:

$$(\varepsilon_0 - \varepsilon_1)^{*M} * T_m^j \varepsilon_0 = 2^{-Mj} (\varepsilon_0 - \varepsilon_{2^j})^{*M} * T_{\widetilde{m}}^j \varepsilon_0.$$

A convolution with γ yields

$$\alpha_0 * T_m^j \varepsilon_0 = \gamma * (\varepsilon_0 - \varepsilon_{2^j})^{*M} * 2^{-Mj} T_{\widetilde{m}}^j \varepsilon_0. \tag{2.4}$$

The first part of the lemma now follows with $C = 2^M \|\gamma\|_1$, since $\|\gamma * \eta\|_p \le \|\gamma\|_1 \|\eta\|_p$ for all $p \in [1,\infty]$.

To show the second part, we proceed in two steps. First, we construct ψ such that γ becomes invertible with respect to convolution in $\ell^1(\mathbb{Z})$. Next, we use the fact that $\widetilde{m}$ is a trigonometric polynomial to find a $\delta > 0$ such that

$$\|(\varepsilon_0 - \varepsilon_{2^j})^{*M} * T_{\widetilde{m}}^j \varepsilon_0\|_p \ge \delta \|T_{\widetilde{m}}^j \varepsilon_0\|_p. \tag{2.5}$$

Once this is done, $C = \delta / \|\gamma^{-1}\|_1$ can be used for the second estimate of the lemma.

Step 1. Since $\widetilde{m}$ satisfies Cohen's criterion, $\hat{\widetilde{g}}(\xi) \neq 0$ for $\xi \in K$, where K is as in Definition 2.3. Moreover, we cannot have $\hat{\widetilde{g}}(2\pi(2p+1)) = 0$ for every integer p, because if this were true, $\hat{\widetilde{g}}(2\pi(2p+1)) = \widetilde{m}(\pi)\hat{\widetilde{g}}(\pi(2p+1))$ and $\widetilde{m}(\pi) \neq 0$ would imply that $\hat{\widetilde{g}} = 0$ on $\pi(2\mathbb{Z}+1)$, contradicting the assumption by (1) of Proposition 2.4.

Hence, $\hat{\tilde{g}}(2\pi p_0) \neq 0$ for some integer $p_0 \neq 0$. By continuity, $\hat{\tilde{g}}$ will not vanish on

$$K_\epsilon = (K \setminus]-\epsilon, \epsilon[) \cup (2\pi p_0 + [-\epsilon, \epsilon]),$$

for sufficiently small $\epsilon > 0$. Fix ϵ small enough to have also $K_\epsilon \cong [-\pi, \pi]$ in the sense of Definition 2.3(1), and choose a compactly supported $\nu \in C^\infty$ with $\nu = 0$ on $]-\epsilon/2, \epsilon/2[$, $\nu = 1$ on K_ϵ and $\nu \geq 0$. Define ψ by

$$\hat{\psi}(\xi) = (-i\xi)^M \nu(\xi) \hat{\tilde{g}}(\xi).$$

Certainly, this $\hat{\psi}$ is smooth with compact support disjoint from the origin. Going back to the definition of γ we find

$$\Gamma(\xi) = \sum_{k \in \mathbb{Z}} \gamma(k) e^{-ik\xi} = \sum_{k \in \mathbb{Z}} \nu |\hat{\tilde{g}}|^2 (\xi + 2\pi k).$$

By construction, Γ is a smooth strictly positive 2π-periodic function. Therefore γ is ℓ^1-invertible with $\gamma^{-1}(k)$ being the Fourier coefficients of $1/\Gamma(\xi)$.

Step 2. The following "separating lemma" will be useful also in the sequel.

LEMMA 2.7. *Let $p \in [1, \infty]$ and $n \in \{1, 2, \dots\}$. If $w \in \ell^p(\mathbb{Z})$ and the sequence y is supported by $\{0, 1, \dots, n-1\}$, then*

$$\|D_n w * y\|_p = \|w\|_p \|y\|_p.$$

Proof of Lemma 2.7. Note that $(D_n w * y)(k) = \sum_l w(l) y(k - nl)$, where the sum has only one term different from zero. For $p < \infty$ this gives

$$\|D_n w * y\|_p^p = \sum_k \Big| \sum_l w(l) y(k - nl) \Big|^p = \sum_l |w(l)|^p \sum_k |y(k - nl)|^p = \|w\|_p^p \|y\|_p^p.$$

The case $p = \infty$ is similar.

Assume, without loss of generality, that $\widetilde{m}(\xi) = \sum_{k=0}^{L-1} \tilde{c}_k e^{-ik\xi}$ for some $L > 0$ and put $\eta = \sum_{k=0}^{L-1} \varepsilon_k$. Then $\eta * (\varepsilon_0 - \varepsilon_1) = \varepsilon_0 - \varepsilon_L$ and the support of $T_{\widetilde{m}}^j \varepsilon_0$ is included in $\{0, 1, \dots, 2^j L - 1\}$. (In fact it is also in the smaller set $\{0, 1, \dots, (L-1)(2^j - 1)\}$.) An application of Lemma 2.7 with $n = 2^j L$, $w = (\varepsilon_0 - \varepsilon_1)^{*M}$ and $y = T_{\widetilde{m}}^j \varepsilon_0$ then gives

$$\begin{aligned} \|(D_{2^j} \eta)^{*M} * (\varepsilon_0 - \varepsilon_{2^j})^{*M} * T_{\widetilde{m}}^j \varepsilon_0\|_p &= \|D_{(2^j L)} \big((\varepsilon_0 - \varepsilon_1)^{*M} \big) * T_{\widetilde{m}}^j \varepsilon_0\|_p \\ &= \|(\varepsilon_0 - \varepsilon_1)^{*M}\|_p \, \|T_{\widetilde{m}}^j \varepsilon_0\|_p, \end{aligned}$$

so

$$\|\eta^{*M}\|_1 \, \|(\varepsilon_0 - \varepsilon_{2^j})^{*M} * T_{\widetilde{m}}^j \varepsilon_0\|_p \geq \|(\varepsilon_0 - \varepsilon_1)^{*M}\|_p \, \|T_{\widetilde{m}}^j \varepsilon_0\|_p.$$

We can therefore use $\delta = \|(\varepsilon_0 - \varepsilon_1)^{*M}\|_p / \|\eta^{*M}\|_1$ in (2.5) and the proof of Lemma 2.6 is complete. □

The two simplest cases can already be analyzed completely by Theorem 2.5.

Example 2.8. Assume $m(\xi) = \big(\frac{1 + e^{-i\xi}}{2}\big)^M$. Keeping in mind that the results of this paper are not changed by a translation of the sequence (c_k), we will also write $c = (\frac{1}{2}, \frac{1}{2})^{*M}$. Then $\widetilde{m}(\xi) = 1$ and $T_{\widetilde{m}} y = 2y$. With notation as in Theorem 2.5, $r(j) = 2^{j(1+s-M-(1/p))}$ and by checking whether this sequence is in ℓ^q we get

$$g \in B_p^{s,q}(\mathbb{R}) \Leftrightarrow \begin{cases} s < M - 1 + \frac{1}{p}, & q < \infty, \\ s \leq M - 1 + \frac{1}{p}, & q = \infty. \end{cases} \tag{2.6}$$

This result is well known because g can be found explicitly. When $M = 0$ the distribution g is the Dirac mass at $x = 0$. For $M \geq 1$ it is the M-fold convolution of the characteristic function of the unit interval, $g = \mathbf{1}_{[0,1]}^{*M}$, which is a B-spline of degree $M - 1$ if $M \geq 2$:

$$g(x) = \frac{1}{(M-1)!} \sum_{k=0}^{M} (-1)^k \binom{M}{k} (x-k)_+^{M-1}.$$

Here $(x)_+ = \max\{0, x\}$. □

Example 2.9. Suppose $m(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^M \widetilde{m}(\xi)$, where

$$\widetilde{m}(\xi) = \widetilde{c}_0 + \widetilde{c}_1 e^{-i\xi}.$$

If $\widetilde{m}(\pi) = 0$ we are in the situation of the preceding example, so let us assume $\widetilde{m}(\pi) \neq 0$. That is $\widetilde{c} = (\widetilde{c}_1, \widetilde{c}_2) \neq (\frac{1}{2}, \frac{1}{2})$.

Since $\widetilde{m}(\xi)$ has at most one zero in $]-\pi, \pi]$, Cohen's criterion is satisfied with K equal to either $[-\epsilon, 2\pi - \epsilon]$ or $[-2\pi + \epsilon, \epsilon]$ for some $\epsilon > 0$. Note that with $T = T_{\widetilde{m}}$,

$$T^{j+1}\varepsilon_0 = T^j \varepsilon_0 * D_{2^j} T \varepsilon_0,$$

and as the support of $T^j\varepsilon_0$ is included in $\{0, 1, \ldots, 2^j - 1\}$, Lemma 2.7 gives

$$\|T^{j+1}\varepsilon_0\|_p = \|T^j\varepsilon_0\|_p \, \|T\varepsilon_0\|_p.$$

Here $T\varepsilon_0 = 2\widetilde{c}$, so by induction, $\|T^j\varepsilon_0\|_p = \|2\widetilde{c}\|_p^j$ and Theorem 2.5 reads

$$g \in B_p^{s,q}(\mathbb{R}) \Leftrightarrow \begin{cases} s < M + \frac{1}{p} - \log_2 \|2\widetilde{c}\|_p, & q < \infty, \\ s \leq M + \frac{1}{p} - \log_2 \|2\widetilde{c}\|_p, & q = \infty. \end{cases} \tag{2.7}$$

We have denoted by $\log_2 x = \frac{\log x}{\log 2}$ the logarithm of base 2. □

3. Spectral radii and the critical exponent. From now on we will always assume the factorization

$$m(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^M \widetilde{m}(\xi), \qquad \widetilde{m}(\pi) \neq 0, \tag{3.1}$$

where $\widetilde{m}$ is a trigonometric polynomial. In other words, M is just the order of the zero of m at $\xi = \pi$. Possibly $M = 0$. It will often be convenient to assume that $c_k = 0$ for negative k. By simple translations, this assumption implies no loss of generality.

For every $p \in [1, \infty]$, $T_{\widetilde{m}}$ is a bounded linear operator $\ell^p(\mathbb{Z}) \to \ell^p(\mathbb{Z})$. The spectral radius $\rho_p(T_{\widetilde{m}})$ of this operator is the supremum of $|\lambda|$ over all $\lambda \in \mathbb{C}$ such that $T_{\widetilde{m}} - \lambda I$ does not have a bounded inverse. The spectral radius formula states that

$$\rho_p(T_{\widetilde{m}}) = \lim_{j\to\infty} \|T_{\widetilde{m}}^j\|_p^{1/j} = \inf_{j \in \mathbb{N}_1} \|T_{\widetilde{m}}^j\|_p^{1/j}, \tag{3.2}$$

where $\|A\|_p$ denotes the operator norm of $A : \ell^p(\mathbb{Z}) \to \ell^p(\mathbb{Z})$. Since $\|T_{\widetilde{m}}^j \varepsilon_0\|_p \leq \|T_{\widetilde{m}}^j\|_p$, the sequence r of Theorem 2.5 will be in all ℓ^q if $\rho_p(T_{\widetilde{m}}) < 2^{s-M-\frac{1}{p}}$, by the root test. We are thus led to the following definition, where it will be convenient to introduce the parameter $v = \frac{1}{p} \in [0, 1]$, putting $\frac{1}{\infty} = 0$ and $\frac{1}{0} = \infty$.

DEFINITION 3.1. *For $v \in [0, 1]$, we define the* critical exponent $s_0(v)$ *of T_m by*

$$s_0(v) = M + v - \log_2 \rho_{\frac{1}{v}}(T_{\widetilde{m}}).$$

Note that $s_0(v)$ depends continuously on the coefficients of $\widetilde{m}$ when $\widetilde{m}(\pi) \neq 0$, since spectral radii depend continuously on operators. The critical exponent deserves its name because of the following.

THEOREM 3.2. *Let $v \in [0,1]$ and $q \in [1,\infty]$. Then $g \in B^{s,q}_{1/v}(\mathbb{R})$ for all $s < s_0(v)$. If $\widetilde{m}$ satisfies Cohen's criterion, we have*

$$\sup\{s \mid g \in B^{s,q}_{\frac{1}{v}}(\mathbb{R})\} = s_0(v).$$

We have already seen that the supremum in Theorem 3.2 is at least $s_0(v)$. The opposite inequality follows from (3.2) and the next lemma.

LEMMA 3.3. *Let $T = T_{\widetilde{m}}$ be defined from N coefficients $c_0, \dots, c_{N-1}$. Then for all $p \in [1,\infty]$,*

$$\|T^j\|_p \leq N^{1-\frac{1}{p}} \|T^j \varepsilon_0\|_p. \tag{3.3}$$

For $p = 1$, equality holds in (3.3).

Proof. The last statement is a consequence of the first since $\|T^j\varepsilon_0\|_1 \leq \|T^j\|_1$.

To show (3.3), we perform what in the language of signal processing is called a "polyphase" decomposition of $y \in \ell^p(\mathbb{Z})$:

$$y = \sum_{k=0}^{N-1} \tau_k D_N y^{(k)},$$

where τ_k denotes translation by k, $\tau_k z(l) = z(l-k)$, and the discrete dilation D_N was defined in the beginning of §2. Then we find

$$T^j y = T^j \varepsilon_0 * D_{2^j} y = \sum_{k=0}^{N-1} \tau_{2^j k}(T^j \varepsilon_0 * D_{(2^j N)} y^{(k)}).$$

The support of $T^j\varepsilon_0$ is included in $\{0, 1, \dots, 2^j N - 1\}$. Hence, by Lemma 2.7 and the triangle inequality we get

$$\|T^j y\|_p \leq \|T^j \varepsilon_0\|_p \sum_{k=0}^{N-1} \|y^{(k)}\|_p,$$

and to complete the proof we only have to note that Hölder's inequality in $\mathbb{C}^N$ furnishes $\sum_{k=0}^{N-1} \|y^{(k)}\|_p \leq N^{1-(1/p)} \|y\|_p$. □

Some elementary properties of the critical exponent $s_0(v)$, regarded as a function of $v \in [0,1]$, are listed in the next proposition.

PROPOSITION 3.4. *$s_0 : [0,1] \to \mathbb{R}$ has the following properties:*

(1) *s_0 is nondecreasing.*

(2) *$s_0(v+h) \leq s_0(v) + h$ for $0 \leq v, h, v+h \leq 1$.*

(3) *$s_0\big((1-t)v_0 + tv_1\big) \geq (1-t)s_0(v_0) + ts_0(v_1)$ for $0 \leq v_0, v_1, t \leq 1$.*

In other words, s_0 is concave.

(4) *$s_0 \leq M$.*

Proof. Put $T = T_{\widetilde{m}}$ and let the support of the coefficients of $\widetilde{m}$ be included in $\{0, 1, \dots, L-1\}$.

(1) Let $p = \frac{1}{v}$ and $p_1 = \frac{1}{v+h}$ with $h \geq 0$. Since the support of $T^j\varepsilon_0$ is included in $\{0, 1, \dots, 2^j L - 1\}$, Hölder's inequality in $\mathbb{C}^{2^j L}$ has the consequence that $\|T^j\varepsilon_0\|_{p_1} \leq (2^j L)^h \|T^j\varepsilon_0\|_p$. Taking jth roots, using Lemma 3.3 and the spectral radius formula

(3.2) we get $\rho_{p_1}(T) \le 2^h \rho_p(T)$, which implies $s_0(v+h) \ge s_0(v)$ by the definition of s_0.

(2) Since ℓ^p-norms decrease with p we have, with notations as above, that $\|T^j\varepsilon_0\|_{p_1} \ge \|T^j\varepsilon_0\|_p$, so $\rho_{p_1}(T) \ge \rho_p(T)$ and the result follows.

(3) If A is a bounded linear operator $\ell^p(\mathbb{Z}) \to \ell^p(\mathbb{Z})$ for all $p \in [1,\infty]$, then $\|A\|_p \le \|A\|_{p_0}^{1-t}\|A\|_{p_1}^t$ provided that $\frac{1}{p} = (1-t)\frac{1}{p_0} + t\frac{1}{p_1}$. This is a consequence of the Riesz–Thorin interpolation theorem [BL, Thm. 1.1.1]. Apply this result to $A = T^j$, take jth roots and let $j \to \infty$. It follows that $v \mapsto \log_2 \rho_{1/v}(T)$ is convex, and s_0 is concave.

(4) Clearly, $\|T^j\|_1 = \|T^j\varepsilon_0\|_1 \ge \sum_k (T^j\varepsilon_0)(k) = 2^j$, so $s_0(1) \le M$ and we know from above that s_0 is nondecreasing. □

Remark. In view of Theorem 3.2, the properties (1)–(3) of Proposition 3.4 contain no surprises when Cohen's criterion is satisfied. (1) holds because of the compactness of the support of g, (2) is a consequence of the Besov embedding theorem $B^{s,q}_{1/(v+h)}(\mathbb{R}) \subset B^{s-h,q}_{1/v}(\mathbb{R})$, [P, p. 63] and (3) follows from interpolation of Besov spaces as described in [P, p. 106] or [BL, §6.4].

4. Estimates for $s_0(0)$. Having seen that the regularity of g is determined by the spectral radius $\rho_p(T)$ of the operator $T = T_{\widetilde{m}}$ acting on $\ell^p(\mathbb{Z})$, the natural task is now to find upper and lower bounds on $\rho_p(T)$. In this section we consider the case $v = \frac{1}{p} = 0$, that is, $p = \infty$. Remember that if $s > 0$ is not an integer $B^{s,\infty}_\infty(\mathbb{R})$ is exactly the Hölder space $C^s(\mathbb{R})$. As a consequence of Theorem 3.2, the critical exponent $s_0(0) = M - \log_2 \rho_\infty(T)$ is also the critical Hölder exponent in the sense that $g \in C^s(\mathbb{R})$ implies $s \le s_0(0)$ and is implied by $s < s_0(0)$. To analyze the limit case $s = s_0(0)$, one has to back to Theorem 2.5. By Lemma 3.3, this reduces to a more refined study of $\|T^j\|_\infty$ as $j \to \infty$. Apparently, Theorem 3.2 does not give an exact criterion for g to be continuous or to have n continuous derivatives. However, we will see in §7 that this problem has an easy solution.

The first upper bound for $\rho_\infty(T)$ is furnished by the spectral radius formula, since $\rho_\infty(T) \le \|T^j\|_\infty^{1/j}$ for all $j = 1, 2, \ldots$. Here, the operator norm can be found explicitly: Regarding T^j as an infinite matrix with entries $a(k,l) = T^j\varepsilon_0(k - 2^j l)$ we have

$$\|T^j\|_\infty = \sup_{k\in\mathbb{Z}} \sum_{l\in\mathbb{Z}} |a(k,l)| = \max_{k\in\{0,1,\ldots,2^j-1\}} \sum_{l\in\mathbb{Z}} |T^j\varepsilon_0(k - 2^j l)|. \tag{4.1}$$

Before we discuss other means of finding upper bounds for $\rho_\infty(T)$, we turn to the lower bounds. A natural way to bound $\rho_\infty(T)$ from beneath would be to seek for elements in the spectrum of $T : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$. If we could restrict T to a finite-dimensional subspace of $\ell^\infty(\mathbb{Z})$, then this restriction would be just a matrix and the modulus of its largest eigenvalue would bound $\rho_\infty(T)$ from beneath.

However, T fail to leave the simplest finite-dimensional subspaces of $\ell^\infty(\mathbb{Z})$ invariant. These candidates are subspaces of finitely supported sequences and subspaces corresponding to finitely supported measures on the Fourier side. The obstruction is the "expansive" nature of T due to the discrete dilation D_2 in its definition. For this reason, we switch to the adjoint operator.

First define S_m associated to $m(\xi) = \sum_k c_k e^{-ik\xi}$ by

$$S_m y(k) = \sum_l 2c_{2k-l} y(l) = D_{\frac{1}{2}}(2c * y)(l), \tag{4.2}$$

where $D_{1/2}y(k) = y(2k)$. Let $\langle\cdot,\cdot\rangle$ be the duality bracket extending the scalar product

in $\ell^2 = \ell^2(\mathbb{Z}, \mathbb{C})$, that is,

$$\langle y, z \rangle = \sum_{k \in \mathbb{Z}} y(k)\overline{z}(k).$$

Then the adjoint of D_2 is $D_{1/2}$, and the adjoint of a convolution with a sequence, is the convolution with the reversed complex conjugated sequence. For the operators T_m and S_m this gives the next lemma.

LEMMA 4.1. *The adjoint of T_m is $T_m^* = S_{\overline{m}}$. That is, if $1 \le p, p' \le \infty$ with $\frac{1}{p} + \frac{1}{p'} = 1$, $y \in \ell^p$ and $z \in \ell^{p'}$, then*

$$\langle T_m y, z \rangle = \langle y, S_{\overline{m}} z \rangle.$$

If $1 \le p, p' \le \infty$ with $\frac{1}{p} + \frac{1}{p'} = 1$, then $\|y\|_p = \sup\{|\langle y, z \rangle| \mid z \in \ell^{p'} \wedge \|z\|_{p'} = 1\}$. Hence, by Lemma 4.1,

$$\|T^j\|_p = \|S^j\|_{p'}, \tag{4.3}$$

where $T = T_{\widetilde{m}}$ and $S = S_{\widetilde{m}}$. (A priori with $S = S_{\widetilde{\widetilde{m}}}$, but this complex conjugation makes no difference.) In the present case, we can therefore study the action of S on $\ell^1(\mathbb{Z})$ instead of the action of T on $\ell^\infty(\mathbb{Z})$.

For $n > 0$, let V_n denote the n-dimensional subspace of $\ell^1(\mathbb{Z})$ consisting of sequences supported by $\{0, 1, \ldots, n-1\}$. Assuming the support of the coefficients $\widetilde{c}$ of $\widetilde{m}$ is contained in V_L we find that $S(V_n) \subset V_n$ as long as $n \ge L - 1$, (choose $L \ge 2$). Moreover, if y is a finitely supported eigenvector for S corresponding to an eigenvalue $\lambda \ne 0$, one finds by iteration that $y \in V_L$. A first lower bound for $\rho_1(S)$ is then the spectral radius of the restriction of S to V_L, which is just an $L \times L$-matrix. In the case $L = 3$ this matrix is

$$\begin{pmatrix} 2\widetilde{c}_0 & 0 & 0 \\ 2\widetilde{c}_2 & 2\widetilde{c}_1 & 2\widetilde{c}_0 \\ 0 & 0 & 2\widetilde{c}_2 \end{pmatrix}.$$

Clearly, the spectral radius of this matrix is the maximum of the spectral radii of the two submatrices

$$S_0 = \begin{pmatrix} 2\widetilde{c}_0 & 0 \\ 2\widetilde{c}_2 & 2\widetilde{c}_1 \end{pmatrix} \quad \text{and} \quad S_1 = \begin{pmatrix} 2\widetilde{c}_1 & 2\widetilde{c}_0 \\ 0 & 2\widetilde{c}_2 \end{pmatrix}.$$

In the general case we define these to be representations of the restrictions $S_{|V_{L-1}}$ and $\tau_{-1} S \tau_{1|V_{L-1}}$, respectively, in the natural basis $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{L-2}$ for V_{L-1}.

While S is not translation invariant, we have

$$S\tau_{2k} = \tau_k S, \qquad k \in \mathbb{Z}. \tag{4.4}$$

To get better lower bounds for $\rho_1(S)$, we can therefore also use the eigenvalues of $S^j \tau_{-k}$. Indeed, if λ is such an eigenvalue, and y is the corresponding eigenvector, then for all $n = 1, 2, \ldots$,

$$\lambda^n y = (S^j \tau_{-k})^n y = S^{jn} \tau_{-k_n} y,$$

for some integer k_n. By the isometry of translations, this yields $\|S^{jn}\|_1 \ge |\lambda|^n$ so $\rho_1(S) \ge |\lambda|^{1/j}$.

Another consequence of (4.4) is that $S^j \tau_{-k_1}$ is similar to $S^j \tau_{-k_2}$ whenever $k_1 - k_2 \in (2^j - 1)\mathbb{Z}$. Hence, we only need to consider the spectrum of $S^j \tau_{-k}$ for $k \in \{0, 1, \ldots, 2^j - 2\}$. If y is a finitely supported eigenvector for $S^j \tau_{-k}$ corresponding

to an eigenvalue different from zero, iteration gives $y \in V_{L-1}$ in all the cases $k \in \{1, \dots, 2^j - 2\}$, and $y \in V_L$ for $k = 0$. On V_{L-1}, we find from (4.4) that

$$S^j \tau_{-k} \sim S_{d_1} S_{d_2} \dots S_{d_j}, \tag{4.5}$$

where $k = \sum_{l=1}^{j} d_l 2^{j-l}$ with $d_l \in \{0, 1\}$. Equivalently, $0.d_1 d_2 \dots d_j$ is the finite binary expansion of $k2^{-j}$.

By a separate treatment of the case $k = 0$, we obtain that the set of nonzero eigenvalues corresponding to finitely supported eigenvectors for $S^j \tau_{-k}$ with fixed j and all $k \in \mathbb{Z}$ is exactly the set of nonzero eigenvalues for the collection of 2^j matrices $S_{d_1} S_{d_2} \dots S_{d_j}$, $d_k \in \{0, 1\}$.

Returning to the formula (4.1) for $\|T^j\|_\infty$ we see that

$$\|S^j\|_1 = \max_{k \in \{0,1,\dots,2^j-1\}} \|S^j \tau_{-k} \varepsilon_0\|_1 = \max_{d_k \in \{0,1\}} \|S_{d_1} S_{d_2} \dots S_{d_j} \varepsilon_0\|_1. \tag{4.6}$$

We collect the obtained results in the next proposition.

PROPOSITION 4.2. *Define u_j and l_j for each $j = 1, 2, \dots$ by*

$$u_j^j = \max_{k \in \{0,1,\dots,2^j-1\}} \sum_{l \in \mathbb{Z}} |T^j \varepsilon_0(k + 2^j l)| = \max_{d_k \in \{0,1\}} \|S_{d_1} S_{d_2} \dots S_{d_j} \varepsilon_0\|_1,$$

$$l_j^j = \max_{d_k \in \{0,1\}} \rho(S_{d_1} S_{d_2} \dots S_{d_j}).$$

Then $\|T^j\|_\infty = u_j^j$ and $\|T^{jn}\|_\infty \geq l_j^{jn}$ for all $j, n = 1, 2, \dots$. As a consequence, we have $\sup_j l_j \leq \rho_\infty(T) = \inf_j u_j$.

Remarks. Clearly, $u_j = \max_{d_k \in \{0,1\}} \|S_{d_1} S_{d_2} \dots S_{d_j}\|_1^{1/j}$ and if we define the joint spectral radius of two matrices as in [DL2] to be $\hat{\rho}(S_0, S_1) = \limsup_j u_j$, we can exchange the 1-norm with any another matrix norm without altering the definition. In the present case, we see that this joint spectral radius is just the ordinary $\ell^1(\mathbb{Z})$-spectral radius of S. Due to a result of Berger and Wang [BW] we have also $\hat{\rho}(S_0, S_1) = \limsup_j l_j$. The last inequality of Proposition 4.2 is therefore never strict.

From Proposition 4.2 and Theorem 2.5 we get Corollary 4.3.

COROLLARY 4.3. *Assume $\widetilde{m}$ satisfies Cohen's criterion, let $j \in \{1, 2, \dots\}$ and fix notations as in Proposition 4.2. Then*

(1) *For $q < \infty$, we have $g \in B_\infty^{s,q}(\mathbb{R})$ if $s < M - \log_2 u_j$, but not if $s \geq M - \log_2 l_j$.*
(2) *$g \in B_\infty^{s,\infty}(\mathbb{R})$ if $s \leq M - \log_2 u_j$, but not if $s > M - \log_2 l_j$.*

In particular, if $s > 0$ is not an integer, then $g \in C^s(\mathbb{R})$ if $s \leq M - \log_2 u_j$, but not if $s > M - \log_2 l_j$.

Remarks. Rioul introduced the use of (4.1) for Hölder estimates in [R]. With slightly different sets of assumptions and definitions of u_j, the results of Proposition 4.2 and Corollary 4.3 about Hölder regularity are also well known from [DL2], [CH2], and [W].

The convergence of $M - \log_2 u_j$ towards $s_0(0)$ can be rather slow. To improve the upper bound for $\rho_\infty(T)$ given by Proposition 4.2, different techniques can be applied. First we describe a method due to Daubechies and Lagarias [DL2]. (See also [CH1] for applications of this trick.)

Suppose $\|S_{d_1} S_{d_2} \dots S_{d_j}\|_1^{1/j} \leq u$ for a finite collection $\mathcal{W}$ of binary "words" $(d_1, d_2, \dots, d_j)$, complete in the sense that every binary sequence $d : \mathbb{N} \to \{0, 1\}$ can be written $d = (w_1, w_2, \dots, w_k, \dots)$ with $w_k \in \mathcal{W}$. If j_0 is the largest word length, then $u_j^j \leq \|S^{j_0}\|_1 u^j$ for all $j = 1, 2, \dots$, so $\rho_\infty(T) \leq u$.

Another method is described by the following proposition, which is based on the idea of modeling the dynamics of $u_j^j = \|S^j\|_1$ by a second-order recursion. We will see in §9 that Proposition 4.4 can be quite useful in sufficiently simple situations. Of course, generalization of the method to recursions of order higher than two is possible.

PROPOSITION 4.4. *For each $d = (d_1, d_2) \in \{0,1\}^2$, let the vectors $\beta_d, \gamma_d, \eta_d \in V_{L-1} \simeq \mathbb{C}^{L-1}$ be such that $S_{d_1} S_{d_2} \varepsilon_0 = S_0 \beta_d + S_1 \gamma_d + \eta_d$. Put $a_d = \|\beta_d\|_1 + \|\gamma_d\|_1$, $b_d = \|\eta_d\|_1$ and*

$$u = \max_{d \in \{0,1\}^2} \left\{ \frac{a_d}{2} + \sqrt{\left(\frac{a_d}{2}\right)^2 + b_d} \right\}.$$

Then $\|T^j\|_\infty = \|S^j\|_1 \le Cu^j$ for all $j = 1, 2, \ldots$ and some positive constant C. Consequently, $\rho_\infty(T) \le u$.

Proof. Since $\|S^j\|_1 = \max_{d_k \in \{0,1\}} \|S_{d_1} S_{d_2} \ldots S_{d_j} \varepsilon_0\|_1$, checking the four cases for (d_{j-1}, d_j) and applying the triangle inequality yields

$$\|S^j\|_1 \le \max_{d \in \{0,1\}^2} \{a_d \|S^{j-1}\|_1 + b_d \|S^{j-2}\|_1\}, \tag{4.7}$$

for $j = 3, 4, \ldots$.

Observe that $u^2 - a_d u - b_d \ge 0$ for all $d \in \{0,1\}^2$, and $u > 0$ since $S \ne 0$. Define $y(j) = \|S^j\|_1 - Cu^j$ with $C > 0$ large enough to have $y(j) \le 0$ for $j \in \{1, 2\}$. Then (4.7) still holds when $\|S^k\|_1$ is replaced by $y(k)$ for $k \in \{j, j-1, j-2\}$. By induction, $y(j) \le 0$ for all $j = 1, 2, \ldots$. □

Finally, we describe a special case where $\rho_\infty(T)$ can be found from a single finite dimensional spectral radius. This result is essentially due to Deslauriers and Dubuc [DD2]. See also [R] and [CD] for related results.

PROPOSITION 4.5. *Assume $e^{ik\xi} \widetilde{m}(\xi) \ge 0$ for all ξ and some integer k. Let W be a finite dimemsional subspace of $\ell^1(\mathbb{Z})$ such that $S(W) \subset W$ and $\varepsilon_k \in W$, and let $\|\cdot\|$ denote the operator norm induced by any fixed norm on W. Then there are constants $0 < C_1 \le C_2 < \infty$ such that*

$$C_1 \|(S_{|W})^j\| \le \|S^j\|_1 \le C_2 \|(S_{|W})^j\|$$

for all $j = 1, 2, \ldots$. In particular, $\rho_1(S) = \rho(S_{|W})$.

Proof. Since all norms on W are equivalent it will be sufficient to prove the result with $\|\cdot\| = \|\cdot\|_1$. Also, we can assume $k = 0$, since the general case then follows from considering $r(\xi) = e^{ik\xi} \widetilde{m}(\xi)$ and $S_r = S\tau_{-k} = \tau_{2k} S \tau_{-2k}$.

The inequality $\|(S_{|W})^j\|_1 \le \|S^j\|_1$ is trivial and the last statement of the proposition follows from the spectral radius formula.

The main observation is that when $\widetilde{m} \ge 0$ we see from

$$T^j \varepsilon_0(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi} \prod_{l=0}^{j-1} 2\widetilde{m}(2^l \xi)\, d\xi,$$

that $\|T^j \varepsilon_0\|_\infty = T^j \varepsilon_0(0) = \langle T^j \varepsilon_0, \varepsilon_0 \rangle = \langle \varepsilon_0, S^j \varepsilon_0 \rangle = S^j \varepsilon_0(0)$. From Lemma 3.3 and the hypotheses on W we have therefore a finite constant C such that $\|S^j\|_1 = \|T^j\|_\infty \le C \|(S_{|W})^j\|_1$. □

An immediate consequence of Proposition 4.5 with $W = V_L$ is that if $e^{ik\xi} \widetilde{m}(\xi)$ is nonnegative for some k then $\rho_1(S) = l_1 = \max\{\rho(S_0), \rho(S_1)\}$. In fact we have necessarily $L = 2k+1$ and we only have to consider the spectral radius of the restriction of S to $W = \operatorname{span}\{\varepsilon_1, \ldots, \varepsilon_{L-2}\}$, represented by a common submatrix of S_0 and S_1.

5. Calculating $s_0(\frac{1}{2})$. The case $p = 2$ is particularly simple, since $\ell^2(\mathbb{Z})$ is a Hilbert space and a convolution operator on $\ell^2(\mathbb{Z})$ has an easily computable norm. Our starting point is the next observation.

LEMMA 5.1. *With $S = S_{\widetilde{m}}$ we have $S^j S^{*j} y = S^j_{2|\widetilde{m}|^2}\varepsilon_0 * y$, for all $y \in \ell^2(\mathbb{Z})$ and $j = 1, 2, \ldots$.*

Proof. From Lemma 4.1 we know that $S^*_{\widetilde{m}} = T_{\widetilde{\widetilde{m}}}$, so $S^j S^{*j} = S^j T^j_{\widetilde{\widetilde{m}}}$. This operator is invariant under translation, as it can be seen from (4.4) and the corresponding adjoint formula. Therefore it is just the operator of convolution with $S^j(T^j_{\widetilde{\widetilde{m}}}\varepsilon_0) = D_{2^{-j}}(T^j_{\widetilde{m}}\varepsilon_0 * T^j_{\widetilde{\widetilde{m}}}\varepsilon_0) = D_{2^{-j}}(T^j_{2|\widetilde{m}|^2}\varepsilon_0) = S^j_{2|\widetilde{m}|^2}\varepsilon_0$. □

By Parseval's formula for Fourier series, a convolution operator $y \mapsto z * y$, where $z \in \ell^1(\mathbb{R})$, is continuous $\ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ with operator norm equal to $\|z\|_{C(\mathbb{T})} = \sup_\xi \left|\sum_k z(k)e^{-ik\xi}\right|$. Hence, Lemma 5.1 implies that

$$\|S^{2j}\|_2 = \|S^j S^{*j}\|_2 = \|S^j_{2|\widetilde{m}|^2}\varepsilon_0\|_{C(\mathbb{T})}. \tag{5.1}$$

Let W be a finite dimensional subspace of $\ell^1(\mathbb{Z})$, such that $S_{|\widetilde{m}|^2}(W) \subset W$ and $\varepsilon_0 \in W$. Because $|\widetilde{m}|^2 \geq 0$, the proof of Proposition 4.5 shows that we have the equivalence

$$C_1\, 2^j \|(A_W)^j\| \leq \|S^{2j}\|_2 \leq C_2\, 2^j \|(A_W)^j\|, \tag{5.2}$$

where we have denoted by A_W the restriction of $S_{|\widetilde{m}|^2}$ to W. $\|\cdot\|$ can be any fixed matrix norm. From (5.2) and the spectral radius formula we get $\rho_2(S)^2 = 2\rho(A_W)$ and the following theorem.

THEOREM 5.2. *With W and A_W as above, we have $s_0(\frac{1}{2}) = M - \log_4 \rho(A_W)$. When $\widetilde{m}$ satisfies Cohen's criterion and $q < \infty$ we have $g \in B_2^{s,q}(\mathbb{R})$ if and only if $s < s_0(\frac{1}{2})$.*

Remarks. If $\widetilde{c}$ has support in $\{0, 1, \ldots, L-1\}$ we can choose W to be the space spanned by $\{\varepsilon_{2-L}, \ldots, \varepsilon_{L-2}\}$. In this case A_W is represented by the matrix with entries

$$a_{kl} = 2(\widetilde{c} * \check{\widetilde{c}})(2k-l) = \sum_n 2\widetilde{c}_n \check{\widetilde{c}}_{n-2k+l}, \qquad k, l \in \{2-L, \ldots, L-2\}.$$

In the case $q = 2$, Theorem 5.2 gives a result about whether or not g belongs to the Sobolev space $H^s(\mathbb{R})$. Eirola showed in [E] that $\sup\{s \mid g \in H^s(\mathbb{R})\} = M - \log_4 \rho(A_W)$ under the hypothesis that $\widetilde{m}(\xi) \neq 0$ for all ξ. For a different proof of Theorem 5.2 in the case $q = 2$, see also [V]. The case where infinitely many $c_k \neq 0$ has been treated independently by Hervé [H] and Gripenberg [G].

We have $g \in B_2^{s,\infty}(\mathbb{R})$ if and only if $\|(\rho(A_W)^{-1}A_W)^j\|$ is bounded for $j \to \infty$. By the Jordan decomposition theorem, this is the case exactly when each eigenvalue λ for A_W, with $|\lambda| = \rho(A_W)$, has equal algebraic and geometric multiplicity.

A simple bound for $\|T^j\|_2$ can also be found from $T^j y = T^j\varepsilon_0 * D_{2^j} y$, which gives

$$\|T^j\|_2 \leq \|T^j\varepsilon_0\|_{C(\mathbb{T})} = 2^j \sup_\xi \prod_{l=0}^{j-1} |\widetilde{m}(2^l\xi)|$$

and consequently, we have the following.

PROPOSITION 5.3. *$s_0(\frac{1}{2}) \geq M - \frac{1}{2} - \frac{1}{j}\log_2\bigl(\sup_\xi \prod_{l=0}^{j-1} |\widetilde{m}(2^l\xi)|\bigr)$ for $j = 1, 2, \ldots$.*

Combining Proposition 3.4(2) and Proposition 5.3 for $j = 1$ we find that g is continuous if $|\widetilde{m}(\xi)| < 2^{M-1}$ for all ξ. This sufficient criterion for continuity was

proved in [D1] by considering the decay of the Fourier transform of g.

6. Estimates for $s_0(1)$. As in §§4 and 5, fix $T = T_{\widetilde{m}}$ and $S = S_{\widetilde{m}}$ with $\widetilde{m}$ defined by (3.1). From Lemma 3.3 we know that $\|T^j\|_1 = \|T^j\varepsilon_0\|_1$. The simplest lower bounds for $s_0(1)$ are therefore given by

$$s_0(1) \geq M + 1 - \frac{1}{j}\log_2 \|T^j\varepsilon_0\|_1, \qquad j = 1, 2, \ldots. \tag{6.1}$$

The right-hand side of (6.1) increases with j and converges to $s_0(1)$ as $j \to \infty$. In a special case, $s_0(1)$ can be found explicitly.

PROPOSITION 6.1. *If $\widetilde{m}$ has nonnegative coefficients $\widetilde{c}_k \geq 0$, then $s_0(1) = M$ and $g \in B_1^{M,\infty}(\mathbb{R})$, but $g \notin B_1^{M,q}(\mathbb{R})$ for any $q < \infty$ if $\widetilde{m}$ satisfies Cohen's criterion.*

Proof. Observe that $\|T^j\varepsilon_0\|_1 = \sum_k (T^j\varepsilon_0)(k) = 2^j$, and apply Theorem 2.5 with $p = 1$ and $s = M$. □

We know from (4.3), that instead of studying the action of T on $\ell^1(\mathbb{Z})$, we can study the action of S on $\ell^\infty(\mathbb{Z})$. To find lower bounds for $\rho_1(T) = \rho_\infty(S)$, it is then natural to search for eigenvectors of $S^j : \ell^\infty(\mathbb{Z}) \to \ell^\infty(\mathbb{Z})$.

Define $\omega_{\xi_0} \in \ell^\infty$ for each $\xi_0 \in \mathbb{R}$ by $\omega_{\xi_0}(k) = e^{i\xi_0 k}$, $k \in \mathbb{Z}$. A straightforward calculation shows that

$$S\omega_{\xi_0} = 2\widetilde{m}(\xi_0)\omega_{2\xi_0}. \tag{6.2}$$

Since $\omega_{\xi_0+2\pi} = \omega_{\xi_0}$, every j-cycle $\xi_1 \mapsto \xi_2 \mapsto \cdots \mapsto \xi_j \mapsto \xi_1$ for the map $\xi \mapsto 2\xi$ mod 2π gives rise to an eigenvector for S^j, namely ω_{ξ_1}:

$$S^j\omega_{\xi_1} = 2^j\left(\prod_{l=1}^{j}\widetilde{m}(\xi_l)\right)\omega_{\xi_1}.$$

These observations lead to the following collection of upper bounds for $s_0(1)$.

PROPOSITION 6.2. *Let $\xi_1 \mapsto \xi_2 \mapsto \cdots \mapsto \xi_j \mapsto \xi_1$ be a j-cycle for the map $\xi \mapsto 2\xi$ mod 2π. Then*

$$s_0(1) \leq s_j = M - \frac{1}{j}\sum_{l=1}^{j}\log_2|\widetilde{m}(\xi_l)|.$$

If $\widetilde{m}$ satisfies Cohen's criterion we have $g \notin B_1^{s_j,q}$ for $q < \infty$.

For a j-cycle as in Proposition 6.2 we have necessarily $2^j\xi_1 = \xi_1 + 2\pi k$ for some $k \in \mathbb{Z}$. That is, $\xi_1 \in \frac{2\pi}{2^j-1}\mathbb{Z}$. Apart from the trivial cycle $0 \mapsto 0$, the simplest example of such a cycle is $\frac{2\pi}{3} \mapsto -\frac{2\pi}{3} \mapsto \frac{2\pi}{3}$. A few others are: $\frac{2\pi}{7} \mapsto \frac{4\pi}{7} \mapsto -\frac{6\pi}{7} \mapsto \frac{2\pi}{7}$ and $\frac{2\pi}{5} \mapsto \frac{4\pi}{5} \mapsto -\frac{2\pi}{5} \mapsto -\frac{4\pi}{5} \mapsto \frac{2\pi}{5}$. Note that with $\mathcal{N} = \{e^{i\xi_l}\}_{l=1}^{j}$ we have $\mathcal{N}^2 = \mathcal{N}$ as in (3) of Proposition 2.4.

The idea of using such cycles to bound regularity from above is due to Cohen [C2]. Using the present notation he proved the estimate of Proposition 6.2 with $s_0(1)$ replaced by $s_0(0)$, under the further hypothesis that $|\widetilde{m}(\pi)| > 1$.

For each $j = 1, 2, \ldots$, we can write the set $\{\frac{2\pi k}{2^j-1} \mid k = 0, 1, \ldots, 2^j - 2\}$ as a disjoint union of cycles. The bound

$$s_0(1) \leq M - \frac{1}{2\pi}\int_{-\pi}^{\pi}\log_2|\widetilde{m}(\xi)|\,d\xi \tag{6.3}$$

then follows from Proposition 6.2 by letting $j \to \infty$. However, the individual bounds s_j of Proposition 6.2 tend to give better results than (6.3), which averages out the good estimates with the bad.

In [CC], other results relating the regularity of g to invariant measures for the map $\xi \mapsto 2\xi \mod 2\pi$ are obtained.

Recall that Proposition 3.4 offers a variety of inequalities from which upper and lower bounds for $s_0(v)$ for different v can be related. For example, if $s_0(\frac{1}{2})$ is known from Theorem 5.2, a lower bound for $s_0(0)$ can imply an upper bound for $s_0(1)$ via the concavity of the graph of $v \mapsto s_0(v)$. Also, the inequality $s_0(\frac{1}{2}) \le s_0(1)$ actually performs better than (6.1) for finite j in many cases.

When the coefficients (c_k) are real, there is an interesting connection between the the fractal (box counting) dimension of the graph of g and the critical exponent $s_0(1)$, as described in [DJ]. With the present notation, we have the result $\dim(\text{graph}\, g) = 2 - s_0(1)$, if g is continuous, $0 < s_0(1) < 1$ and Cohen's criterion is satisfied.

7. Continuity, differentiability and integrability. Theorem 3.2 does not seem to give an exact criterion for g to be continuous, or to have n continuous derivatives, $g \in C^n(\mathbb{R})$. Nor do we automatically obtain criteria for g to be integrable or to have n integrable derivatives, $g \in W^{n,1}(\mathbb{R})$. However, the following theorem will remove these apparent weaknesses. With different notation and methods of proof, the results below concerning $C^n(\mathbb{R})$ are well known; see [CDM], [CH2], [DDD], [DGL], [MP], and [W].

THEOREM 7.1. *Put $T = T_{\widetilde{m}}$, and assume $\widetilde{m}$ sastisfies Cohen's criterion. For every $n = 0, 1, 2, \ldots$ we have then the following.*

(1) *$g \in C^n(\mathbb{R}) \Leftrightarrow s_0(0) > n$. Equivalently, if there exists $j = 1, 2, \ldots$ such that $\|(2^{n-M}T)^j\|_\infty < 1$.*

(2) *$g \in W^{n,1}(\mathbb{R}) \Leftrightarrow s_0(1) > n$. Equivalently, if there exists $j = 1, 2, \ldots$ such that $\|(2^{n-1-M}T)^j\|_1 < 1$.*

Consequently, if $g \in C^n(\mathbb{R})$, then $g \in C^{n+\epsilon}(\mathbb{R})$ for some $\epsilon > 0$, and $g \in W^{n,1}(\mathbb{R})$ implies $g \in W^{n,p}(\mathbb{R})$ for some $p > 1$.

Proof. Let $\psi \in \mathcal{S}(\mathbb{R})$ have a Fourier transform with compact support disjoint from the origin, as in Corollary 2.2. For a sufficiently small $h > 0$, we can then write $\psi(x) = \theta(x) - \theta(x + h)$ where $\hat{\theta}$ has the same smoothness and support properties as $\hat{\psi}$. Explicitly, put $\hat{\theta}(\xi) = \hat{\psi}(\xi)/(1 - e^{ih\xi})$.

Assume g is integrable. With α_j defined from ψ as in Lemma 2.1 we then get,

$$2^{-j/2}\alpha_j(k) = \int g(x)\overline{\psi}(2^j x - k)\, dx = \int \big(g(x) - g(x - 2^{-j}h)\big)\overline{\theta}(2^j x - k)\, dx. \tag{7.1}$$

Summing over k we obtain $2^{-j/2}\|\alpha_j\|_1 \le C\|g - \tau_{(2^{-j}h)}g\|_{L^1}$ with the finite constant $C = \sup_x \sum_k |\theta(x - k)|$. It is well known that $a \mapsto \tau_a g$ is continuous $\mathbb{R} \to L^1(\mathbb{R})$ whenever $g \in L^1(\mathbb{R})$, so we conclude that $2^{-j/2}\|\alpha_j\|_1 \to 0$ for $j \to \infty$. By a combination of the Lemmas 2.1, 2.6, and 3.3, it follows that $\|T^j\|_1 2^{-j(M+1)} \to 0$ for $j \to \infty$ so $s_0(1) > 0$.

On the other hand, if $s_0(1) > 0$ then Theorem 3.2 implies $g \in B_1^{0,1}(\mathbb{R})$ and therefore $g \in L^1(\mathbb{R})$ by (1.7). Even better, we have $s_0(v) > 0$ for some $v < 1$ by (2) of Proposition 3.4 so $g \in B_p^{0,1}(\mathbb{R}) \subset W^{0,p}(\mathbb{R}) = L^p(\mathbb{R})$ for some $p > 1$. We have just proved (2) in the case $n = 0$. The general case follows easily by partial integration in (7.1).

Suppose now that g is continuous. Then (7.1) gives the inequality $2^{j/2}\|\alpha_j\|_\infty \le \|\theta\|_{L^1} \sup_x |g(x) - g(x - 2^{-j}h)|$. Since g has compact support, it is also uniformly continuous and we see that $2^{j/2}\|\alpha_j\|_\infty \to 0$ for $j \to \infty$. This leads to $s_0(0) > 0$, after using the same lemmas as above. The special case $n = 0$ of (1) is now obvious, and

the general case follows again from partial integration in (7.1). □

From (4) of Proposition 3.4 we can now derive a result about the necessity of zeros at $\xi = \pi$.

COROLLARY 7.2. *Assume m satisfies Cohen's criterion. Then $g \in W^{n,1}(\mathbb{R})$ implies $M \geq n+1$. In particular, $m(\pi) = 0$ if g is integrable.*

The last statement is sharp in the sense that $m(\pi) = 0$ is not necessary for g to be a complex measure: g is the Dirac mass at $x = 0$ if $m(\xi) = 1$. With the help of (4) of Proposition 2.4, Corollary 7.2 can be extended to the case where $m(\xi)$ and $m(\xi+\pi)$ are never both zero, but this condition can not be omitted. Indeed, $m(\xi) = \frac{1+e^{-i2\xi}}{2}$ corresponds to $g = \frac{1}{2}\mathbf{1}_{[0,2]} \in L^1(\mathbb{R})$.

Note that when $g \in W^{n,1}(\mathbb{R})$ with $n \geq 1$, the $(n-1)$th derivative of g is an absolutely continuous function, differentiable almost everywhere, by the Lebesgue differentiation theorem. A sufficient criterion for differentiability almost everywhere is therefore contained in Theorem 7.1. For a detailed treatment of pointwise and local regularity we refer to [DL2].

8. Quality of convergence. Until now we have considered the distribution solution g to (1.1) as given by Theorem 1.1. When Cohen's criterion is satisfied, the regularity of g is then measured exactly by the critical exponent $s_0(v)$ defined in §3. However, g can also be regarded as a fixed point for the operator

$$P_m f(x) = \sum_k 2c_k f(2x-k). \tag{8.1}$$

Choosing a sufficiently nice initial function f_0, one could hope that $P_m^n f_0 \to g$ as $n \to \infty$. If f_0 is integrable with $\int f_0\,dx = 1$, this convergence holds at least in the sense of tempered distributions. We shall see (Theorem 8.2) that for a reasonable class of initial functions, convergence holds in $B_p^{s,q}(\mathbb{R})$ as long as $s < s_0(\frac{1}{p})$, and this result does not depend on Cohen's criterion. Loosely speaking, the critical exponent measures the quality of convergence rather than the regularity of the abstract solution g. When Cohen's criterion is not satisfied, it can happen that there is a gap between this quality and the regularity of g.

Assume $f_0 \in \mathcal{S}(\mathbb{R})$ is *interpolating*, that is, $f_0(k) = \varepsilon_0(k)$ for $k \in \mathbb{Z}$. Then (8.1) gives for $k \in \mathbb{Z}$ and $n = 1, 2, \ldots$

$$(P_m^n f_0)(2^{-n}k) = (T_m^n \varepsilon_0)(k). \tag{8.2}$$

Therefore results about the convergence of the iterates $P_m^n f_0$ will lead to corresponding results about discrete approximations to $g(2^{-n}k)$. In the applications, one never sees g, only the iterates $T_m^n \varepsilon_0$.

DEFINITION 8.1. *Let X be a normed space with $\mathcal{S}(\mathbb{R}) \subset X \subset \mathcal{S}'(\mathbb{R})$. We say that P_m* converges in X, *if $P_m^n f_0 \to g$ in X as $n \to \infty$ for all $f_0 \in \mathcal{S}(\mathbb{R})$ such that $\hat{f}_0(0) = 1$ and $\operatorname{supp} \hat{f}_0 \subset [-2\pi, 2\pi]$.*

The class of permitted initial functions described by Definition 8.1 certainly contains the Meyer scaling function φ, by (1.5). For a permitted interpolating function, one could take $\hat{f}_0 = |\hat{\varphi}|^2$ to obtain $\sum_k \hat{f}_0(\xi + 2\pi k) = 1$ since $\{\tau_k \varphi\}_{k \in \mathbb{Z}}$ is an orthonormal sequence in $L^2(\mathbb{R})$.

THEOREM 8.2. *Let $p, q \in [1, \infty]$. Then P_m converges in $B_p^{s,q}(\mathbb{R})$ if $s < s_0(\frac{1}{p})$.*

Proof. First of all, we know from Theorem 3.2 that $g \in B_p^{s,q}(\mathbb{R})$. Let f_0 be as in Definition 8.1 and put $u = f_0 - g$. We have to show that $P_m^n f_0 - g = P_m^n u$ converges to 0 in $B_p^{s,q}(\mathbb{R})$ as $n \to \infty$.

Assume this holds in the case $M = 0$. As usual we can write $g = \mathbf{1}_{[0,1]}^{*M} * \widetilde{g}$, where $\widetilde{g}$ is defined from $\widetilde{m}$ in the sense of Theorem 1.1. By the support properties of $\hat{f}_0$ we can also write $f_0 = \mathbf{1}_{[0,1]}^{*M} * \widetilde{f}_0$, where $\widetilde{f}_0$ is still an allowed initial function in the sense of Definition 8.1. The critical exponent corresponding to $\widetilde{m}$ is $s_0(\frac{1}{p}) - M$, and with $\widetilde{u} = \widetilde{f}_0 - \widetilde{g}$ the assumption implies that $P_{\widetilde{m}}^n \widetilde{u}$ converges to 0 in $B_p^{s-M,q}(\mathbb{R})$.

The identity $(1+z)(1-z) = 1 - z^2$ leads to

$$P_{\mu m}(\mathbf{1}_{[0,1]} * f) = \mathbf{1}_{[0,1]} * P_m f,$$

where $\mu(\xi) = \frac{1+e^{-i\xi}}{2}$. In our case this gives $P_m^n u = \mathbf{1}_{[0,1]}^{*M} * P_{\widetilde{m}}^n \widetilde{u}$. It remains only to observe that convolution with $\mathbf{1}_{[0,1]}$ defines a continuous operator $B_p^{t,q}(\mathbb{R}) \to B_p^{t+1,q}(\mathbb{R})$ in order to conclude that $P_m^n u$ converges to 0 in $B_p^{s,q}(\mathbb{R})$.

Without loss of generality, we can therefore suppose $M = 0$. Also, it will be sufficient to show the convergence for $q = \infty$ since $B_p^{s_1,\infty}(\mathbb{R})$ is continuously embedded in $B_p^{s,q}(\mathbb{R})$ for $s_1 > s$; see [P]. Among equivalent norms on $B_p^{s,\infty}(\mathbb{R})$, we choose the one induced by the wavelet characterization of Theorem 1.2. If the wavelet coefficients of $f \in B_p^{s,\infty}(\mathbb{R})$ are $\beta^f(k) = \langle f, \tau_k \varphi \rangle$ and $\alpha_j^f(k) = \langle f, \psi_{jk} \rangle$, then

$$\|f\|_{B_p^{s,\infty}} = \|\beta^f\|_p + \sup_{j \geq 0} (2^{j(\frac{1}{2} - \frac{1}{p} + s)} \|\alpha_j^f\|_p). \tag{8.3}$$

Putting $\beta^n = \beta^{P_m^n u}$ and $\alpha_j^n = \alpha_j^{P_m^n u}$, simple calculations as those leading to Lemma 2.1 give

$$\alpha_j^n = \begin{cases} 2^{-j/2} \alpha_0^{n-j} * T^j \varepsilon_0, & 0 \leq j \leq n, \\ 2^{-n/2} \alpha_{j-n}^0 * D_{2^{j-n}} T^n \varepsilon_0, & j > n. \end{cases} \tag{8.4}$$

Admit for the moment a result which we prove in the appendix.

LEMMA 8.3. *There is a constant C_0 such that $\|\beta^n\|_{p_1} + \|\alpha_0^n\|_{p_1} \leq C_0 \, 2^{-n}$ for all $p_1 \in [1, \infty]$ and $n = 1, 2, \ldots$.*

Then the first term of the right-hand side of (8.3) with $f = P_m^n u$ clearly tends to 0 as $n \to \infty$. Choosing $\Delta s > 0$ with $s + \Delta s < s_0(\frac{1}{p})$, we obtain that $\|T^j \varepsilon_0\|_p \leq C_1 2^{j((1/p) - s - \Delta s)}$ for some $C_1 > 0$, by the definition of the critical exponent. An application of Lemma 8.3 with $p_1 = 1$ in the first case of (8.4) gives us

$$2^{j(\frac{1}{2} - \frac{1}{p} + s)} \|\alpha_j^n\|_p \leq C_0 C_1 \, 2^{-(n-j)} 2^{-j\Delta s} \leq C_0 C_1 \, 2^{-n \min\{1, \Delta s\}}.$$

Since $\hat{\psi}(\xi) = 0$ for $|\xi| \leq \frac{2\pi}{3}$ and the support of the Fourier transform of $P_m^n f_0$ is included in $[-2^{n+1}\pi, 2^{n+1}\pi]$, it follows that $\alpha_j^n = -\alpha_j^g$ for $j \geq n+2$. In this case we have therefore

$$2^{j(\frac{1}{2} - \frac{1}{p} + s)} \|\alpha_j^n\|_p \leq 2^{-j\Delta s} \|g\|_{B_p^{s+\Delta s,\infty}} \leq 2^{-n\Delta s} \|g\|_{B_p^{s+\Delta s,\infty}}.$$

For the remaining case $j = n+1$, (8.4) gives

$$2^{j(\frac{1}{2} - \frac{1}{p} + s)} \|\alpha_j^n\|_p \leq 2^{\frac{1}{2} - \frac{1}{p} + s} \|\alpha_1^0\|_1 C_1 \, 2^{-n\Delta s}.$$

Collecting the different cases we see that

$$\|P_m^n u\|_{B_p^{s,\infty}} \leq C \, 2^{-n \min\{1, \Delta s\}}, \tag{8.5}$$

for some new constant $C > 0$, and the desired convergence follows. □

There are two important cases where Cohen's criterion is in fact necessary to have even only a reasonable convergence.

Dyadic interpolation. Assume

$$m(\xi) + m(\xi + \pi) = 1, \tag{8.6}$$

or equivalently, $c_{2k} = \frac{1}{2}\varepsilon_0(k)$. It is then an easy matter to verify that P_m preserves interpolating functions. In this case c_{2k+1} are the coefficients of an *interpolating subdivision scheme* [DLG], [DD1], and [DD2]. If P_m converges uniformly, (in supremum norm) we see using an interpolating initial function f_0 that g itself must be continuous and interpolating. Hence m satisfies Cohen's criterion by (2) of Proposition 2.4. But then we know from (1) of Theorem 7.1 that $s_0(0) > 0$ and we have shown all implications which are not trivial consequences of Theorem 8.2 in the next proposition.

PROPOSITION 8.4. *If m satisfies* (8.6), *the following four statements are equivalent and imply that m satisfies Cohen's criterion:*

(1) *P_m converges uniformly (in sup-norm).*
(2) *$s_0(0) > 0$.*
(3) *P_m converges in $C^\epsilon(\mathbb{R})$ for some $\epsilon > 0$.*
(4) *g is a continuous interpolating function.*

Biorthogonal filters. This second case is important for the construction of biorthogonal and orthonormal bases of wavelets in $L^2(\mathbb{R})$, [C3], [VH], [D]. The assumption is here that m_1 and m_2 form a biorthogonal filter pair in the sense that

$$m_1(\xi)\overline{m_2}(\xi) + m_1(\xi + \pi)\overline{m_2}(\xi + \pi) = 1. \tag{8.7}$$

The basic observation is then that the integer translates of $P_{m_1} f_1$ and $P_{m_2} f_2$ form biorthogonal sequences in $L^2(\mathbb{R})$, provided this is the case for the integer translates of f_1 and f_2. (More precisely, $\langle \tau_k f_1, \tau_l f_2 \rangle_{L^2(\mathbb{R})} = \varepsilon_0(k-l)$ for $k, l \in \mathbb{Z}$.) To see this, just note that $m = m_1\overline{m_2}$ satisfies (8.6).

Taking the Meyer scaling function φ as initial function for both P_{m_1} and P_{m_2}, we see that if we have convergence in $L^2(\mathbb{R})$, the integer translates of g_1 and g_2 form biorthogonal sequences in $L^2(\mathbb{R})$. Again (2) of Proposition 2.4 allows us to conclude that both m_1 and m_2 satisfy Cohen's criterion. Since $L^2(\mathbb{R}) = B_2^{0,2}(\mathbb{R})$, Theorem 5.2 now gives that both $s_0^1(\frac{1}{2}) > 0$ and $s_0^2(\frac{1}{2}) > 0$ where $s_0^1(\frac{1}{2})$ and $s_0^2(\frac{1}{2})$ are critical exponents of m_1 and m_2, respectively.

The conjugate quadrature filter case is when we have $m_1 = m_2 = m$. The integer translates of $g_n = P_m^n \varphi$ then form an orthonormal sequence in $L^2(\mathbb{R})$ for each fixed $n \in \mathbb{N}$. In particular, Fatou's lemma applied to $|\hat{g}_n|^2$ gives that $\|g\|_{L^2} \leq \liminf_n \|g_n\|_{L^2} = 1$ so $g \in L^2(\mathbb{R})$. Now by Theorem 5.2, $s_0(\frac{1}{2}) > 0$ if m satisfies Cohen's criterion, so L^2-convergence is in fact equivalent to this criterion. With a slightly different definition of L^2-convergence, this result can be found in [C1]. Thus we have the following.

PROPOSITION 8.5. *If the symbols m_1 and m_2 satisfy* (8.7), *the following four statements are equivalent and imply that both m_1 and m_2 satisfy Cohen's criterion.*

(1) *P_{m_1} and P_{m_2} both converge in $L^2(\mathbb{R})$.*
(2) *$s_0^1(\frac{1}{2}) > 0$ and $s_0^2(\frac{1}{2}) > 0$.*
(3) *P_{m_1} and P_{m_2} both converge in $H^\epsilon(\mathbb{R})$ for some $\epsilon > 0$.*
(4) *The integer translates of g_1 and g_2 form biorthogonal sequences in $L^2(\mathbb{R})$.*

Moreover, if $m_1 = m_2 = m$, then (1)–(4) *hold if and only if m satisfies Cohen's criterion.*

Let us finally describe some discrete convergence properties inherited from Theorem 8.2. We expect from (8.2) that the sequence $T_m^n \varepsilon_0$ should be a good discrete approximation of $g(2^{-n}k)_{k\in\mathbb{Z}}$. We have the following result about the quality of this approximation. Note that $2^{nl}(\varepsilon_0 - \varepsilon_1)^{*l} * T_m^n \varepsilon_0$ is just a way of writing the lth divided difference of $T_m^n \varepsilon_0$. The natural "step size" is 2^{-n}.

COROLLARY 8.6. *Assume $s_0(0) > r$, where r is a nonnegative integer. Then $g \in C^r(\mathbb{R})$ and*

$$\sup_{k\in\mathbb{Z}} \left| \left(2^{nl}(\varepsilon_0 - \varepsilon_1)^{*l} * T_m^n \varepsilon_0\right)(k) - g^{(l)}(2^{-n}k) \right| = \mathrm{O}(2^{-n\min\{1,\Delta s\}})$$

for $n \to \infty$ whenever $l \in \{0, 1, \ldots, r\}$ and $0 < \Delta s < s_0(0) - l$.

Proof. Note that $\frac{d}{dx} \circ P_m = 2P_m \circ \frac{d}{dx}$ and $P_m \circ \tau_a = \tau_{a/2} \circ P_m$. From these rules we obtain

$$\frac{d^l}{dx^l}\left(P_m^n(\mathbf{1}_{[0,1]}^{*l} * f_0)\right) = 2^{nl}(\delta_0 - \delta_{2^{-n}})^{*l} * P_m^n f_0. \tag{8.8}$$

Let f_0 be an interpolating initial function in the class of Definition 8.1. Then $f_1 = \mathbf{1}_{[0,1]}^{*l} * f_0$ is still a permitted initial function. Choose $0 < \epsilon < 1$ such that $l + \epsilon + \Delta s < s_0(0)$. Since d^l/dx^l is continuous $B_\infty^{l+\epsilon,\infty}(\mathbb{R}) \to B_\infty^{\epsilon,\infty}(\mathbb{R}) = C^\epsilon(\mathbb{R})$, (8.5) of the proof of Theorem 8.2 with $s = l + \epsilon$ gives us a constant $C > 0$ such that

$$\sup_{x\in\mathbb{R}} \left| \frac{d^l(P_m^n f_1 - g)}{dx^l}(x) \right| \le C\, 2^{-n\min\{1,\Delta s\}}.$$

The corollary now follows by restriction to $x \in 2^{-n}\mathbb{Z}$ since (8.8) and (8.2) give

$$\frac{d^l P_m^n f_1}{dx^l}(2^{-n}k) = \left(2^{nl}(\varepsilon_0 - \varepsilon_1)^{*l} * T_m^n \varepsilon_0\right)(k). \qquad \square$$

Remarks. The upper bound $\mathrm{O}(2^{-n})$ for the speed of convergence in Corollary 8.6 comes from Lemma 8.3. If $\hat{f}_0^{(l)}(0) = \hat{g}^{(l)}(0)$ for $l = 0, 1, \ldots, q$, we can replace $\min\{1, \Delta s\}$ with $\min\{q + 1, \Delta s\}$ in the corollary. One way to achieve this is to use $\hat{f}_0 = R|\hat{\varphi}|^2$ in the above proof, where φ is the Meyer scaling function and $R(\xi) = \sum_k r(k)e^{-ik\xi}$ is a trigonometric polynomial with $R^{(l)}(0) = \hat{g}^{(l)}(0)$ for $l = 0, 1, \ldots, q$. (Recall that $\hat{\varphi} = 1$ in a neighbourhood of the origin.) f_0 is then no longer interpolating but $f_0(k) = r(k)$ for $k \in \mathbb{Z}$. In terms of discrete approximations, we consider the sequences $r * T_m^n \varepsilon_0$ in stead of $T_m^n \varepsilon_0$ for the approximation of $g(2^{-n}k)_{k\in\mathbb{Z}}$.

If $S_m r = r$ with $\sum_k r(k) = 1$ the described method works with $q = M - 1$, because then

$$\begin{cases} R(2\xi) = m(\xi)R(\xi) + m(\xi + \pi)R(\xi + \pi), \\ \hat{g}(2\xi) = m(\xi)\hat{g}(\xi), \end{cases}$$

and m vanishes up to order $M - 1$ at $\xi = \pi$.

If we know the values of g at the integers, a much more direct result can be obtained by putting $r(k) = g(k)$, since we then have

$$(r * T_m^n \varepsilon_0)(k) = g(2^{-n}k) \tag{8.9}$$

for $k \in \mathbb{Z}$. In fact (8.9) is a local version of what is called the nonlocal algorithm in [DL1]. Standard results about spline interpolation then give easy and powerful convergence results depending only on the regularity of g. Note that in this case $S_m r = r$ is a consequence of (1.1).

9. Examples. In this section we will apply the obtained results to three relatively simple and well known examples.

Example 9.1. Let $c = (\frac{1}{2}, \frac{1}{2})^{*2} * (\frac{1+\sqrt{3}}{2}, \frac{1-\sqrt{3}}{2})$. Cohen's criterion is trivially satisfied and $|m(\xi)|^2 + |m(\xi+\pi)|^2 = 1$. By (4) of Proposition 8.5 the integer translates of g form an orthonormal sequence in $L^2(\mathbb{R})$, and if we put

$$\psi(x) = \sum_k 2(-1)^{1-k} c_{1-k} g(2x-k),$$

then ψ is the first compactly supported orthonormal wavelet of Daubechies [D]. The regularity properties of ψ are exactly the same as those of g.

We are here in the situation described by Example 2.9 with $\widetilde{c} = (\frac{1+\sqrt{3}}{2}, \frac{1-\sqrt{3}}{2})$. In terms of the critical exponent,

$$s_0(v) = \begin{cases} 2 + v - v\log_2\left((1+\sqrt{3})^{\frac{1}{v}} + (\sqrt{3}-1)^{\frac{1}{v}}\right), & 0 < v \le 1, \\ 2 - \log_2(1+\sqrt{3}), & v = 0. \end{cases}$$

We know also that this exponent is exact in the sense that $g \in B_{\frac{1}{v}}^{s_0(v),q}(\mathbb{R})$ if and only if $q = \infty$. In particular, g is Hölder continuous with best possible exponent $2 - \log_2(1+\sqrt{3}) \approx 0.5500$. A well-known result from [DL2]. We have $s_0(\frac{1}{2}) = 1$ and $v \mapsto s_0(v)$ is strictly increasing. Hence, Theorem 5.2 gives that $g \in H^s(\mathbb{R})$ if and only if $s < 1$, and from (1.7) we see that g is absolutely continuous with derivative in L^p exactly when $p < 2$.

Thus, g is differentiable almost everywhere. For a description of the set of points where g fails to be differentiable we refer to the pointwise analysis of [DL2]. □

Example 9.2. Let $c = \frac{1}{2}(\beta, 1-\beta, 1-\beta, \beta)$ where β is a real parameter. This is exactly the real symmetric 4-coefficient case. The family was considered by de Rham in [dR] with $0 < \beta < \frac{1}{2}$, where the equation (1.1) arises from considering the limit curve of a symmetric "corner cutting" scheme. (Lending the modern terminology from [DGL].)

If m_1 is defined similarly from β_1 with $(1-2\beta)(1-2\beta_1) = 1$ it follows that

$$m(\xi)\overline{m_1}(\xi) + m(\xi+\pi)\overline{m_1}(\xi+\pi) = 1.$$

Hence m and m_1 define biorthogonal filters in the sense of Proposition 8.5. This filter pair was studied by Vetterli and Herley in [VH].

We have the factorization $c = (\frac{1}{2}, \frac{1}{2}) * (\beta, 1-2\beta, \beta)$. Fixing $M = 1$, we can then choose

$$\widetilde{m}(\xi) = 1 - 4\beta\sin^2(\xi/2).$$

For $\beta < \frac{1}{4}$ this $\widetilde{m}$ is strictly positive. For $\beta = \frac{1}{4}$ the solution g is a quadratic spline as in the case $M = 3$ of Example 2.8. When $\beta > \frac{1}{4}$, $\widetilde{m}$ has two real zeros in $[-\pi, \pi]$ located symmetrically around $\xi = 0$.

By (3) of Proposition 2.4, $\widetilde{m}$ fail to satisfy Cohen's criterion if and only if either $\widetilde{m}(\frac{\pi}{2}) = 0$ or $\widetilde{m}(\frac{\pi}{3}) = 0$. That is, if $\beta \in \{\frac{1}{2}, 1\}$. In the first case, $c = (\frac{1}{2}, \frac{1}{2}) * (\frac{1}{2}, 0, \frac{1}{2})$ giving $g = \frac{1}{2}\mathbf{1}_{[0,1]} * \mathbf{1}_{[0,2]}$. In the second case $c = (\frac{1}{2}, 0, 0, \frac{1}{2})$ and $g = \frac{1}{3}\mathbf{1}_{[0,3]}$.

It turns out that $s_0(0)$ can be found explicitly for all $\beta \in \mathbb{R}$. When $\beta \notin \{\frac{1}{2}, 1\}$ this critical exponent is exact in the sense that $g \in B_\infty^{s_0(0),q}(\mathbb{R})$ if and only if $q = \infty$.

The result is:

$$s_0(0) = \begin{cases} 2, & \beta = \frac{1}{4}, \\ 1 - \log_2(\beta + \sqrt{4\beta - 7\beta^2}), & \frac{1}{4} < \beta < \frac{1}{2}, \\ -\log_2 \beta, & \frac{1}{2} \le \beta \le 1, \\ -\log_2 |1 - 2\beta|, & \beta < 0 \vee \beta > 1. \end{cases} \tag{9.1}$$

The case $\beta = \frac{1}{4}$ is evident since $c = (\frac{1}{2}, \frac{1}{2})^{*3}$. To show the other cases we first compute, with notation as in §4,

$$S_0 = \begin{pmatrix} 2\beta & 0 \\ 2\beta & 2 - 4\beta \end{pmatrix}, \qquad S_1 = \begin{pmatrix} 2 - 4\beta & 2\beta \\ 0 & 2\beta \end{pmatrix}.$$

When $\beta < \frac{1}{4}$ we have $\widetilde{m} > 0$ and Proposition 4.5 immediately gives $s_0(0) = 1 - \log_2(2 - 4\beta)$ which is exact since $\|T^j_{\widetilde{m}}\|_\infty \sim (2 - 4\beta)^j$ for $j \to \infty$.

For $\beta \ge \frac{1}{2}$, we first compute l_1 of Proposition 4.2,

$$l_1 = \max\{2\beta, 4\beta - 2\}.$$

Next, we observe that

$$\begin{cases} S_0^2 \varepsilon_0 &= (2 - 2\beta) S_0 \varepsilon_0 + (8\beta^2 - 4\beta)\varepsilon_0, \\ S_1^2 \varepsilon_0 &= (2 - 4\beta)^2 \varepsilon_0, \\ S_0 S_1 \varepsilon_0 &= (2 - 4\beta) S_0 \varepsilon_0, \\ S_1 S_0 \varepsilon_0 &= (2 - 2\beta) S_0 \varepsilon_0 + (8\beta^2 - 4\beta)\varepsilon_1. \end{cases} \tag{9.2}$$

An application of Proposition 4.4 then gives $u = \max\{|1 - \beta| + 3\beta - 1, 4\beta - 2\} = l_1$ and $s_0(0) = 1 - \log_2(\max\{2\beta, 4\beta - 2\})$ which is exact (when $\beta \notin \{\frac{1}{2}, 1\}$) since $\|T^j_{\widetilde{m}}\|_\infty \sim u^j$ for $j \to \infty$.

The only remaining case of (9.1) is $\frac{1}{4} < \beta < \frac{1}{2}$. Here it turns out that

$$l_2 = \sqrt{\rho(S_0 S_1)} = \sqrt{2\beta\big(2 - 3\beta + \sqrt{4\beta - 7\beta^2}\big)} = \beta + \sqrt{4\beta - 7\beta^2}$$

and $l_2 > l_1$. If we replace the first and last line of (9.2) with

$$\begin{cases} S_0^2 \varepsilon_0 &= 2\beta S_0 \varepsilon_0 + (4\beta - 8\beta^2)\varepsilon_1, \\ S_1 S_0 \varepsilon_0 &= 2\beta S_0 \varepsilon_0 + (4\beta - 8\beta^2)\varepsilon_0, \end{cases}$$

Proposition 4.4 now gives $u = l_2$, completing the proof of (9.1).

The solid graph of Fig. 9.1 is a plot of $s_0(0)$ as a function of the parameter $\beta \ne \frac{1}{4}$. We know that g is continuous if and only if $0 < \beta < 1$, and g is continuously differentiable only in the case $\beta = \frac{1}{4}$ where it is a quadratic spline. In the interval $\frac{1}{4} < \beta < \frac{1}{2}$ there is a local minimum at $\beta = \frac{4+\sqrt{2}}{14} \approx 0.3867$ where $s_0(0) = 1 - \log_2(\frac{2+4\sqrt{2}}{7}) \approx 0.8706$.

When $(1 - 2\beta)(1 - 2\beta_1) = 1$ we see that β and β_1 can never both be between 0 and 1. Hence for the biorthogonal filters mentioned in the beginning of this example, the corresponding g and g_1 cannot both be continuous. A fact that was observed numerically in [VH].

Except in the case $\beta = \frac{1}{4}$, where Example 2.8 gives $s_0(\frac{1}{2}) = \frac{5}{2}$, the critical exponent $s_0(\frac{1}{2})$ is determined via Theorem 5.2 with $W = \operatorname{span}\{\varepsilon_0, \varepsilon_{-1} + \varepsilon_1\}$ by the

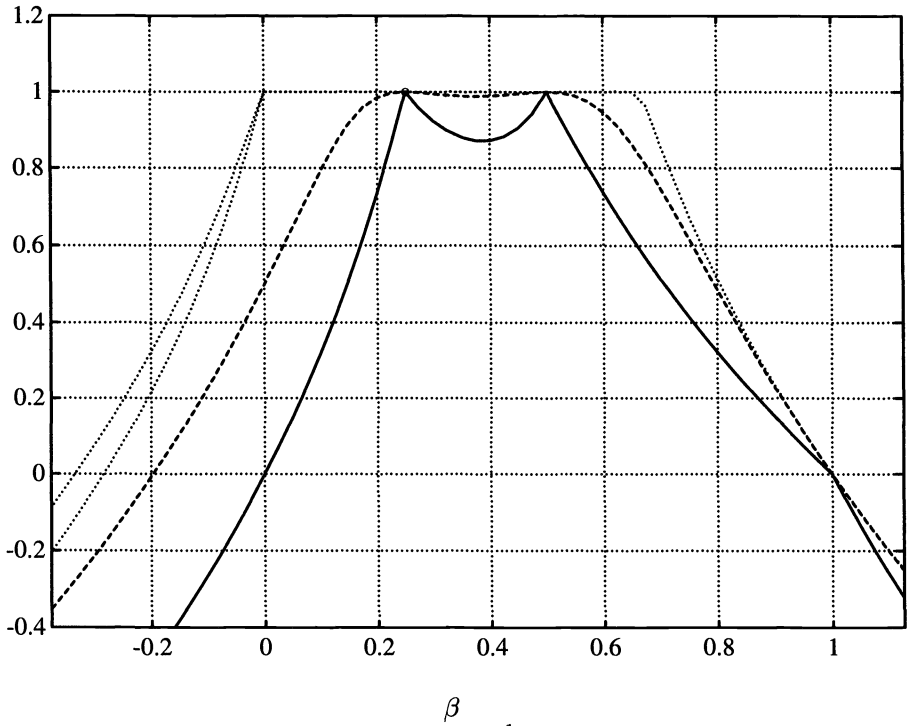


FIG. 9.1. *The critical exponents* $s_0(0)$ (*solid*), $s_0(\frac{1}{2})$ (*dashed*) *and bounds for* $s_0(1)$ (*dotted*) *as a function of* β *for the family of coefficients considered in Example* 9.2, $c = \frac{1}{2}(\beta, 1-\beta, 1-\beta, \beta)$. *For* $\beta = \frac{1}{4}$ *the result,* $s_0(v) = 2 + v$, *is not marked.*

spectral radius of the matrix

$$\begin{pmatrix} 12\beta^2 - 8\beta + 2 & 8\beta(1-2\beta) \\ 2\beta^2 & 4\beta(1-2\beta) \end{pmatrix}.$$

Computing this spectral radius yields, for $\beta \neq \frac{1}{4}$:

$$s_0(\tfrac{1}{2}) = 1 - \log_4 \left(2\beta^2 - 2\beta + 1 + \sqrt{(2\beta^2 - 2\beta + 1)^2 - 8\beta(1-2\beta)^3}\right). \tag{9.3}$$

The dashed graph on Fig. 9.1 is a plot of this critical exponent. Solving a third-order equation we see that g is square integrable if and only if $\beta_0 < \beta < 1$ where

$$\beta_0 = \frac{1}{6}\left(1 + \left(\frac{-43 + 9\sqrt{29}}{2}\right)^{1/3} - \left(\frac{43 + 9\sqrt{29}}{2}\right)^{1/3}\right) \approx -0.1963.$$

We saw above that for biorthogonal filters, the corresponding g and g_1 cannot both be continuous. However they are both square integrable exactly when $\beta_0 < \beta < \frac{\beta_0}{2\beta_0 - 1} \approx 0.1410$. The upper bounds for the regularity of g_1 when $g \in L^2(\mathbb{R})$ are $s_0(\frac{1}{2}) \approx 0.8969$ and $s_0(0) = \log_2(1 - 2\beta_0) \approx 0.4778$.

For the critical exponent $s_0(1)$, an application of (6.1) with $j = 1$ and Proposition 6.2 with the 2-cycle $\frac{2\pi}{3} \mapsto -\frac{2\pi}{3} \mapsto \frac{2\pi}{3}$ gives for $\beta \neq \frac{1}{4}$,

$$1 - \log_2(2|\beta| + |1 - 2\beta|) \leq s_0(1) \leq 1 - \log_2|1 - 3\beta|. \tag{9.4}$$

When $\beta = \frac{1}{4}$ we have $s_0(1) = 3$ and when $\beta \in [0, \frac{1}{2}] \setminus \{\frac{1}{4}\}$, Proposition 6.2 gives $s_0(1) = 1$ since then all elements of $\widetilde{c} = (\beta, 1-2\beta, \beta)$ are nonnegative.

We see from (9.3)–(9.4) and Theorem 7.1, treating the cases $\beta \in \{\frac{1}{2}, 1\}$ separately, that g is integrable if $-\frac{1}{4} < \beta \leq 1$ but not if $\beta \leq -\frac{1}{3}$ or $\beta > 1$. Also g is not absolutely continuous except in the cases $\beta \in \{\frac{1}{4}, \frac{1}{2}\}$. On Fig. 9.1, the dotted graphs give upper and lower bounds for $s_0(1)$ when $\beta < 0$, the exact value when $\beta \in [0, \frac{1}{2}] \setminus \{\frac{1}{4}\}$ and an upper bound when $\beta > \frac{1}{2}$. The lower bound are computed numerically from (6.1) with $j = 10$ and the upper bound is the one from (9.4). Numerical experiments suggest that no better upper bound can be obtained from Proposition 6.2 by testing all j-cycles up to order 7. For $\beta > \frac{1}{2}$ the lower bound of (6.1) with $j = 10$ is still very poor in comparison with the inequality $s_0(1) \geq s_0(\frac{1}{2})$ from (1) of Proposition 3.4. □

Example 9.3. Let $c = \frac{1}{2}(-w, 0, \frac{1}{2}+w, 1, \frac{1}{2}+w, 0, -w)$ where w is a real parameter. This case corresponds to a 4-point interpolation scheme for curve design considered by Deslauriers and Dubuc [DD1] and Dyn, Levin and Gregory [DLG].

We have the factorization $c = (\frac{1}{2}, \frac{1}{2})^{*2} * (-2w, 4w, 1-4w, 4w, -2w)$. Fixing $M = 2$ then gives

$$\widetilde{m}(\xi) = 1 + 8w(1 - \cos\xi)\cos\xi,$$

$\widetilde{m} > 0$ for $-\frac{1}{2} < w < \frac{1}{16}$, and $\widetilde{m}$ has at most 4 zeros in $[-\pi, \pi]$, symmetrically located around $\xi = 0$. From (3) of Proposition 2.4 and the fact that $c_{2k} = \frac{1}{2}\varepsilon_0(k)$ so $m(\xi) + m(\xi + \pi) = 1$, we see that Cohen's criterion fails to hold exactly when $\widetilde{m}(\frac{\pi}{3}) = 0$. That is, when $w = -\frac{1}{2}$. In this special case $c = (\frac{1}{2}, 0, 0, \frac{1}{2})^{*2}$ and therefore $g = \frac{1}{9}\mathbf{1}_{[0,3]}^{*2}$.

We will concentrate on estimates of $s_0(0)$. If $w = \frac{1}{16}$ then $c = (\frac{1}{2}, \frac{1}{2})^{*4} * (-\frac{1}{2}, 2, -\frac{1}{2})$ and $s_0(0) = 2$ as it can be seen by adding 3 to the result in (9.1) for $\beta = -\frac{1}{2}$. For $w \neq \frac{1}{16}$ the matrices S_0 and S_1 are

$$S_0 = \begin{pmatrix} -4w & 0 & 0 & 0 \\ 2-8w & 8w & -4w & 0 \\ -4w & 8w & 2-8w & 8w \\ 0 & 0 & -4w & 8w \end{pmatrix}, \qquad S_1 = \begin{pmatrix} 8w & -4w & 0 & 0 \\ 8w & 2-8w & 8w & -4w \\ 0 & -4w & 8w & 2-8w \\ 0 & 0 & 0 & -4w \end{pmatrix}.$$

With notation as in Proposition 4.2 we find

$$l_1 = \max\{8|w|, |1 \pm \sqrt{1-16w}|\} = \begin{cases} 1+\sqrt{1-16w}, & -\frac{1}{2} \leq w < \frac{1}{16}, \\ 4\sqrt{w}, & \frac{1}{16} < w \leq \frac{1}{4}, \\ 8|w|, & w < -\frac{1}{2} \vee w > \frac{1}{4}. \end{cases} \tag{9.5}$$

Proposition 4.5 and the special treatment of the case $w = \frac{1}{16}$ gives the result

$$s_0(0) = 2 - \log_2(1 + \sqrt{1-16w}) \quad \text{for} \quad -\tfrac{1}{2} \leq w \leq \tfrac{1}{16}. \tag{9.6}$$

This could also have been found with the methods of [DD2], see [R]. We see from Theorem 7.1 and (9.5) that g can only be continuous if $-\frac{1}{2} \leq w < \frac{1}{2}$ and continuously differentiable if $0 < w < \frac{1}{4}$. A well-known fact [DLG]. We will now apply the method of Proposition 4.4 to show that g is continuous if and only if $-\frac{1}{2} \leq w < \frac{1}{2}$. The problem is to assert the continuity of g when $\frac{1}{16} < w < \frac{1}{2}$.

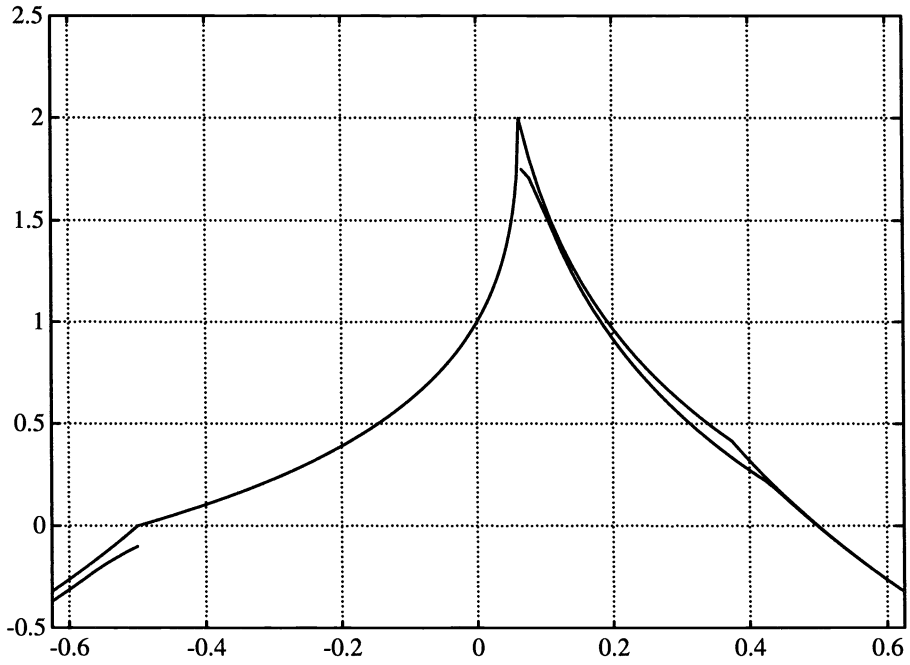


FIG. 9.2. *The critical exponent* $s_0(0)$ *for the interpolating family* $c = \frac{1}{2}(-w, 0, \frac{1}{2}+w, 1, \frac{1}{2}+w, 0, -w)$ *is given by the graph when* $-\frac{1}{2} \le w \le \frac{1}{16}$. *For other values of* w, *the two graphs are plots of the upper and lower bounds given by* (9.9).

A consideration of the limit case $w = \frac{1}{2}$ leads to the idea of writing

$$(9.7)\qquad \begin{cases} S_0^2\varepsilon_0 &= 8w\,(2w(\varepsilon_0+\varepsilon_3)+(1-2w)(\varepsilon_1+\varepsilon_2))\,,\\ S_1^2\varepsilon_0 &= 2S_1\varepsilon_0+16w\,((2w-1)\varepsilon_0-2w\varepsilon_2)\,,\\ S_0S_1\varepsilon_0 &= 2S_0\varepsilon_1+16w\,(-2w\varepsilon_0+(2w-1)\varepsilon_2)\,,\\ S_1S_0\varepsilon_0 &= 2S_0\varepsilon_0-16w\varepsilon_1. \end{cases}$$

For $0 \le w \le \frac{1}{2}$ Proposition 4.5 then gives $\rho_1(S) \le u = \max\{4\sqrt{w}, 1+\sqrt{1+16w}\} = 1+\sqrt{1+16w}$. Hence the lower bound

$$(9.8)\qquad s_0(0) \ge 2 - \log_2(1+\sqrt{1+16w}) \quad \text{for} \quad 0 \le w \le \tfrac{1}{2}.$$

Clearly, this proves that $s_0(0) > 0$ for $0 \le w < \frac{1}{2}$ and the above claim about the continuity of g follows.

A plot of numerically computed upper and lower bounds for $s_0(0)$ is given in Fig. 9.2. In the interval $-\frac{1}{2} \le w \le \frac{1}{16}$ the single graph is given by (9.6). For other values of w the pair of graphs are given by the bounds

$$(9.9)\qquad 2 - \log_2 u_{10} \le s_0(0) \le 2 - \log_2 l_2,$$

using the notation of Proposition 4.2. Numerical experiments suggest that it should be possible to show that $s_0(0) = 2 - \log_2 l_2$, at least in a neighborhood of $w = \frac{3}{16}$, with a generalization of Proposition 4.4 to fourth order recursions. Therefore we conjecture

that $g \in C^1(\mathbb{R})$ if and only if $0 < w < w_0$ where w_0 is defined by $l_2(w_0) = 2$ and $w_0 \approx 0.19272925$. □

Appendix.

Proof of Proposition 2.4. Assume (1) holds and m satisfies Cohen's criterion. If K is the compact of Definition 2.3, the intersection of K and $\xi + 2\pi\mathbb{Z}$ must be empty. As K is compact and $\xi + 2\pi\mathbb{Z}$ is closed, it follows that all points in a whole neighborhood of $\xi \in [-\pi, \pi]$ can not be congruent to any $\eta \in K$, contradicting Definition 2.3(1). We conclude that (1) implies the failure of Cohen's criterion.

Conversely, assume (1) does not hold. For all $\xi \in [-\pi, \pi]$ we then have a $k_\xi \in \mathbb{Z}$ such that $\hat{g}(\xi + 2\pi k_\xi) \neq 0$. By continuity, $\hat{g}$ does not vanish on $I_\xi + 2\pi k_\xi$, where I_ξ is an open interval containing ξ. By compactness, the interval $[-\pi, \pi]$ is covered by a finite subcollection of $\{I_\xi \mid \xi \in [-\pi, \pi]\}$. A modification of the corresponding finite collection of intervals $I_\xi + 2\pi k_\xi$ by set subtractions and closure operations will now give a compact K as in Definition 2.3. In other words, Cohen's criterion can only fail if (1) holds.

The equivalence (1) $\Leftrightarrow$ (2) is a consequence of the Poisson summation formula, in the form that the Fourier transform of the tempered distribution $\sum_k \delta_k$ is $2\pi \sum_k \delta_{2\pi k}$. Indeed, a convolution with the compactly supported distribution $e^{i\xi \cdot} g$ corresponds on the Fourier side to a multiplication with the smooth function of polynomial growth $\tau_{-\xi}\hat{g}$. Hence,

$$\left(\sum_k e^{i\xi k} \tau_k g\right)\widehat{\ } = 2\pi \sum_k \hat{g}(\xi + 2\pi k)\delta_{\xi + 2\pi k},$$

and the the equivalence (1) $\Leftrightarrow$ (2) follows.

To show (1) $\Rightarrow$ (3), the main tool will be that

$$\hat{g}(2\xi + 2\pi k) = \begin{cases} m(\xi)\hat{g}(\xi + \pi k), & k \text{ even}, \\ m(\xi + \pi)\hat{g}(\xi + \pi k), & k \text{ odd}. \end{cases} \tag{A.1}$$

If $m(\xi) = m(\xi + \pi) = 0$ for some ξ we clearly see that $\hat{g}(2\xi + 2\pi k) = 0$ for all $k \in \mathbb{Z}$. Let us therefore assume that $m(\xi)$ and $m(\xi + \pi)$ are never both zero.

Suppose (1) holds and define $\mathcal{N} = \{e^{i\xi} \mid \text{for all } k \in \mathbb{Z} : \hat{g}(\xi + 2\pi k) = 0\}$. By assumption, this set is not empty and it does not contain 1 because $\hat{g}(0) = 1$. If $e^{2i\xi} \in \mathcal{N}$, (A.1) shows that either $e^{i\xi}$ or $e^{i(\xi+\pi)}$ is also in $\mathcal{N}$. That is, $\mathcal{N} \subset \mathcal{N}^2$. We also know that $\mathcal{N}$ must be finite since $\hat{g}$ is holomorphic. Since $\mathcal{N}^2$ cannot have more elements than $\mathcal{N}$, it follows that $\mathcal{N}^2 = \mathcal{N}$. Now the set $\sqrt{\mathcal{N}} = \sqrt{\mathcal{N}^2} = \mathcal{N} \cup (-\mathcal{N})$ must have twice as many elements as $\mathcal{N}$, so $\mathcal{N} \cap (-\mathcal{N}) = \emptyset$. If $e^{i\xi} \in -\mathcal{N}$, then $e^{2i\xi} \in \mathcal{N}$ but $e^{i\xi} \notin \mathcal{N}$ and the first case of (A.1) shows that $m(\xi) = 0$. But this is exactly the last statement of (3).

Next, we show that (3) $\Rightarrow$ (4). Observe that if the factorization $Q(\xi)m(\xi) = Q(2\xi)m_0(\xi)$ holds, the degree of m_0 must be less than that of m, provided Q is non-trivial. By induction, it is therefore not necessary to show that m_0 satisfies Cohen's criterion.

Assume (3) holds. If $m(\xi_0) = m(\xi_0 + \pi) = 0$, we simply use

$$Q(\xi) = \frac{e^{-i\xi} - e^{-2i\xi_0}}{1 - e^{-2i\xi_0}}, \qquad m_0(\xi) = \frac{e^{-i\xi} - e^{-2i\xi_0}}{e^{-2i\xi} - e^{-2i\xi_0}} m(\xi).$$

If $m(\xi)$ and $m(\xi + \pi)$ never both vanish, put $P(\xi) = \prod_{z \in \mathcal{N}} (e^{-i\xi} - z)$. By the property that $\mathcal{N}^2 = \mathcal{N}$, we have $P(2\xi) = P(\xi)P(\xi + \pi)$. Since m vanishes when $P(\xi + \pi) = 0$

we can now use

$$Q(\xi) = \frac{P(\xi)}{P(0)}, \qquad m_0(\xi) = \frac{m(\xi)}{P(\xi+\pi)}.$$

Clearly, $Q(\xi)m(\xi) = Q(2\xi)m_0(\xi)$ and $\hat{g}_0(2\xi) = m_0(\xi)\hat{g}_0(\xi)$ imply that $(Q\hat{g}_0)(2\xi) = m(\xi)(Q\hat{g}_0)(\xi)$ so $\hat{g} = Q\hat{g}_0$ by Theorem 1.1. But this is the Fourier transform of the last identity of (4).

Finally the implication (4) $\Rightarrow$ (1) is evident, since $\hat{g}(\xi) = Q(\xi)\hat{g}_0(\xi) = 0$ whenever $Q(\xi) = 0$. □

Proof of Lemma 8.3. Since ℓ^p-norms decrease with p it suffices to show the lemma for $p_1 = 1$. Also, we only prove the result for β^n, the argument for α_0^n being the same.

Let $\nu \in \mathcal{S}(\mathbb{R})$ have a compactly supported Fourier transform with $\hat{\nu} = 1$ on the support of $\hat{\varphi}$. Then $\hat{\nu}\hat{\varphi} = \hat{\varphi}$ so

$$\|\beta^f\|_1 = \|\beta^{f*\nu}\|_1 \le C\,\|\nu * f\|_{L^1},$$

where $C = \sup_x \sum_k |\varphi(x-k)| < \infty$. With this sampling result in hand, we just have to show that $\|\nu * P_m^n u\|_{L^1} \le C\,2^{-n}$ for some $C > 0$.

By definition $\hat{u}$ is a smooth function with $\hat{u}(0) = 0$. Writing $\hat{u}(\xi) = \xi F(\xi)$ with F smooth, and putting $w_n = \nu * P_m^n u$ we obtain

$$\hat{w}_n(\xi) = 2^{-n}\xi F(2^{-n}\xi)\hat{\nu}(\xi)\prod_{j=1}^{n} m(2^{-j}\xi).$$

The product $\prod_{j=1}^n m(2^{-j}\xi)$ converges uniformly on every compact subset of $\mathbb{C}$ to the entire function $\hat{g}$ as $n \to \infty$. In particular, both the product and its derivative are uniformly bounded on the support of $\hat{\nu}$. By the Leibniz rule, we find the same result for $2^n\hat{w}_n$. Hence $\|\hat{w}_n\|_{H^1} \le C\,2^{-n}$ for some $C > 0$, and the lemma follows from the well known inequality $\|f\|_{L^1} \le \|\hat{f}\|_{H^1}$. □

Acknowledgments. This work was done during a stay at the Department of Mathematics of the Royal Institute of Technology in Stockholm. The author wishes to express his thanks for their warm hospitality. A grant from the Nordic Council of Ministers made the stay possible and is also gratefully acknowledged.

REFERENCES

[BL] J. Bergh and J. Löfström, *Interpolation Spaces, An Introduction*, Springer-Verlag, New York, 1976.

[BW] M.A. Berger and Y. Wang, *Bounded semi-groups of matrices*, preprint, Georgia Inst. Tech., Atlanta, GA 30332.

[C1] A. Cohen, *Ondelettes, analyses multirésolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré, Analyse non linéaire, 7 (1990), pp. 439–459.

[C2] ———, *Ondelettes, analyses multirésolutions et traitement numérique du signal*, thèse, Université Paris IX Dauphine, 1990.

[C3] ———, *Biorthogonal wavelets*, in Wavelets and Their Application, C. K. Chui, ed. (1992), Academic Press, pp. 123–152.

[CC] A. Cohen and J. P. Conze, *Régularité des bases d'ondelettes et mesures ergodiques*, Rev. Mat. Iberoamericana, 8, pages 351–365 (1992).

[CD] A. Cohen and I. Daubechies, *Non-separable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–137.

[CDM] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991), No. 453.

[CH1] D. COLELLA AND C. HEIL, *The characterization of continuous four-coefficient scaling functions and wavelets*, IEEE Trans. Info. Theory, 38 (1992), pp. 876–880.

[CH2] ———, *Characterizations of scaling functions* I. *Continuous solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 496–518.

[D] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[DD1] G. DESLAURIERS AND S. DUBUC, *Dyadic interpolation*, in Fractals: Non-integral Dimensions and Applications, G. Cherbit, ed., John Wiley, Chichester, 1991.

[DD2] ———, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.

[DDD] G. DESLAURIERS, J. DUBOIS, AND S. DUBUC, *Multidimensional iterative interpolation*, Canad. J. Math., 43 (1991), pp. 297–312.

[DGL] N. DYN, J. A. GREGORY, AND D. LEVIN., *Analysis of uniform binary subdivision schemes for curve design*, Constr. Approx., 7 (1991), pp. 127–147.

[DJ] A. DELIU AND B. JAWERTH, *Geometrical dimension versus smoothness*, Constr. Approx, 8 (1992), pp. 211–222.

[DL1] I. DAUBECHIES AND J. LAGARIAS, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[DL2] ———, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[DLG] N. DYN, D. LEVIN, AND J. A. GREGORY, *A 4-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Design, 4 (1987), pp. 257–268.

[E] T. EIROLA, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[G] G. GRIPENBERG, *Unconditional bases of wavelets for Sobolev spaces*, preprint, Dept. of Math., University of Helsinki, 1991.

[H] L. HERVÉ, *Construction et regularité des fonctions d'échelle*, preprint Institut de Recherche Mathématique de Rennes, Université de Rennes 1, Rennes, France, 1992.

[L] P. G. LEMARIÉ, *Fonctions a support compact dans les analyses multi-résolutions*, Rev. Mat. Iberoamericana, 7 (1991), pp. 157–182.

[M] Y. MEYER, *Ondelettes et opérateurs* I: *ondelettes*, Hermann, 1990.

[Ma] S. MALLAT, *Multiresolution approximation and wavelets*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–88.

[MP] C. A. MICCHELLI AND H. PRAUTZSCH, *Uniform refinement of curves*, Linear Algebra Appl., 114-115 (1989), pp. 841–870.

[P] J. PEETRE, *New thoughts on Besov spaces*, Math. Dept., Duke Univ., Durham, NC, 1976.

[R] O. RIOUL, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.

[dR] G. DE RHAM, *Sur une courbe plane*, J. Math. Pures Appl., 35 (1956), pp. 25–42.

[V] L. F. VILLEMOES, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

[VH] M. VETTERLI AND C. HERLEY, *Wavelets and Filter Banks: Theory and Design*, IEEE Trans. on Signal Proc., 40(9) (1992), pp. 2207–2232.

[W] Y. WANG, *On two-scale dilation equations*, preprint, Georgia Inst. Tech., Atlanta, GA, 1991.

SIAM J. MATH. ANAL.
Vol. 25, No. 5, p. 1461, September 1994

012

ERRATUM: GENERALIZED JACOBI WEIGHTS, CHRISTOFFEL FUNCTIONS, AND JACOBI POLYNOMIALS*

TÁMAS ERDÉLYI†, ALPHONSE P. MAGNUS ‡, AND PAUL NEVAI§

The authors were inadvertently listed out of alphabetical order in the previous publication of this article [1].

REFERENCES

[1] P. NEVAI, T. ERDÉLYI, AND A.P. MAGNUS, *Generalized Jacobi weights, Christoffel functions, and Jacobi polynomials*, SIAM J. Math. Anal., 25 (1994), pp. 602–614.

* Received by the editors April 13, 1994; accepted for publication April 14, 1994.

† Present address, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada. Department of Mathematics, Ohio State University, 231 West 18th Avenue, Columbus, Ohio 43210-1174 (terdelyi@math.ohio-state.edu).

‡ Institut de Mathématique Pure et Appliquée, Université Catholique de Louvain, Chemin du Cyclotron, 2, Louvain-la-Neuve, B-1348, Belgium (magnus@anma.ucl.ac.be).

§ Department of Mathematics, Ohio State University, 231 West 18th Avenue, Columbus, Ohio 43210-1174 (nevai@math.ohio-state.edu).

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1463–1467, November 1994

A REMARK ON THE STABILITY OF VISCOUS SHOCK WAVES*

JONATHAN GOODMAN†, ANDERS SZEPESSY‡, AND KEVIN ZUMBRUN§

Abstract. Recently, Szepessy and Xin gave a new proof of stability of viscous shock waves. A curious aspect of their argument is a possible disturbance of zero mass, but $\log(t)t^{-1/2}$ amplitude in the vicinity of the shock wave. This would represent a previously unobserved phenomenon. However, only an upper bound is established in their proof. Here, we present an example of a system for which this phenomenon can be verified by explicit calculation. The disturbance near the shock is shown to be precisely of order $t^{-1/2}$ in amplitude.

Key words. stability, viscous conservation laws, shock waves, diffusion waves

AMS subject classifications. 35K55, 35L65, 76L05

In [4], Szepessy and Xin study the stability of a weak travelling shock wave $\Phi(x - st)$ of a strictly hyperbolic system of conservation laws

$$u_t + f(u)_x = u_{xx}, \qquad t > 0,$$

$$u(\cdot, 0) = u_0. \tag{1}$$

The perturbed solution $u(x,t) \in R^m$ is decomposed into a translated shock wave Φ, a sum of scalar diffusion waves θ_i, and a linear coupled diffusion wave η. Following Liu [2], the initial excess mass $\int_{-\infty}^{+\infty}(u_0 - \Phi)dx$ determines the translation of the shock and the masses of the $m - 1$ scalar diffusion waves. The scalar diffusion waves are given by self-similar solutions of convected Burgers and/or heat equations with the appropriate mass. The coupled linear diffusion wave η, introduced in [4], has zero total mass and is the solution of (1) linearized around the shock wave Φ, with coupling from the scalar diffusion waves as source terms. Theorem 2.2 in [4] shows that the coupled wave η decays time asymptotically as $\log(t)t^{-1/2}$ in the shock region.

The purpose of this paper is to give an example for which the decay estimates in [4] of η in the shock region are sharp modulo a logarithmic factor, $\log(t)$. We demonstrate this by explicit analysis of the example. The significance of this result is to verify that the coupled diffusion wave is a physical phenomenon on the order of the scalar diffusion waves, and not only a technical device for proving stability.

For the choice

$$f(u) = \begin{pmatrix} u_1^2/2 - u_2^2/2 \\ u_2 \end{pmatrix}, \tag{2}$$

(1) has a stationary shock

$$\Phi = \begin{pmatrix} \phi_1 \\ 0 \end{pmatrix} = \begin{pmatrix} -\epsilon \tanh(\epsilon x/2) \\ 0 \end{pmatrix}, \tag{3}$$

* Received by the editors November 2, 1992; accepted for publication (in revised form) July 15, 1993. This work was supported in part by National Science Foundation Postdoctoral Program grant DMS-9107990 and Air Force Office of Scientific Research grant AFOSR-91-0063.

† Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012 (goodman@goodman.cims.nyu.edu).

‡ NADA, Royal Institute of Technology, S-100 44 Stockholm, Sweden (szepessy@nada.kth.se).

§ Department of Mathematics, Indiana University, Bloomington, Indiana 47405 (kzumbrun@indiana.edu).

$$\lim_{x\to\pm\infty}\Phi(x)=\begin{pmatrix}\mp\epsilon\\0\end{pmatrix}.$$

For a constant m, $|m|<\epsilon$, we will consider a perturbation of form

$$u_0=\Phi+\begin{pmatrix}v_x\\m\delta(x)\end{pmatrix}, \tag{4}$$

where $\delta(x)$ represents a dirac mass at the origin, and $v\in H^2(R)$, the Sobolev space of functions with two derivatives in L^2, is assumed to be small.

With these choices, the decomposition of the solution described in [4] takes the simple form

$$\begin{pmatrix}u_1\\u_2\end{pmatrix}=\Phi+\theta+\eta+\begin{pmatrix}v_x\\0\end{pmatrix},$$

$$\theta=\begin{pmatrix}0\\\theta_2\end{pmatrix},\qquad \eta=\begin{pmatrix}\eta_1\\0\end{pmatrix}. \tag{5}$$

Here, the scalar diffusion wave θ_2 satisfies a convected heat equation with a dirac mass as initial data; hence it is simply a convected heat kernel moving with speed $+1$,

$$\theta_2(x,t)=mK(x-t,t),$$

$$K(y,s)\equiv(1/\sqrt{4\pi s})\exp(-y^2/4s), \tag{6}$$

while the coupled diffusion wave η_1 satisfies

$$L\eta_1\equiv\eta_{1t}+(\phi_1(x)\eta_1)_x-\eta_{1_{xx}}=(\theta_2^2/2)_x, \tag{7}$$

$$\eta_1(x,t=0)\equiv 0.$$

It is shown in [4] that

$$\lim_{t\to\infty}||u-\Phi-\theta-\eta||_{L^p(R)}=0,\qquad p\geq 2, \tag{8}$$

$$\lim_{t\to\infty}\int_t^\infty\int_{-\infty}^{+\infty}|u-\Phi-\theta-\eta|^2dx\,dt=0. \tag{9}$$

In the following theorem, we prove that neither θ nor η is negligible in the sense of (9).

THEOREM 1. *Let f and u_0 be given by* (2), (3), *and* (4). *Then there are positive constants c and C independent of ϵ, m, and t, such that if u is a solution of* (1) *and θ and η are given by* (5), (6), *and* (7), *then*

$$Cm^2\epsilon^{1/2}t^{-1/2}>|\eta(x,t)|>cm^2\epsilon^{1/2}t^{-1/2}\quad \text{for } |x|<\epsilon^{-1}\quad\text{and}\quad t>\epsilon^{-2}, \tag{10}$$

and, for ϵ sufficiently small,

$$\lim_{\tau\to\infty}\int_\tau^\infty\int_{-\infty}^{+\infty}|u-\Phi-\theta-\eta|^2dx\,dt=0, \tag{11}$$

$$\lim_{\tau\to\infty}\int_\tau^\infty\int_{-\infty}^{+\infty}|u-\Phi|^2dx\,dt=\infty, \tag{12}$$

$$\lim_{\tau\to\infty}\int_{\tau}^{\infty}\int_{-\infty}^{+\infty}|u-\Phi-\theta|^2dx\,dt=\infty. \tag{13}$$

Remark 1. The estimate of [4] is in this case

$$\eta_1(x,t)\le\begin{cases}c'\log(t)/t^{1/2}, & \text{if } |\epsilon x|<\log(t);\\ c'\log(t)/t, & \text{if } \log(t)/\epsilon<x<t^N+t^{1/2}\log(t);\\ c'\log(t)/[(1+t^N)(1+x^2)], & \text{otherwise.}\end{cases}$$

We note that the limit (9) from [4] is a direct consequence of the energy method used there and the Ansatz $u=\Phi+\theta+\eta+error$. The limits (12) and (13) show that the simpler Ansätze $u=\Phi+error$ or $u=\Phi+\theta+error$ are not compatible with the energy method of [4]. □

The proof of the theorem is based on an exact solution found in [5].

Proof. Following [4] we solve (7) using the fact that the fundamental solution of L satisfies an equation involving L^*, the dual of L. Indeed, let Ψ satisfy

$$L^*\Psi\equiv-\Psi_t-\phi_1\Psi_x-\Psi_{xx}=0,\qquad t<T,$$
$$\Psi(y,t=T)\equiv\delta(y-x).$$

Then,

$$0=\int_0^T\int_{-\infty}^{+\infty}L^*\Psi\eta_1dxdt=\int_0^T\int_{-\infty}^{+\infty}\Psi L\eta_1dxdt-\int_{-\infty}^{+\infty}\eta_1(x,T)\Psi(x,T)dx+\int_{-\infty}^{+\infty}\eta_1(x,0)\Psi(x,0)dx.$$

Hence,

$$\eta_1(x,T)=\int_0^T\int_{-\infty}^{+\infty}\Psi(y,t)\partial_y\theta^2(y,t)/2dydt=\int_0^T\int_{-\infty}^{+\infty}-\Psi_x\theta^2/2dydt, \tag{14}$$

where $z\equiv\Psi_x$ satisfies

$$-z_t-(\phi_1(x)z)_x-z_{xx}=0, \tag{15}$$
$$z(y,t=T)=\partial_y\delta(y-x).$$

In [4], $z=\Psi_x$ is solved approximately, giving the bounds in Remark 1. In the present case, ϕ_1 is a solution of the Burgers equation and exact solutions can be found; cf. [5]. Here, we derive these solutions by differentiating the Hopf–Cole transformation.

The dual equation of (15),

$$w_t+\phi_1w_x-w_{xx}=0, \tag{16}$$

is the linearization of

$$v_t+v_x^2/2-v_{xx}=a \tag{17}$$

around $v=\int_0^x\phi_1(y,t)dy$ for $a=\epsilon^2/2$. The Hopf–Cole transformation

$$v=-2\log(H) \tag{18}$$

reduces (17) to

$$MH=H_t-H_{xx}+aH/2=0.$$

Hence, by differentiating (18) at H, we see that

$$w = H^{-1}h = \exp(v/2)h$$

is a solution of (16) for any h satisfying $Mh = 0$. Setting $w(x, s = t) \equiv \delta(x - y)$ and using the fact that

$$v = \int_0^x \phi_1(y,t)dy = -2\log(\cosh(\epsilon x/2)),$$

we obtain

$$\begin{aligned} w(x,s) &= \exp\left(\int_y^x \phi_1(y')dy'\right)(1/\sqrt{4\pi(s-t)})\exp(-(x-y)^2/4(s-t) - \epsilon^2(s-t)/4) \\ &= \frac{e^{\epsilon y/2} + e^{-\epsilon y/2}}{e^{\epsilon x/2} + e^{-\epsilon x/2}}(1/\sqrt{4\pi(s-t)})\exp(-(x-y)^2/4(s-t) - \epsilon^2(s-t)/4) \\ &= \left(\frac{e^{-\epsilon x/2}}{e^{\epsilon x/2} + e^{-\epsilon x/2}}\right)(1/\sqrt{4\pi(s-t)})\exp(-(x-y-\epsilon(s-t))^2/4(s-t)) \\ &\quad + \left(\frac{e^{\epsilon x/2}}{e^{\epsilon x/2} + e^{-\epsilon x/2}}\right)(1/\sqrt{4\pi(s-t)})\exp(-(x-y+\epsilon(s-t))^2/4(s-t)) \end{aligned}$$

for $t < s$. The purpose of using the dual equation is so that certain inner products are independent of time. For example,

$$\int_R \Psi_x(x,T)w(x,T)dx = \int_R \Psi_x(x,t)w(x,t)dx$$

when $t < T$. From $\Psi(\cdot, T) = \delta_x$ and $w(\cdot, t) = \delta_y$, we obtain

(19)

$$\begin{aligned} &\Psi_x(y,t) = -w_x(x,T) \\ &= \left[\left(\frac{e^{-\epsilon x/2}}{e^{\epsilon x/2} + e^{-\epsilon x/2}}\right)K_y^+(x-y,T-t) + \left(\frac{\epsilon}{(e^{\epsilon x/2} + e^{-\epsilon x/2})^2}\right)K^+(x-y,T-t)\right] \\ &\quad + \left[\left(\frac{e^{\epsilon x/2}}{e^{\epsilon x/2} + e^{-\epsilon x/2}}\right)K_y^-(x-y,T-t) - \left(\frac{\epsilon}{(e^{\epsilon x/2} + e^{-\epsilon x/2})^2}\right)K^-(x-y,T-t)\right], \end{aligned}$$

where

$$K^{\pm}(z,s) = K(z \mp \epsilon s, s) = (1/\sqrt{4\pi s})\exp(-(z \mp \epsilon s)^2/4s)$$

are heat kernels moving with speed $\pm\epsilon$. This means that although Ψ_x initially is a derivative of a dirac mass, it decreases as $t^{-1/2}$ for $|\epsilon x| < 1$.

Equation (19) is equivalent to the fact that the backward problem for L^* has Green's function

$$\left(\frac{e^{-\epsilon x/2}}{e^{\epsilon x/2} + e^{-\epsilon x/2}}\right)K^+(x-y,T-t) + \left(\frac{e^{\epsilon x/2}}{e^{\epsilon x/2} + e^{-\epsilon x/2}}\right)K^-(x-y,T-t),$$

a convex sum of two heat kernels moving outward with speeds ± 1; cf. [5]. This follows, at least approximately, from very general principles. The convection field $-\phi_1(x)$ insures that, in the backward direction, mass moves out to $\pm\infty$. In the far field, ϕ_1 is essentially constant, so that the behavior is like that of a convected heat equation, and asymptotically the solution separates into two humps.

Combining (4) and (19), we have an expression for η of which each term is the convolution of a heat kernel or derivative of a heat kernel with the square of a heat kernel, moving with (possibly) different speeds.

In particular, for $|x| < \epsilon^{-1}$ and $t > \epsilon^{-2}$, the term

$$I \equiv \frac{\epsilon}{(e^{\epsilon x/2} + e^{-\epsilon x/2})^2} \int_0^T \int_{-\infty}^{+\infty} K^-(x-y, T-t)\theta_2(y,t)^2/2 dy dt$$
$$= \frac{m^2\epsilon}{(e^{\epsilon x/2} + e^{-\epsilon x/2})^2} \int_0^T \int_{-\infty}^{+\infty} K(x-y+\epsilon(T-t), T-t)K^2(y-t,t)/2 dy dt$$

is dominant. The integrand of this term is everywhere positive and, in the region

$$A = \{(y,t) : |(y-t)| < \sqrt{t},\ |(y-x) - \epsilon(T-t)| < \sqrt{(T-t)}\},$$

it is greater than $C'(T-t')^{-1/2}t'^{-1}$, where $t' = \frac{x+\epsilon T}{1+\epsilon}$ and $C' = (1/2)e^{-3/4}(4\pi)^{-3/2}$. Hence, there is a positive constant c for which

$$I > \frac{m^2\epsilon}{(e^{\epsilon x/2} + e^{-\epsilon x/2})^2} \int_A K(x-y+\epsilon(T-t), T-t)K^2(y-t,t)/2 dy dt > c\epsilon m^2/\sqrt{t'}$$
$$> cm^2\sqrt{\epsilon/T}.$$

The other terms of (19) inserted into (14) can easily be estimated (cf. [3] and [1]) and yield exponential decay in T from K^+, K_y^+, and at least T^{-1} decay from K_y^-. This proves the lower bound on η in (10), which together with (9) implies (11)–(13). The upper bound in (10) follows by estimating the term I, also, by the methods of [3].

Remark 2. The choice of a linearly degenerate second field was made only for convenience in exposition. For instance, the same argument shows that the theorem remains valid for the flux

$$f(u) = \begin{pmatrix} u_1^2/2 - u_2^2/2 \\ u_2 + u_2^2/2 \end{pmatrix},$$

which has two genuinely nonlinear fields.

Acknowledgments. The authors are grateful to Stanford University and in particular to Joe Keller and Tai-Ping Liu for their hospitality while this work was being done.

REFERENCES

[1] I.-L. CHERN, *Multiple-Mode Diffusion Waves for Viscous Nonstrictly Hyperbolic Conservation Laws*, Tech. Rep., Department of Mathematics, The University of Chicago, 1992.

[2] T.-P. LIU, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., 56 (1985), pp. 1–108.

[3] ———, *Interactions of nonlinear hyperbolic waves*, in 1989 Conference on Nonlinear Analysis, Academica Sinica, Taipei, World Scientific, Singapore, 1989.

[4] A. SZEPESSY AND Z. XIN, *Nonlinear stability of viscous shock waves*, Arch. Rational Mech. Anal., 122 (1993), pp. 55–103.

[5] K. ZUMBRUN, *Formation of diffusion waves in a scalar conservation law with convection*, in Proc. AMS, to appear.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1468–1487, November 1994

002

INTERFACE PROBLEMS IN VISCOPLASTICITY AND PLASTICITY*

CARSTEN CARSTENSEN†

Abstract. This paper is concerned with three-dimensional interface (or transmission) problems in solid mechanics which consist of the (quasi-static) equilibrium condition and a first order evolution inclusion in a bounded Lipschitz domain Ω and the homogeneous linear elasticity problem in an unbounded exterior domain Ω_2. The evolution problem in Ω models viscoplasticity and Prandtl–Reuß plasticity with hardening as well as perfect plasticity. The exterior part of the interface problem is rewritten in terms of boundary integral operators using the Poincaré–Steklov operator. This symmetric coupling approach takes the total system of the Calderon projector into account. Then, existence and uniqueness results are obtained in a mixed variational formulation.

Key words. interface problem, transmission problem, solid mechanics, viscoplasticity, plasticity, perfect plasticity, first order evolution inclusion, Poincaré–Stecklov operator, Calderon projector

AMS subject classifications. 73E60, 73E50, 73C99

1. Introduction.

In this paper we analyze the following interface problem in three-dimensional solid mechanics. Let Ω_2 be an exterior unbounded domain with Lipschitz boundary Γ where the displacements satisfy the homogeneous Navier–Lamé equations in three-dimensional elasticity and the radiation condition $u_2(x) = O(1/|x|)$ for $|x| \to \infty$. In the bounded interior Lipschitz domain $\Omega \subseteq \mathbb{R}^3 \setminus \bar{\Omega}_2$ we consider the quasi-static equilibrium equation with a time-dependent body force and a first order evolution inclusion modeling the material behavior. The latter is related to viscoplasticity and plasticity with hardening as well as perfect plasticity (von Mises yield condition, Prandtl–Reuß flow rule without or with kinematic or isotropic hardening). Finally, on the interface $\Gamma = \bar{\Omega} \cap \bar{\Omega}_2$ we have equality of the traces of the related displacements $u|_\Gamma = u_2|_\Gamma$ and tractions $t = T_2 u_2$ (T_2 is the conormal derivative).

The fundamental solution for the Lamé operator yields a representation formula for u_2 so that u_2 is known as far as its Cauchy data $(u_2|_\Gamma, T_2 u_2|_\Gamma)$ are known. The Cauchy data are coupled by the Poincaré–Steklov operator (or Dirichlet–Neumann map) S_2 through $T_2 u_2 = -S_2 u_2|_\Gamma$. Hence, the interface problem can be reformulated to some problem (P): Given some body force f, find—under suitable initial conditions—time dependent vector fields on Ω, namely the displacements $u \in H$, the stresses $\sigma \in L$ and (possibly) the internal variables $q \in L'$, satisfying almost everywhere in time

$$\epsilon^* \sigma + S u = f, \tag{1}$$

$$\begin{pmatrix} \epsilon u' - A\sigma' \\ -q' \end{pmatrix} \in \partial\varphi \begin{pmatrix} \sigma \\ q \end{pmatrix}. \tag{2}$$

Here, the prime denotes the time derivative, A and S are linear, bounded, symmetric, and time-independent operators; A is positive definite and models elasticity constants and S is positive semidefinite and related to the Poincaré–Steklov operator S_2. The linear operator ϵ maps the displacement field u to the related (linear) Green strains ϵu, the symmetric part of grad u. The dissipation functional φ is a lower semicontinuous convex uniform proper functional, $\partial\varphi$ is its subgradient. This paper covers three particular cases in the class of materials of the Maxwell type defined by (1) and (2):

* Received by the editors August 31, 1992; accepted for publication July 7, 1993.
† Department of Mathematics, Heriot–Watt University, Edinburgh EH14 4AS, United Kingdom.

1. φ is given by

$$\varphi\binom{\sigma}{q} := \frac{1}{2\mu}\operatorname{dist}\left(\binom{\sigma}{q}, B\right)^2$$

for *Perzyna's viscoplasticity* (cf. [20] and the references given there) and

$$\varphi\binom{\sigma}{q} := 0 \quad \text{if } \binom{\sigma}{q} \in B, \qquad \varphi\binom{\sigma}{q} := \infty \quad \text{if } \binom{\sigma}{q} \notin B$$

for the following.

2. *Plasticity with hardening* (q describes hardening).

3. For *perfect plasticity* (where q disappears). In all three cases B is the closed convex set of admissible stress parameters $\binom{\sigma}{q}$ (cf. Definition 7.1).

Note that—in contrast to the elasto-viscoplastic material of the Gröger type [26] or Burger type [23] treated in [4] for interface problems—the quasi-static equilibrium condition (1) and the evolution inclusion (2) (based on the normality rule for the dissipation functional) are not coupled such that no simple substitution gives a single first order evolution inclusion. Therefore, in order to prove existence of solutions, a laborious regularization technique [11], [16], [17], [21], [22]—outlined in the sequel—must be applied: We start considering a regularization of problem (P) and prove that the new problem (P_ν) has a unique solution $(u_\nu, \sigma_\nu, q_\nu)$. Then we prove some estimates of these solutions which are independent of ν. Thus, for a sequence $(\nu_n) \to 0$ we have sequences of solutions of the problems (P_{ν_n}) which are bounded in a reflexive Banach space $W^{1,2}(0,T; H \times L \times L')$. According to Banach–Alaoglu's theorem we select a weakly convergent subsequence. Finally, the weak limit solves problem (P). In the case of plasticity a second regularization is required leading to problems $(P_{\mu,\nu})$.

The crucial point here lies in the behavior of the displacements: Its traces (used in the interface conditions) appear in (1) while the related strain rate $\epsilon u'$ appears in (2). For example, this leads to uniqueness results for the traces of the displacements whereas the displacements are in general not unique in the case of perfect plasticity.

This note presents existence and uniqueness of solutions of the problem (P)—and of the interface problem as well—and considers a modified weak form of problem (P) for perfect plasticity. The question of asymptotic stability of the solutions is left open (i.e., $0 < T < \infty$ will be fixed in the sequel). The paper is organized as follows. The exterior problem as well as some boundary integral operators are introduced in §2 while the general form of the interior problem is described in §3. The interface problem and the above mentioned problem (P) are formulated in §4 where we prove equivalence of the two problems. In §5 we start with proving some general tools and then consider existence and uniqueness for the problems with viscoplastic material in §6 and plasticity with hardening in §7. Perfect plasticity is included in §8 giving existence of solutions of the interface problem in a very weak sense.

In a further paper the numerical treatment of the problem (P) will be analyzed consisting in coupling boundary element and finite element methods in space and, e.g., the implicit Euler method in time.

2. The exterior problem. In this section we report the exterior problem (EP) which is part of the interface problem.

Let $\Omega_0 \subset \Omega_1 \subset \mathbb{R}^3$ be bounded Lipschitz domains in three dimensions such that Ω_0 lies compactly in Ω_1. Then, $\Omega := \Omega_1 \setminus \bar{\Omega}_0$ is the interior domain and $\Omega_2 := \mathbb{R}^3 \setminus \bar{\Omega}_1$ is the exterior domain. The boundary of Ω is divided into two parts, namely the interior boundary $\Gamma_0 := \partial\Omega_0$ and the interface $\Gamma := \partial\Omega_1$; cf. Fig. 1. We consider

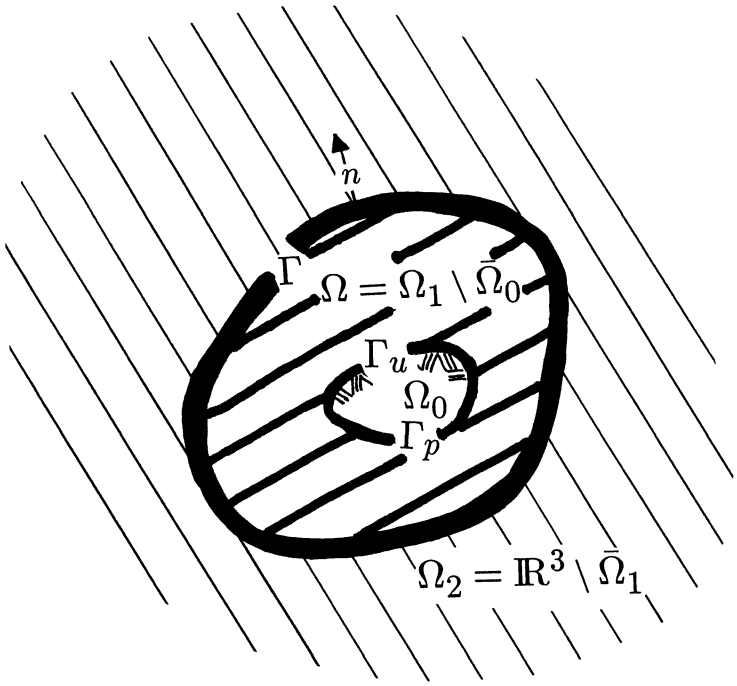


FIG. 1.

Dirichlet, Neumann, or mixed boundary conditions on Γ_0 and allow the case $\Gamma_0 = \emptyset$ (whence $\Omega_0 = \emptyset$).

The exterior problem is the homogeneous Lamé system of linear elasticity [9], [10] for regular displacements [8]–[10], [12], [19],

$$\Delta^* u := -\mu_2 \Delta u - (\lambda_2 + \mu_2) \text{ grad div } u = 0 \quad \text{in } \Omega_2 \tag{3}$$

$$u_2 = O\left(\frac{1}{|x|}\right) \quad \text{as} \quad |x| \to \infty, \tag{4}$$

with $\Delta = \text{div grad}$ denoting the Laplace operator, μ_2, λ_2 being the positive Lamé constants.

Let $T_2(u_2)$ be the conormal derivative defined by

$$T_2(u_2) := 2\mu_2 \partial_n u_2 + \lambda_2 n \operatorname{div} u_2 + \mu_2 n \times \operatorname{curl} u_2.$$

∂_n denotes the normal derivative, n being the unit normal pointing into Ω_2; cf. Fig. 1. Then, the interface conditions at the interface Γ read

$$(\gamma u, t) = (u_2|_\Gamma, T_2(u_2)|_\Gamma), \tag{5}$$

where u are the displacements in Ω, γ denotes the trace mapping, $\gamma := \cdot|_\Gamma$, and t are the (unknown) surface tractions at Γ acting like a surface force for the interior problem.

In order to give a weak formulation of the exterior problem we introduce the space of regular solutions [9], [10]

$$\mathcal{L}_2 := \{u_2 \in H^1_{loc}(\Omega_2; \mathbb{R}^3) : u_2 \text{ satisfies (4) and } \Delta^* u_2 = 0\}. \tag{6}$$

Then the trace-mapping places $u_2 \in \mathcal{L}_2$ onto $\gamma u_2 \in H^{1/2} := H^{1/2}(\Gamma; \mathbb{R}^3)$. The tractions $T_2 u_2$ can be defined by the first Green formula [8], [9], [10]

$$\int_{\Omega_2} \Delta^* u_2 v \, d\Omega_2 = \langle T_2 u_2, \gamma v \rangle + \Phi_2(u_2, v) \tag{7}$$

for any $v \in H^1(\Omega_2; \mathbb{R}^3)$ with compact support and

$$\Phi_2(u_2, v) = \int_{\Omega_2} \sum_{ijkl=1}^{3} a_{ijkl} \epsilon_{kl}(u_2) \epsilon_{ij}(v) \, d\Omega_2,$$

where

$$a_{ijkl} := \lambda_2 \delta_{ij} \delta_{kl} + \mu_2 (\delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk}),$$

$\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. In (7), the strain tensor ϵu is defined by

$$\epsilon_{ij}(u) := \tfrac{1}{2}(u_{i,j} + u_{j,i}), \tag{8}$$

$(u_{i,j}) := (u_{i,j})_{i,j=1,2,3} := \operatorname{grad} u$. The brackets $\langle \cdot, \cdot \rangle$ always denote the duality between $H^{1/2} := H^{1/2}(\Gamma; \mathbb{R}^3)$ and $H^{-1/2} := H^{-1/2}(\Gamma; \mathbb{R}^3) = (H^{1/2})^*$ such that for $v \in H^{1/2}(\Gamma; \mathbb{R}^3)$ and $w \in L^2(\Gamma; \mathbb{R}^3)$

$$\langle w, v \rangle = \int_\Gamma w \cdot v \, d\Gamma.$$

Note that by (7), $T_2 u_2 \in H^{-1/2}$.

DEFINITION (EP). *Given $(v, t) \in H^{1/2} \times H^{-1/2}$ the exterior problem* (EP) *consists in finding $u_2 \in \mathcal{L}_2$ with $(v, t) = (\gamma u_2, T_2 u_2)$.*

In order to solve the exterior problem we need some boundary integral operators concerning the fundamental solution G_2 for the Lamé operator. The Kelvin matrix $G_2(x, y)$ is the kernel of G_2,

$$G_2(x, y) = \frac{\lambda_2 + 3\mu_2}{8\pi\mu_2(\lambda_2 + 2\mu_2)} \left\{ \frac{1}{|x - y|} I + \frac{\lambda_2 + \mu_2}{\lambda_2 + 3\mu_2} \frac{(x - y)(x - y)^T}{|x - y|^3} \right\}.$$

I is the unit matrix in $\mathbb{R}^3$ and ^{T} denotes the transposed matrix. Since G is analytic in $\mathbb{R}^3 \times \mathbb{R}^3$ without the diagonal we may define its traction

$$T_2(x, y) := T_{2,y}(G_2(x, y))^T, \quad x \neq y.$$

Then, the single layer potential V_2, the double layer potential Λ_2 and its dual Λ_2', and the hypersingular operator D_2 are defined by

$$\begin{aligned}
(V_2 \phi)(x) &= \langle G_2(x, \cdot), \phi \rangle, \\
(\Lambda_2 v)(x) &= \langle T_2(x, \cdot), v \rangle, \\
(D_2 v)(x) &= -T_{2,x}(\langle T_2(x, \cdot), v \rangle), \\
(\Lambda_2' \phi)(x) &= -T_{2,x}(\langle G_2(x, \cdot), \phi \rangle, v) \quad (x \in \Gamma).
\end{aligned}$$

From [5]–[8] and [10] we have the following properties of the above boundary potentials. For real Banach spaces X and Y let $\mathcal{L}(X; Y)$ denote the real vector space of bounded linear mappings from X into Y.

LEMMA 2.1. *For* $H^{1/2} := H^{1/2}(\Gamma, \mathbb{R}^3)$ *and* $H^{-1/2} := H^{-1/2}(\Gamma, \mathbb{R}^3)$,

$$V_2 \in \mathcal{L}(H^{-1/2}; H^{1/2}),$$
$$\Lambda_2 \in \mathcal{L}(H^{1/2}; H^{1/2}),$$
$$\Lambda_2' \in \mathcal{L}(H^{-1/2}; H^{-1/2}),$$
$$D_2 \in \mathcal{L}(H^{1/2}; H^{-1/2}).$$

D_2 *is positive semidefinite and* V_2 *is positive definite, i.e., there exists a constant* $c > 0$ *such that for all* $v \in H^{1/2}$ *and all* $\phi \in H^{-1/2}$ *there holds*

$$\langle D_2 v, v\rangle \geq 0 \quad \text{and} \quad \langle \phi, V_2\phi\rangle \geq c\|\phi\|^2_{H^{-1/2}}.$$

D_2 *and* V_2 *are symmetric,* Λ' *is the dual of* Λ. □

We are now in the position to state the following equivalence result concerning the exterior problem. The proof can be found in [2] and [4].

THEOREM 2.2. *For any* $(v,t) \in H^{1/2} \times H^{-1/2} := H^{1/2}(\Gamma, \mathbb{R}^3) \times H^{-1/2}(\Gamma, \mathbb{R}^3)$ *the exterior problem has a solution* u_2 *if and only if*

$$t = -S_2 v, \tag{9}$$

with the

$$S_2 := D_2 + (1/2 - \Lambda_2')V_2^{-1}(1/2 - \Lambda_2) \in \mathcal{L}(H^{1/2}; H^{-1/2}). \tag{10}$$

In this case the solution u_2 *of the exterior problem is unique and given by the representation formula*

$$u_2(x) = \langle T_2(x,\cdot), v\rangle - \langle G_2(x,\cdot), \phi\rangle \quad (x \in \Omega_2), \tag{11}$$

where $\phi := V_2^{-1}(\Lambda_2 - 1/2)v$. □

The proof of the following lemma can be found in [2]–[4].

LEMMA 2.3. *The Poincaré–Steklov operator* S_2 *is positive definite.* □

3. The interior problem. In this section we report the interior problem in plasticity which consists of the equilibrium equation and the evolution inclusion.

In order to state the equilibrium condition, we define

$$H := \{u \in H^1(\Omega; \mathbb{R}) : u|_{\Gamma_u} = 0\};$$

H is the space of the displacements, H^* being the dual of H, and

$$L := L^2(\Omega; \mathbb{R}^{3\times 3}_{sym}), \quad L' := L^2(\Omega; \mathbb{R}^m)$$

are the spaces of stresses and hardening parameters, $\mathbb{R}^{3\times 3}_{sym}$ being the six-dimensional real vector space of symmetric 3×3 matrices, m being a natural number. We identify the duals of L, L', $L \times L'$ with themselves.

Note that the trace map satisfies $\gamma \in \mathcal{L}(H; H^{1/2})$. Let $\gamma^* \in \mathcal{L}(H^{-1/2}; H^*)$ denote that dual of γ. Given $u \in H$, define the strain ϵu through (8) such that $\epsilon \in \mathcal{L}(H; L)$, $\epsilon^* \in \mathcal{L}(L; H^*)$ being the dual of ϵ ($L = L^*$).

Given a body force $b \in H^*$, the strong form of equilibrium

$$\mathrm{div}\sigma + b = 0, \quad \sigma n|_{\Gamma_p} = \bar{t}, \quad \sigma n|_{\Gamma} = t$$

can be rewritten as follows. Multiply $\mathrm{div}\sigma + b = 0$ with a test function $v \in H$, integrate $v \cdot \mathrm{div}\sigma$ over Ω, use Green's formula taking all the boundary conditions as well as the symmetry of σ into account and finally obtain

$$\epsilon^*\sigma = f + \gamma^* t, \tag{12}$$

where $f \in H^*$ is given through b and $\bar{t}$ assuming sufficient regularity of the given tractions $\bar{t} : \Gamma_p \to \mathbb{R}^3$ such that f is bounded.

The evolution inclusion causes time dependence of the problem. Given a real separable Hilbert space X, a fixed real number $T > 0$ and $1 \leq p \leq \infty$, let $L^p(0,T;X)$ denote the space of all measurable functions $h : [0,T] \to X$ such that $\| h \|_X \in L^p(0,T;\mathbb{R})$. Then $W^{m,p}(0,T;X)$ denotes the space of all $h \in L^p(0,T;X)$ such that the derivatives $h, h', \ldots, h^{(m-1)}$ are absolutely continuous (such that they are differentiable almost everywhere in $[0,T]$ in the classical sense as a limit of quotients of differences and the derivative satisfies the main theorem of calculus) and $h^{(m)} \in L^p(0,T;X)$. The norm of h in $W^{m,p}(0,T;X)$ is the sum of $\| h^{(j)} \|_{L^p(0,T;X)}$ over $j = 0, \ldots, m$.

In order to describe the material law let $\varphi : L \times L' \to [0, +\infty]$ be a convex lower semicontinuous proper functional; particular cases are under consideration below. Then the constitutive relation—modeling the normality rule of the dissipation functional—reads

$$(\epsilon u', 0) \in (A\sigma', q') + \partial\varphi(\sigma, q) \quad \text{a.e. in } (0,T), \tag{13}$$

where $A \in \mathcal{L}(L;L)$ is the inverse of a linear elasticity operator, i.e., A is positive definite and symmetric. The subgradient of φ is defined by

$$\begin{aligned} \partial\varphi(\sigma,q) := \{(\tau,p) \in \mathrm{Dom}(\varphi) : \forall (\rho,r) \in L \times L' \\ \langle \binom{\tau}{p}, \binom{\rho-\sigma}{r-q} \rangle_{L\times L'} \leq \varphi(\rho,r) - \varphi(\sigma,q)\}, \\ \mathrm{Dom}(\varphi) := \{(\tau,p) \in L \times L' | \varphi(\tau,p) < \infty\}. \end{aligned}$$

DEFINITION (IP). *Given $f \in W^{2,2}(0,T;H^*)$ and $t \in W^{2,2}(0,T;H^{-1/2})$ with $f(0) = 0$, the interior problem* (IP) *consists in finding $(u,\sigma,q) \in W^{1,2}(0,T;H \times L \times L')$ with $(u,\sigma,q)(0) = 0$ satisfying* (12) *and* (13).

Remark 1. Since $(u,\sigma,q) \in W^{1,2}(0,T;H \times L \times L')$ is absolutely continuous it is continuous so that the initial condition $(u,\sigma,q)(0) = 0$ makes sense. The restriction to homogeneous initial data is only for convenience of notation.

Remark 2. It is referred to the literature for the physical background [21]–[23] and particular cases.

Remark 3. Assuming Dirichlet boundary conditions (i.e., Γ_u has positive surface measure, cf. Fig. 1) as in [10], we get existence and uniqueness of a solution of problem (IP) for the three particular examples considered in this paper following the proofs below.

4. The interface problem. In this section we formulate the interface problem in plasticity and rewrite it in terms of boundary integral operators using Theorem 2.2 in the problem (P) which is analyzed in the following sections.

In (IP) the tractions t are given data. In the interface problem it is unknown.

DEFINITION (Interface Problem). *Given $f \in W^{2,2}(0,T;H^*)$ with $f(0) = 0$, the interface problem consists in finding*

$$(u, u_2, t, \sigma, q) \in W^{1,2}(0,T; H \times \mathcal{L}_2 \times H^{-1/2} \times L \times L')$$

with $(u, u_2, t, \sigma, q)(0) = 0$ *satisfying* (5), (12), *and* (13).

Eliminating t with (9) one obtains the following equivalent problem (P).

DEFINITION (P). *Given* $f \in W^{2,2}(0,T;H^*)$ *with* $f(0) = 0$, *the problem* (P) *consists in finding*

$$(u, \sigma, q) \in W^{1,2}(0,T; H \times L \times L')$$

with $(u, \sigma, q)(0) = 0$ *satisfying* (13) *and*

$$\epsilon^* \sigma + Su = f, \tag{14}$$

where $S := \gamma^* S_2 \gamma \in \mathcal{L}(H; H^*)$.

THEOREM 4.1. $(u, u_2, \sigma, q, t) \in W_2^1(0,T; H \times \mathcal{L}_2 \times L \times L' \times H^{-1/2})$ *solves the interface problem if and only if* $(u, \sigma, q) \in W_2^1(0,T; H \times L \times L')$ *solves problem* (P). *In the latter case* $t = T_2 u_2$ *and* u_2 *is given through* (11).

Proof. The proof is straightforward using Theorem 2.2. The details are omitted. □

Remark 4. Using the general Korn's inequality and Lemma 2.3 one concludes that

$$\epsilon^* \epsilon + S \in \mathcal{L}(H; H^*)$$

is symmetric and positive definite. See, e.g., [3] for details. Thus, given $\nu > 0$, the operator $(\nu \epsilon^* \epsilon + S)^{-1}$ exists and is bounded.

5. Preliminary results. In this section, problem (P) is analyzed using ideas of [16], [17], [21], and [22]. The results are used in the following three sections (concerning viscoplasticity, plasticity with hardening and perfect plasticity) for the proof that problem (P) and hence the interface problem has a unique solution. As it is outlined in §1 we consider some auxiliary problems and study weak limits as well as uniform bounds of their solutions.

PROBLEM (P_ν). *Given* $\nu > 0$, *find* $(u_\nu, \sigma_\nu, q_\nu) \in W^{1,2}(0,T; H \times L \times L')$ *with homogeneous initial values satisfying almost everywhere in* $(0,T)$

$$\epsilon^* \sigma_\nu + S u_\nu + \nu \cdot \epsilon^* \epsilon u_\nu = f \tag{15}$$

$$(\epsilon u_\nu', 0) \in (A\sigma_\nu', q_\nu') + \partial\varphi(\sigma_\nu, q_\nu). \tag{16}$$

Remark 5. In (15), the additional operator $\nu \cdot \epsilon^* \epsilon$ may be replaced by any other positive semidefinite operator A_ν provided that $S + A_\nu$ is positive definite and A_ν tends towards zero if $\nu \to 0^+$.

LEMMA 5.1. *Let* $\nu > 0$. *Then there exists exactly one solution* $(u_\nu, \sigma_\nu, q_\nu)$ *of problem* (P_ν).

Proof. According to Remark 4, (15) is equivalent to

$$u_\nu = (S + \nu \cdot \epsilon^* \epsilon)^{-1} (f - \epsilon^* \sigma_\nu) \in W^{1,2}(0,T; H) \tag{17}$$

which can be substituted in (16), which is then equivalent to

$$(g, 0) \in (\hat{A}\sigma_\nu', q_\nu') + \partial\varphi(\sigma_\nu, q_\nu) \tag{18}$$

with

$$g := \epsilon (S + \nu \epsilon^* \epsilon)^{-1} f' \in W^{1,2}(0,T; L),$$
$$\hat{A} := A + \epsilon (S + \nu \epsilon^* \epsilon)^{-1} \epsilon^*.$$

According to the main theorem on first order evolution equations (cf., e.g., [25] and in particular [26, p. 357] if $\hat{A}$ is not the identity), (18) has a unique solution (σ_ν, q_ν) which proves the lemma. □

The following lemma states that the limits of weakly convergent sequences of solutions of problem (P_ν) with $\nu \to 0$ are solutions of problem (P).

LEMMA 5.2. *Assume that (ν_n) is a sequence of positive real numbers tending to zero as n tends towards infinity such that*

$$(u_{\nu_n}, \sigma_{\nu_n}, q_{\nu_n}) \rightharpoonup (u, \sigma, q) \quad \text{in} \quad L^2(0,T; H \times L \times L') \tag{19}$$

as well as

$$(u'_{\nu_n}, \sigma'_{\nu_n}, q'_{\nu_n}) \rightharpoonup (u', \sigma', q') \quad \text{in} \quad L^2(0,T; H \times L \times L'). \tag{20}$$

We have that $(u_{\nu_n}, \sigma_{\nu_n}, q_{\nu_n})$ are the solutions of the problems (P_{ν_n}) for the parameter sequence (ν_n).

Then $(u, \sigma, q) \in W^{1,2}(0,T; H \times L \times L')$ solves problem (P).

Proof. Because of (19) and $\lim_{n\to\infty} \nu_n = 0$, we obtain (14).

Taking $(\tau, p) \in L^2(0,T; L \times L')$ in (16) and time integration thereof leads to

$$\int_0^T (\langle \epsilon u'_{\nu_n} - A\sigma'_{\nu_n}, \tau - \sigma_{\nu_n}\rangle_L - \langle q'_{\nu_n}, p - q_{\nu_n}\rangle_{L'})\, ds \le \int_0^T (\varphi(\tau, p) - \varphi(\sigma_{\nu_n}, q_{\nu_n}))\, ds,$$

which can be rewritten using (15) as

$$\begin{aligned}\int_0^T (\langle \epsilon u'_{\nu_n} - A\sigma'_{\nu_n}, \tau\rangle_L - \langle q'_{\nu_n}, p\rangle_{L'} - \langle f, u'_{\nu_n}\rangle_H)\, ds \\ \le -\int_0^T (\langle A\sigma_{\nu_n}, \sigma'_{\nu_n}\rangle_L - \langle (S + \nu_n \epsilon^* \epsilon) u_{\nu_n}, u'_{\nu_n}\rangle_H \\ - \langle q_{\nu_n}, q'_{\nu_n}\rangle_{L'} + \varphi(\tau, \sigma) - \varphi(\sigma_{\nu_n}, q_{\nu_n}))\, ds.\end{aligned}$$

The left-hand side converges for $n \to \infty$. Using that $\langle A\sigma, \sigma\rangle_L + \langle Su, u\rangle_H + \langle q, q\rangle_{L'}$ and φ are weakly lower semicontinuous and (14), one finally gets

$$\int_0^T (\langle \epsilon u' - A\sigma', \tau - \sigma\rangle_L - \langle q', p - q\rangle_{L'})\, dt \le \int_0^T (\varphi(\tau, p) - \varphi(\sigma, q))\, dt.$$

From this one can prove (13). □

As mentioned above, we discuss bounds of the solution of problem (P_ν) uniformly in ν. We fix a function

$$(\chi, s) \in W^{2,2}(0,T; L \times L') \quad \text{with} \quad \epsilon^* \chi = f \in W^{2,2}(0,T; L) \tag{21}$$

and $(\chi, s)(0) = 0$. One way to construct the function $\chi(t)$ is to use the solution

$$v \in H^1(\Omega)^3 \quad \text{with} \quad \epsilon^* \epsilon v = f(t) \quad \text{and} \quad v|_{\partial\Omega} = 0$$

and to set $\chi(t) = \epsilon v$.

LEMMA 5.3. *Let* (χ, s) *satisfy* (21) *and assume that* $\varphi(\chi, s) : [0,T] \to [0,\infty]$ *belongs to* $L^1(0,T)$. *Then,*

$$\| (\sigma_\nu, q_\nu) \|_{L^\infty(0,T;L\times L')} \leq c, \tag{22}$$

where the constant $c > 0$ *depends on* $\|A\|$, $\|A^{-1}\|$, $\|(\chi, s)\|_{L^\infty(0,T;L\times L')}$, $\|(\chi', s')\|_{L^1(0,T;L\times L')}$ *and* $\|\varphi(\chi, s)\|_{L^1(0,T)}$ *but not on* ν.

Proof. Define

$$(\tau_\nu, p_\nu) := (\sigma_\nu - \chi, q_\nu - s) \in W^{1,2}(0,T; L \times L'). \tag{23}$$

As seen in the proof of Lemma 5.1, (σ_ν, q_ν) satisfies (18) which shows almost everywhere in $(0,T)$

$$(-A\chi', 0) \in (\hat{A}\tau_\nu', q_\nu') + \partial\varphi(\sigma_\nu, q_\nu). \tag{24}$$

The definition of the subdifferential leads to

$$\langle \begin{pmatrix} \hat{A}\tau_\nu' \\ p_\nu' \end{pmatrix}, \begin{pmatrix} \tau_\nu \\ p_\nu \end{pmatrix} \rangle_{L\times L'} \leq \varphi(\chi, s) - \varphi(\sigma_\nu, q_\nu) + \langle \begin{pmatrix} -A\chi' \\ s' \end{pmatrix}, \begin{pmatrix} \tau_\nu \\ p_\nu \end{pmatrix} \rangle_{L\times L'}.$$

For $t \in (0,T)$, one concludes

$$\int_0^t \langle (\hat{A}\tau_\nu', p_\nu'), (\tau_\nu, p_\nu) \rangle_{L\times L'} \, dt \leq c_1 + c_2 \| (\tau_\nu, p_\nu) \|_{L^\infty(0,T;L\times L')}. \tag{25}$$

The fundamental theorem of calculus, the symmetry of $\hat{A}$ and the homogeneous initial conditions in (25) lead to

$$\frac{1}{2} \langle (\hat{A}\tau_\nu(t), p_\nu(t)), (\tau_\nu(t), p_\nu(t)) \rangle_{L\times L'} \leq c_1 + c_2 \| (\tau_\nu, p_\nu) \|_{L^\infty(0,T;L\times L')}.$$

Since $A \leq \hat{A}$ is positive definite, one obtains

$$\| (\tau_\nu, p_\nu) \|_{L^\infty(0,T;L\times L')} \leq c_3. \qquad \square$$

LEMMA 5.4. *Problem* (P) *determines the "stress parameters" and the traces of the displacements uniquely, i.e., if* (u_i, σ_i, q_i), $i = 1, 2$, *solve problem* (P) *then* $\sigma_1 = \sigma_2$, $q_1 = q_2$ *and* $\gamma u_1 = \gamma u_2$.

Proof. Assume that there exist two solutions (u_i, σ_i, q_i) for $i = 1$ and $i = 2$ of the problem (P). Let $u := u_2 - u_1$, $\sigma := \sigma_2 - \sigma_1$, $q := q_2 - q_1$. Then,

$$\epsilon^* \sigma = -Su \tag{26}$$

and from the definition of the subdifferential for $i = 1, 2$

$$\langle (\epsilon u_i' - A\sigma_i', -q_i'), (\sigma_{i+1} - \sigma_i, q_{i+1} - q_i) \rangle_{L\times L'} \leq \varphi(\sigma_{i+1}, q_{i+1}) - \varphi(\sigma_i, q_i), \tag{27}$$

where the index i is used modulo 2, i.e., $(u_3, \sigma_3, q_3) := (u_1, \sigma_1, q_1)$. Adding the equations in (27) for $i = 1$ and $i = 2$ one obtains

$$\langle (-\epsilon u' + A\sigma', q'), (\sigma, q) \rangle_{L\times L'} \leq 0.$$

By (26), this gives almost everywhere in (0,T)

$$\langle Su', u \rangle_H + \langle A\sigma', \sigma \rangle_L + \langle q', q \rangle_{L'} \leq 0. \tag{28}$$

Due to the fundamental theorem of calculus, the homogeneous initial values and the symmetry of S and A, integration of (28) shows almost everywhere in (0,T)

$$\langle Su, u \rangle_H + \langle A\sigma, \sigma \rangle_L + \langle q, q \rangle_{L'} \leq 0. \tag{29}$$

Since S_2 and A are positive definite, (29) proves $\sigma = 0$, $q = 0$ and $\gamma u = 0$. $\square$

6. Viscoplasticity. The first example for the functional φ describes the viscoplastic law due to Perzyna; we refer to [20] (and the references given there) for the physical derivation and for the justification of the hardening parameters in particular.

Given a convex closed subset B of $L \times L'$ with $0 \in B$, let $\operatorname{dist}((\sigma, q), B)$ denote the distance of the stress parameters (σ, q) from B in $L \times L'$. For $\mu > 0$ define

$$\varphi : L \times L' \to \mathbb{R}, \quad (\sigma, q) \mapsto \frac{1}{2\mu} \operatorname{dist}((\sigma, q), B)^2. \tag{30}$$

It is well known that $\varphi \geq 0$ is convex, continuous, and Gateaux-differentiable:

$$\partial\varphi(\sigma, q) = \left\{ \frac{1}{\mu} \Big((\sigma, q) - \Pi_B(\sigma, q) \Big) \right\}, \tag{31}$$

where $\Pi_B : L \times L' \to B$ is the orthogonal projector onto B in $L \times L'$.

THEOREM 6.1. *If φ is given as in* (30), *then problem* (P) *as well as the interface problem have unique solutions.*

Proof. Since φ is continuous, Lemma 5.3 shows (22) with a constant c independent of $\nu > 0$. Hence $\partial\varphi(\sigma_\nu, q_\nu) : [0, T] \to L \times L'$ is also bounded in $L^\infty(0, T; L \times L')$. By (31),

$$\| (\epsilon u'_\nu - A\sigma'_\nu, -q'_\nu) \|_{L^\infty(0,T;L\times L')} \leq 2c/\mu. \tag{32}$$

Using this, Remark 4, and (15) we obtain almost everywhere in (0,T)

$$\begin{aligned} & c_1 \| u'_\nu \|_H^2 - c_2 \| u'_\nu \|_H \\ & \leq \langle \epsilon u'_\nu, A^{-1} \epsilon u'_\nu \rangle_L + \langle S u'_\nu + \nu \epsilon^* \epsilon u'_\nu - f', u'_\nu \rangle_H \\ & = \langle \epsilon u'_\nu - A\sigma'_\nu, A^{-1} \epsilon u'_\nu \rangle_L \leq c_3 \| u'_\nu \|_H. \end{aligned}$$

Thus $\| u'_\nu \|_{L^\infty(0,T;H)} \leq c_4$ where $c_1, \ldots, c_4$ are independent of ν. Hence u_ν and u'_ν are bounded in $L^2(0, T; H)$ uniformly in ν. According to (32) the same holds for the stress parameters. Then the Banach–Alaoglu theorem and Lemma 5.2 prove the existence of solutions of problem (P). According to (31) and (13), Lemma 5.4 shows uniqueness not only for the stress parameters and γu but also for $\epsilon u'$. By the homogeneous initial values we conclude uniqueness for ϵu and u as well. □

Remark 6. The estimates in the proof of the theorem lead to $(u, \sigma, q) \in W^{1,\infty}(0, T; H \times L \times L')$ as well.

7. Plasticity with hardening. The second example for the functional φ describes the Prandtl–Reuß plasticity with the von Mises yield condition. We refer to [16], [17], [20]–[22] for the physical derivation and for the justification of the hardening parameters in particular.

DEFINITION 7.1. *Identify $\mathbb{R}^{3\times3}_{sym} \times \mathbb{R}^m$ with $\mathbb{R}^p$ and write $L \times L' = L^2(\Omega; \mathbb{R}^p)$, $p := 6 + m$. A convex and continuous function $f : \mathbb{R}^p \to \mathbb{R}$ (f should not be confused with the given body force) with $f(0, 0) < 0$ is called a yield function of φ if*

$$\begin{aligned} C &:= \{(\sigma, q) \in \mathbb{R}^{3\times3}_{sym} \times \mathbb{R}^m = \mathbb{R}^p | f(\sigma, q) \leq 0\}, \\ B &:= \{(\sigma, q) \in L \times L' = L^2(\Omega; \mathbb{R}^p) | (\sigma, q) \in C \text{ a.e. in } \Omega\}, \\ \varphi &:= L \times L' \to [0, \infty], (\sigma, q) \mapsto 0 \text{ if } (\sigma, q) \in B; \infty \text{ if } (\sigma, q) \notin B. \end{aligned}$$

(a) *φ models plasticity with kinematic hardening (and the von Mises yield condition) if $m = 6$, $\mathbb{R}^6$ is identified with $\mathbb{R}^{3\times3}_{sym}$ and its yield function f is given through*

$$f(\sigma, q) = (\sigma^D - q^D) : (\sigma^D - q^D) - 1, \quad \sigma, q \in \mathbb{R}^{3\times3}_{sym}.$$

(b) *φ models plasticity with isotropic hardening (and the von Mises yield condition) if $m = 1$ and its yield function f is given through*

$$f(\sigma, q) = \sigma^D : \sigma^D - (1+q)^2, \quad (\sigma, q) \in \mathbb{R}^{3\times 3}_{sym} \times [0, \infty).$$

Recall that σ^D is the deviatoric part of $\sigma \in \mathbb{R}^{3\times 3}_{sym}$, given through $\sigma^D := \sigma - \frac{1}{3} tr(\sigma) I$, I being the unit matrix in $\mathbb{R}^3$. $A : B := \sum_{i,j=1,2,3} A_{ij} B_{ji}$ for $A, B \in \mathbb{R}^{3\times 3}$.

In the next lemma we state an important relation between σ' and $\epsilon u'$ using the following notation. For $\tau, \rho \in \mathbb{R}^{3\times 3}$ let the fourth order tensor $\tau \otimes \rho$ be the dyadic product of τ and ρ defined by

$$(\tau \otimes \rho)_{ijkl} := \tau_{ij} \cdot \rho_{kl}$$

so that $\tau \otimes \rho \alpha = (\rho : \alpha)\tau$ for $\alpha \in \mathbb{R}^{3\times 3}$.

LEMMA 7.2. *Let $(u, \sigma, q) \in W^{1,2}(0, T; H \times L \times L')$ satisfy (13). For any $t \in [0, T]$ define $\Omega_e(t)$ and $\Omega_p(t)$ (up to a set of measure zero) through*

$$\Omega_e(t) := \{x \in \Omega | f(\sigma(t,x), q(t,x)) < 0\},$$
$$\Omega_p(t) := \{x \in \Omega | f(\sigma(t,x), q(t,x)) = 0\}.$$

(a) *If φ models kinematic hardening, then almost everywhere in $[0, T]$*

$$\begin{aligned} \epsilon u' &= A\sigma', \quad q' = 0 \quad a.e.\ in\ \Omega_e, \\ \epsilon u' &= (A + (\sigma^D - q^D) \otimes (\sigma^D - q^D))\sigma' \quad a.e.\ in\ \Omega_p, \\ q' &= (\sigma^D - q^D) \otimes (\sigma^D - q^D) : \sigma' \quad a.e.\ in\ \Omega_p. \end{aligned}$$

(b) *If φ models isotropic hardening, then a.e. in $[0, T]$*

$$\begin{aligned} \epsilon u' &= A\sigma', \quad q' = 0 \quad a.e.\ in\ \Omega_e, \\ \epsilon u' &= \left(A + \frac{\sigma^D \otimes \sigma^D}{\sigma^D : \sigma^D}\right)\sigma' \quad a.e.\ in\ \Omega_p, \\ q' &= \frac{\sigma^D : \sigma'}{|\sigma^D|} \quad a.e.\ in\ \Omega_p. \end{aligned}$$

In both cases $\epsilon u'$ is uniquely defined through (σ, q) and there exists a constant c depending only on $\|A\|$ and $\| (\sigma, q) \|_{L^\infty(0,T;L\times L')}$ such that

$$\| \epsilon u' \|_{L^2(0,T;L)} \le c \cdot \| (\sigma', q') \|_{L^2(0,T;L\times L')}.$$

Proof. The proof follows the lines of [17, Thm. 2] so that the details are omitted. □

Remark 7. According to Lemmas 5.4 and 7.2, problem (P) has at most one solution.

DEFINITION 7.3. *We say that $f \in W^{2,2}(0, T; H^*)$ is a safe load and that $(\chi, s) \in W^{1,\infty}(0, T; L^\infty(\Omega; \mathbb{R}^{3\times 3}_{sym} \times \mathbb{R}^m)$ satisfies the safe load assumption if $\chi(0) = 0$, $\epsilon^* \chi = f$ and there exists some $\delta > 0$ such that for any $t \in [0, T]$ and for any $(\rho, r) \in L \times L'$ the following holds*

$$\| (\rho - \chi(t), r - s(t)) \|_{L^\infty(\Omega; \mathbb{R}^{3\times 3}_{sym} \times \mathbb{R}^m)} \le \delta \quad \Rightarrow \quad (\rho, r) \in B.$$

Following [17] the safe load assumption is just a regularity assumption if hardening occurs.

LEMMA 7.4. *Assume that there exists* $\chi \in W^{1,\infty}(0,T;L)$ *with* $\epsilon^*\chi = f$ *and*

$$\| \chi \|_{L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^{3\times 3}_{sym}))} < \infty.$$

Then, for kinematic or isotropic hardening, there exists $s \in W^{1,\infty}(0,T;L')$ *such that* (χ, s) *satisfies the safe load assumption.*

Proof. Let $s := \chi$ in the case of kinematic and $s := 2\chi : \chi$ in the case of isotropic hardening and let $\delta := 1/2$. Then, the lemma is proved by straightforward calculations. □

Remark 8. In perfect plasticity (cf. §8) the safe load assumption [16] is some kind of bounded data requirement (bounded through the frozen hardening parameter). We refer to [22] for examples (in one space dimension) without a solution if it is violated. The question of a "limit load" is related to the safe load assumption (cf. [22], [24]). Physically, the safe load assumption states that there exists a fictitious stress field which is "uniformly below yielding" (i.e., it leads to a purely elastic material behavior even for perturbed stresses) and in equilibrium with the exterior load.

As considered in (30), let $\varphi_\mu(\sigma,q) := \frac{1}{2\mu}\mathrm{dist}((\sigma,q),B)^2$. φ_μ is Gateaux-differentiable with the derivative $D\varphi_\mu(\sigma,q) = \frac{1}{\mu}((\sigma,q) - \Pi_B(\sigma,q)) \in L \times L'$ where $\Pi_B : L \times L' \to B$ is the orthogonal projector onto B.

LEMMA 7.5. *Let* $(\chi, s) \in W^{1,\infty}(0,T;L\times L')$ *satisfy the safe load assumption with the related constant* δ. *Then for any* $(\sigma,q) \in W^{1,2}(0,T;L\times L')$ *we have for* $p = 1, 2$

$$\int_0^T \| D\varphi_\mu(\sigma(t),q(t)) \|^p_{L^1(\Omega;\mathbb{R}^{3\times 3}_{sym}\times \mathbb{R}^m)}\, dt$$
$$\leq \frac{1}{\delta}\int_0^T < D\varphi_\mu(\sigma(t),q(t)), (\sigma(t)-\chi(t), s(t)-q(t)) >^p_{L\times L'}\, dt.$$

Proof. The proof is explicitly given in [21, Lemma 2] or [22, Lemma 3.1 p. 306] for the case $m = 0$ and works verbatim for the present case. □

After the above preliminaries the main result reads as follows.

THEOREM 7.6. *If* φ *is given as in Definition* 7.1 *and if the safe load assumption holds then problem* (P), *as well as the interface problem, has a unique solution* $(u,\sigma,q) \in W^{1,2}(0,T;H\times L\times L')$.

The proof is divided in several steps formulated as lemmas. The summary of the proof is given at the end of this section.

According to Remark 7 it remains to prove existence of solutions. In order to regularize problem (P) we introduce two parameters $\mu, \nu > 0$ and consider problem $(P_{\mu,\nu})$.

PROBLEM $(P_{\mu,\nu})$. *Given* $\mu,\nu > 0$, *find* $(u_{\mu,\nu},\sigma_{\mu,\nu},q_{\mu,\nu}) \in W^{1,2}(0,T;H\times L\times L')$ *with homogeneous initial values satisfying almost everywhere in* $(0,T)$

$$\epsilon^*\sigma_{\mu,\nu} + Su_{\mu,\nu} + \nu\cdot\epsilon^*\epsilon u_{\mu,\nu} = f, \tag{33}$$

$$(\epsilon u'_{\mu,\nu} - A\sigma'_{\mu,\nu}, -q'_{\mu,\nu}) = D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}). \tag{34}$$

LEMMA 7.7. *Let* $\mu,\nu > 0$. *Then there exists exactly one solution* $(u_{\mu,\nu},\sigma_{\mu,\nu}, q_{\mu,\nu})$ *of problem* $(P_{\mu,\nu})$, *and* $(u_{\mu,\nu},\sigma_{\mu,\nu},q_{\mu,\nu}) \in W^{1,2}(0,T;H\times L\times L')$. *Under the safe load assumption there exists a constant* c *such that for any* $\mu,\nu > 0$

$$\| (\sigma_{\mu,\nu}, q_{\mu,\nu}) \|_{L^\infty(0,T;L\times L')} \leq c.$$

Proof. Existence and uniqueness of solutions of $(u_{\mu,\nu}, \sigma_{\mu,\nu}, q_{\mu,\nu}) \in W^{1,2}(0,T; H \times L \times L')$ of problem $(P_{\mu,\nu})$ follow from Lemma 5.1. Choosing (χ, s) from the safe load assumption in Lemma 5.3 gives the bound of $(\sigma_{\mu,\nu}, q_{\mu,\nu})$ independent of $\mu, \nu > 0$ and concludes the proof. □

The following two lemmas give uniform bounds for the solutions of problem $(P_{\mu,\nu})$.

LEMMA 7.8. *Under the safe load assumption there exists a constant c such that for any $\mu, \nu > 0$ and for the unique solution $(u_{\mu,\nu}, \sigma_{\mu,\nu}, q_{\mu,\nu})$ of problem $(P_{\mu,\nu})$ we have*

$$\| (\gamma u_{\mu,\nu} \sigma_{\mu,\nu}, q_{\mu,\nu}) \|_{W^{1,2}(0,T;H^{1/2} \times L \times L')} \leq c.$$

Proof. Define $(\tau_{\mu,\nu}, p_{\mu,\nu}) := (\sigma_{\mu,\nu} - \chi, q_{\mu,\nu} - s)$ where (χ, s) satisfies the safe load assumption. Then (33) gives

$$u_{\mu,\nu} = -(S + \nu \cdot \epsilon^* \epsilon)^{-1} \epsilon^* \tau_{\mu,\nu}. \tag{35}$$

This can be differentiated with respect to time and substituted in (34) which gives

$$0 = (\hat{A} \tau'_{\mu,\nu}, p'_{\mu,\nu}) + (A\chi', s') + D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}) \tag{36}$$

with $\hat{A} := A + \epsilon(S + \nu \cdot \epsilon^* \epsilon)^{-1} \epsilon^*$.

In the first part of the proof multiply (36) with $(\tau_{\mu,\nu}, p_{\mu,\nu})$, integrate over $[0,T]$, use the fundamental theorem on calculus, notice the homogeneous initial conditions and use that A is positive definite to obtain

$$\begin{aligned}
&\int_0^T \langle D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}), (\tau_{\mu,\nu}, p_{\mu,\nu}) \rangle_{L \times L'} \, dt \\
&\leq -\frac{1}{2} \langle (\hat{A} \tau_{\mu,\nu}(T), p_{\mu,\nu}(T)), (\tau_{\mu,\nu}(T), p_{\mu,\nu}(T)) \rangle_{L \times L'} \\
&\quad + \| (A\chi', s') \|_{L^2(0,T;L \times L')} \| (\tau_{\mu,\nu}, p_{\mu,\nu}) \|_{L^2(0,T;L \times L')} \\
&\leq \| (A\chi', s') \|_{L^2(0,T;L \times L')} \| (\tau_{\mu,\nu}, p_{\mu,\nu}) \|_{L^2(0,T;L \times L')} \leq c_1,
\end{aligned}$$

where we used Lemma 7.7. Then Lemma 7.5 shows

$$\int_0^T \| D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}) \|_{L^1(\Omega; \mathbb{R}^{3 \times 3}_{sym} \times \mathbb{R}^m)} \, dt \leq c_1/\delta. \tag{37}$$

In the second part of the proof multiply (36) with $(\tau'_{\mu,\nu}, p'_{\mu,\nu})$ and integrate over $[0,T]$ to obtain

$$\begin{aligned}
&\int_0^T (\langle \hat{A} \tau'_{\mu,\nu}, \tau'_{\mu,\nu} \rangle_L + \langle p'_{\mu,\nu}, p'_{\mu,\nu} \rangle_{L'}) \, dt \\
&\leq \int_0^T \langle D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}), (\chi'_{\mu,\nu}, s'_{\mu,\nu}) \rangle_{L \times L'} \, dt \\
&\quad - \int_0^T \langle D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}), (\sigma'_{\mu,\nu}, q'_{\mu,\nu}) \rangle_{L \times L'} \, dt \\
&\quad + \| (A\chi', s') \|_{L^2(0,T;L \times L')} \| (\tau'_{\mu,\nu}, p'_{\mu,\nu}) \|_{L^2(0,T;L \times L')} \\
&\leq \frac{c_1}{\delta} \| (\chi'_{\mu,\nu}, s'_{\mu,\nu}) \|_{L^\infty(0,T;L^\infty(\Omega; \mathbb{R}^{3 \times 3}_{sym} \times \mathbb{R}^m))} \\
&\quad + c_2 \| (\tau'_{\mu,\nu}, p'_{\mu,\nu}) \|_{L^2(0,T;L \times L')},
\end{aligned}$$

where we used the fundamental theorem on calculus, the homogeneous initial conditions, $0 \in B$, $\varphi_\mu \geq 0$ and (37). Since A is positive definite this shows

$$\| (\tau'_{\mu,\nu}, p'_{\mu,\nu}) \|_{L^2(0,T;L\times L')} \leq c_4.$$

Using (35) in the above estimate of

$$\int_0^T \langle \hat{A}\tau'_{\mu,\nu}, \tau'_{\mu,\nu}\rangle_L \, dt \geq \int_0^T \langle Su'_{\mu,\nu}, u'_{\mu,\nu}\rangle_H \, dt$$

we obtain from Lemma 2.3

$$\| \gamma u'_{\mu,\nu} \|_{L^2(0,T;H^{1/2})} \leq c_5.$$

The constants $c_1, \ldots, c_5$ are independent of μ, ν. Since (χ, s) is fixed and by definition of $(\tau_{\mu,\nu}, p_{\mu,\nu})$, this and Lemma 7.7 prove the lemma. □

In order to treat the case $\mu \to \infty$ consider problem (P_ν) where φ (see (16)) is given as in Definition 7.1.

LEMMA 7.9. *For any $\nu > 0$ there exists exactly one solution $(u_\nu, \sigma_\nu, q_\nu)$ of problem (P_ν). Under the safe load assumption there exists a constant c such that for any $\nu > 0$*

$$\| (\gamma u_\nu, \sigma_\nu, q_\nu) \|_{W^{1,2}(0,T;H^{1/2}\times L\times L')} \leq c.$$

Proof. Existence and uniqueness of solutions $(u_\nu, \sigma_\nu, q_\nu) \in W^{1,2}(0,T; H \times L \times L')$ of problem (P_ν) follow from Lemma 5.1. Note that for fixed $\nu > 0$ the family

$$(u_{\mu,\nu}, \sigma_{\mu,\nu}, q_{\mu,\nu})_{\mu>0}$$

is bounded in $W^{1,2}(0,T; H\times L\times L')$; see Lemma 7.8 and (17). Consequently, according to Banach–Alaoglu's theorem, there exists a sequence (μ_n) of positive real numbers with $\lim_{n\to\infty} \mu_n = 0$ and

$$(u_{\mu_n,\nu}, \sigma_{\mu_n,\nu}, q_{\mu_n,\nu}) \rightharpoonup (\hat{u}_\nu, \hat{\sigma}_\nu, \hat{q}_\nu) \quad \text{in} \quad W^{1,2}(0,T; H \times L \times L'). \tag{38}$$

As in the proof of Lemma 5.2 one concludes that $(\hat{u}_\nu, \hat{\sigma}_\nu, \hat{q}_\nu)$ satisfies the homogeneous initial condition and (15). According to (34), Lemma 7.8 and (35) there exists $c_\nu > 0$ such that for any $\mu > 0$

$$\| D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}) \|_{L^2(0,T;L\times L')} \leq c_\nu.$$

By definition of $D\varphi_\mu$ this implies

$$\lim_{n\to\infty} \| (\sigma_{\mu,\nu}, q_{\mu,\nu}) - \Pi_B(\sigma_{\mu,\nu}, q_{\mu,\nu}) \|_{L^2(0,T;L\times L')} = 0. \tag{39}$$

Consider the continuous function

$$g : L^2(0,T; L\times L') \to \mathbb{R}, \quad (\rho, r) \mapsto \int_0^T \| (\rho, r) - \Pi_B(\rho, r) \|^2_{L\times L'} \, dt.$$

Since g is is weakly lower semicontinuous (39) leads to

$$(\hat{\sigma}_\nu, \hat{q}_\nu) \in B \qquad \text{a.e. in } [0,T]. \tag{40}$$

Finally, let $(\tau, p) \in L^2(0,T; L \times L')$ such that $(\tau, p) \in B$ almost everywhere in $[0,T]$. Since φ is convex, $D\varphi_\mu$ is monotone. Hence, almost everywhere in $[0,T]$

$$0 \leq \langle (\sigma_{\mu,\nu}, q_{\mu,\nu}) - (\tau, p), D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu}) \rangle_{L\times L'}.$$

Substitution of $D\varphi_\mu(\sigma_{\mu,\nu}, q_{\mu,\nu})$ from (36), integration over $[0,T]$ and using $(\tau_{\mu,\nu}, p_{\mu,\nu}) := (\sigma_{\mu,\nu} - \chi, q_{\mu,\nu} - s)$ leads to

$$\begin{aligned} (41) \qquad & \int_0^T \langle (\tau'_{\mu_n,\nu}, p'_{\mu_n,\nu}), (\hat{A}\tau_{\mu_n,\nu}, p_{\mu_n,\nu}) \rangle_{L\times L'} \, dt \\ & \leq -\int_0^T \langle (\sigma_{\mu_n,\nu}, q_{\mu_n,\nu}) - (\tau, p), (A\chi', s') \rangle_{L\times L'} \, dt \\ & \quad -\int_0^T \langle (\chi, s) - (\tau, p), (\hat{A}\tau'_{\mu_n,\nu}, p'_{\mu_n,\nu}) \rangle_{L\times L'} \, dt. \end{aligned}$$

Because of (38), the right-hand side converges. The left-hand side is a weakly lower semicontinuous function in $(\tau_{\mu,\nu}, p_{\mu,\nu})$. Hence, writing $(\hat{\tau}_\nu, \hat{p}_\nu) := (\hat{\sigma} - \chi, \hat{q}_\nu - s)$, one concludes

$$\begin{aligned} & \int_0^T \langle (\hat{\tau}_\nu, \hat{p}_\nu), (\hat{A}\hat{\tau}'_\nu, \hat{p}'_\nu) \rangle_{L\times L'} \, dt \\ & \leq -\int_0^T \langle (\hat{\sigma}_\nu, \hat{q}_\nu) - (\tau, p), (A\chi', s') \rangle_{L\times L'} \, dt \\ & \quad -\int_0^T \langle (\chi, s) - (\tau, p), (\hat{A}\hat{\tau}'_\nu, \hat{p}'_\nu) \rangle_{L\times L'} \, dt. \end{aligned}$$

Since $(\hat{u}_\nu, \hat{\sigma}_\nu, \hat{q}_\nu)$ satisfies (15) this gives

$$(42) \qquad \int_0^T \langle (\tau, p) - (\hat{\sigma}_\nu, \hat{q}_\nu), (\epsilon\hat{u}'_\nu - A\hat{\sigma}'_\nu, -\hat{q}'_\nu) \rangle_{L\times L'} \, dt \leq 0.$$

As in the proof of Lemma 5.2, from this one proves (16).

Altogether, the weak limit $(\hat{u}_\nu, \hat{\sigma}_\nu, \hat{q}_\nu)$ from (38) is the unique solution of problem (P_ν). Hence, the bound of Lemma 7.8 holds also for $(\gamma u_\nu, \sigma_\nu, q_\nu)$. □

Proof of Theorem 7.6. According to Lemma 7.9 the solutions $(u_\nu, \sigma_\nu, q_\nu)$ of (P_ν) exist for any $\nu > 0$ and the traces of the displacements and the stress parameters $(\gamma u_\nu, \sigma_\nu, q_\nu)$ are uniformly bounded in $W^{1,2}(0,T; H^{1/2} \times L \times L')$. Since $(u_\nu, \sigma_\nu, q_\nu)$ satisfies the conditions of Lemma 7.2, $\epsilon u'$ is uniformly bounded in $W^{1,2}(0,T;L)$ as well. Thus, $(u_\nu, \sigma_\nu, q_\nu)_{\nu>0}$ is bounded in $W^{1,2}(0,T; H \times L \times L')$. According to Banach–Alaoglu's theorem there exists a sequence of positive real numbers (ν_n) with $\lim_{n\to\infty} \nu_n$ such that

$$(u_{\nu_n}, \sigma_{\nu_n}, q_{\nu_n})_{n=1,2,3,\ldots} \quad \text{and} \quad (u'_{\nu_n}, \sigma'_{\nu_n}, q'_{\nu_n})_{n=1,2,3,\ldots}$$

converges weakly in $L^2(0,T; H \times L \times L')$. Then, using Lemma 5.2 one concludes existence of solutions of problem (P). □

Remark 9. The proof of Theorem 7.6 shows the role of the hardening parameters. In the absence of hardening it is not guaranteed that $\epsilon u'$ is bounded and that a weakly convergent sequence of solutions of the problems (P_{ν_n}) exists.

Note that (37) gives a bound for $\epsilon u'_{\mu,\nu}$ independent of μ, ν in the space $L^1(0,T; L^1(\Omega; \mathbb{R}^{3\times 3}_{sym}))$. This is the tool for proving existence results in the case of perfect plasticity, i.e., $m = 0$, in the space $BD(\Omega)$; cf. the following section.

8. Perfect plasticity. The third example for the functional φ describes perfect plasticity with the von Mises yield condition. In contrast to the previous cases we cannot prove uniqueness of the displacements (cf. [21] for a one-dimensional example in the interior problem with infinite solutions) which may be discontinuous (in $BD(\Omega)$; cf. below) so that the interface conditions must be weakened.

As in Definition 7.1 we define

$$C := \{\sigma \in \mathbb{R}^{3\times 3}_{sym} | f(\sigma) \leq 0\},$$
$$B := \{\sigma \in L | \sigma \in C \text{ a.e. in } \Omega\},$$
$$\varphi := L \to [0,\infty], \qquad \sigma \mapsto 0 \quad \text{if } \sigma \in B; \infty \text{ if } \sigma \notin B,$$

where $m = 0$, i.e., no hardening occurs. The von Mises yield function f (f should not be confused with the given body force) is defined by $f(\sigma) := \sigma^D : \sigma^D - 1$ but other yield functions are also allowed. Assume that B is closed and convex in L and that the safe load assumption of Definition 7.3 is satisfied (for $m = 0$, i.e., s does not appear). For convenience of notation we assume that $\Omega_0 = \emptyset$ or $\emptyset = \Gamma_p \subseteq \Gamma_0 = \partial\Omega_0$. The more general case of mixed or Neumann data on $\partial\Omega_0$ can easily be included following the lines of [22, Thm. 3.3] but, in view of a possible discontinuity of the displacements on $\partial\Omega$, this requires a further variable for the displacements on $\Gamma_p \subseteq \partial\Omega_0$.

In order to consider a weaker formulation of the interface problem we follow [16] and [22] for the interior problem: Assume that (u, u_2, t, σ) satisfies the interface problem. Then, provided e.g., $f \in L^3(\Omega; \mathbb{R}^3)$, we have everywhere in $[0,T]$

$$\sigma \in \Sigma := \{\tau \in L : \operatorname{div} \tau \in L^3(\Omega; \mathbb{R}^3),\ \tau n|_\Gamma \in H^{-1/2}\},$$

τn being defined by Green's formula. Hence, according to Green's formula, for any $\tau \in \Sigma$

$$\langle \epsilon u', \tau - \sigma\rangle_L = \langle (\tau - \sigma)n|_\Gamma, \gamma u'\rangle - \int_\Omega u' \operatorname{div}(\tau - \sigma)\, d\Omega.$$

Note that the right-hand side makes sense if $\gamma u' \in H^{1/2}$ and $u' \in L^{2/3}(\Omega; \mathbb{R}^3)$ which weakens the assumptions. Hence, using (13) one obtains $\sigma \in \Sigma \cap B$ and for all $\tau \in \Sigma \cap B$

$$0 \leq \langle A\sigma', (\tau - \sigma)\rangle_L - \langle (\tau - \sigma)n|_\Gamma, \gamma u'\rangle + \int_\Omega u' \operatorname{div}(\tau - \sigma)\, d\Omega. \tag{43}$$

Thus, any solution of the interface problem solves the weak interface problem (WIP) which reads as follows.

DEFINITION (Weak Interface Problem). *Given* $f \in W^{2,2}(0,T; L^3(\Omega; \mathbb{R}^3))$ *with* $f(0) = 0$, *the weak interface problem* (WIP) *consists in finding* $(u_2, w, t, \sigma) \in W^{1,2}(0,T; \mathcal{L}_2 \times H^{1/2} \times H^{-1/2} \times \Sigma \cap B)$ *and* $v \in L^2_w(0,T; BD(\Omega))$ *with* $(u_2, w, t, \sigma)(0) = 0$ *satisfying*

$$\operatorname{div} \sigma + f = 0 \quad \text{in } \Omega, \tag{44}$$
$$\sigma n|_\Gamma = t \quad \text{on } \Gamma, \tag{45}$$
$$(w,t) = (u_2|_\Gamma, T_2(u_2)|_\Gamma) \quad \text{on } \Gamma, \tag{46}$$

and for all $\tau \in \Sigma \cap B$

$$0 \leq \langle A\sigma', (\tau - \sigma)\rangle_L - \langle (\tau - \sigma)n|_\Gamma, w'\rangle + \int_\Omega v \operatorname{div}(\tau - \sigma)\, d\Omega. \tag{47}$$

Remark 10.

1. The space of *bounded deformations*

$$BD(\Omega) := \{v \in L^1(\Omega;\mathbb{R}^3) : \epsilon_{ij}(v) \in M^1(\Omega) \quad (1 \le i,j \le 3)\},$$

—$M^1(\Omega)$ being the space of bounded measures on Ω, $M^1(\Omega) = C_c^0(\Omega)^*$, i.e., the dual of the space of continuous functions with compact support in Ω (endowed with the max-norm)—was introduced by Strang and Suquet and is now established in the context of plasticity [24]. $BD(\Omega)$ is a nonreflexive Banach space endowed with the norm

$$\| v \|_{BD(\Omega)} = \| v \|_{L^1(\Omega;\mathbb{R}^3)} + \sum_{i,j=1,2,3} \| \epsilon_{ij}(v) \|_{M^1(\Omega)}.$$

The space $BD(\Omega)$ is motivated by the fact that it is the dual space of

$$C(\bar{\Omega};\mathbb{R}^{3\times 3}_{sym})/\{\tau \in C(\bar{\Omega};\mathbb{R}^{3\times 3}_{sym}) : \mathrm{div}\tau = 0,\ \tau n|_\Gamma = 0\}.$$

Consequently, we may consider the weak* topology on $BD(\Omega)$.

2. The embedding

$$BD(\Omega) \hookrightarrow L^{3/2}(\Omega;\mathbb{R}^3)$$

is continuous. Hence, according to $\tau - \sigma \in \Sigma$, (47) makes sense.

3. Comparing (47) with (43) one identifies u' with v, but an additional variable w appears in the role of γu. The possibility of $v|_\Gamma \neq w$ arises from discontinuities of deformations allowed on Γ [22].

4. Concerning $BD(\Omega)$, Green's formula is known in the following form [21], [24]: Provided $\partial\Omega$ is of class C^1 there exists $\hat{\gamma} \in \mathcal{L}(BD(\Omega);L^1(\partial\Omega;\mathbb{R}^3))$ such that for any $\tau \in C^1(\bar{\Omega};\mathbb{R}^{3\times 3}_{sym})$ and any $v \in BD(\Omega)$

$$\int_\Omega v\mathrm{div}\ \tau\, d\Omega + \int_\Omega \tau\, d\epsilon(v) = \int_{\partial\Omega} \hat{\gamma}(v)\tau n\, d\Gamma$$

where $\int_\Omega \tau\, d\epsilon(v) := \sum_{i,j=1,2,3} < \epsilon_{ij}(v), \tau_{ij} >_{M^1(\Omega)}$.

We have $\hat{\gamma}v = v|_{\partial\Omega}$ if $v \in C(\bar{\Omega};\mathbb{R}^3)$.

5. Obviously, the embeddings

$$H \hookrightarrow BD(\Omega) \quad \text{and} \quad H^{1/2} \hookrightarrow L^1(\partial\Omega;\mathbb{R}^3)$$

are continuous. Then a density argument shows $\gamma = \hat{\gamma}|_H$.

6. Note that $BD(\Omega)$ is the dual of a normed linear space X so that $L^2_w(0,T;BD(\Omega)) = L^2(0,T;X)^*$; L^2_w is the space of weakly measurable vector valued functions such that their norms belong to $L^2(0,T;\mathbb{R})$; cf. [15].

Eliminating t with (9) in (45), (46) one obtains the following problem (WP).

DEFINITION (WP). *Given* $f \in W^{2,2}(0,T;L^3(\Omega;\mathbb{R}^3))$ *with* $f(0) = 0$, *the weak problem* (WP) *consists in finding*

$$(w,\sigma) \in W^{1,2}(0,T;H^{1/2} \times \Sigma \cap B) \quad \textit{and} \quad v \in L^2_w(0,T;BD(\Omega))$$

with $(w,\sigma)(0) = 0$ *satisfying* (44),

$$\sigma n|_\Gamma = -S_2 w \quad \textit{on}\ \Gamma \tag{48}$$

and (47) *for all* $\tau \in \Sigma \cap B$.

With Theorem 2.2 one concludes the following theorem.

THEOREM 8.1. (v, u_2, w, t, σ) *solves the weak interface problem* (WIP) *if and only if* (v, w, σ) *solves the weak problem* (WP). *In the latter case* $t = T_2 w$ *and* u_2 *is given through* (11) *having Cauchy data* (w, t). □

According to the previous theorem the following theorem proves existence of solutions of the weak interface problem.

THEOREM 8.2. *Provided that the safe load assumption holds for* $\chi \in W^{1,2}(0,T;\Sigma)$ *the weak problem* (WP) *has a solution* (v, w, σ) *with* (w, σ) *being unique.*

Proof. Let (P_μ) denote the problem (P) in §6, i.e., problem (P) where φ is given by (30) and $m = 0$. According to Theorem 6.1, let (u_μ, σ_μ) denote the unique solution of problem (P_μ) for any $\mu > 0$. From the proof of Theorem 6.1 we know that (u_μ, σ_μ) is the weak limit of some sequence of solutions $(u_{\mu,\nu}, \sigma_{\mu,\nu})$ of the problem $(P_{\mu,\nu})$ for $\nu \to 0$. Hence, we conclude from Lemma 7.8 that $(\gamma u_\mu, \sigma_\mu)_{\mu>0}$ is uniformly bounded in $W^{1,2}(0,T;H^{1/2} \times L)$.

Writing $\tau_\mu := \sigma_\mu - \chi$ so that $\epsilon^* \tau_\mu + S u_\mu = 0$ (χ from the safe load assumption with $\epsilon^* \chi = f$), (P_μ) yields

$$\begin{aligned}\langle D\varphi_\mu(\sigma_\mu), \tau_\mu \rangle_L &= \langle \epsilon u'_\mu - A\sigma'_\mu, \tau_\mu \rangle_L \\ &= -\langle S_2 \gamma u_\mu, \gamma u'_\mu \rangle - \langle A\sigma'_\mu, \tau_\mu \rangle_L.\end{aligned}$$

Since $(\gamma u_\mu, \sigma_\mu)_{\mu>0}$ is uniformly bounded in $W^{1,2}(0,T;H^{1/2} \times L)$ as well as in $L^\infty(0,T;H^{1/2} \times L)$ this leads to a constant $c_1 > 0$ (independent of $\mu > 0$) with

$$\int_0^T \langle D\varphi_\mu(\sigma_\mu), \tau_\mu \rangle_L^2 \, dt \le c_1. \tag{49}$$

Lemma 7.5 with $p = 2$ shows

$$\| D\varphi_\mu(\sigma_\mu) \|_{L^2(0,T;L^1(\Omega;\mathbb{R}^{3\times 3}_{sym}))} \le c_2.$$

Using this in (P_μ) gives that $\epsilon u'_\mu$ is uniformly bounded in $L^2(0,T;L^1(\Omega;\mathbb{R}^{3\times 3}_{sym}))$ and in $L^2(0,T;BD(\Omega))$.

Now—due to the Banach–Alaoglu theorem—we are in the position to select a subsequence (μ_n) of parameters with $\lim_{n\to\infty} \mu_n = 0$ and

$$\begin{aligned}(\gamma u_{\mu_n}, \sigma_{\mu_n}) &\rightharpoonup (w, \sigma) \quad \text{in } W^{1,2}(0,T;H^{1/2} \times L), \\ (u_{\mu_n}) &\rightharpoonup^* v \quad \text{in } L^2_w(0,T;BD(\Omega)), \\ (u_{\mu_n}) &\rightharpoonup v \quad \text{in } L^2(0,T;L^{3/2}(\Omega;\mathbb{R}^3)),\end{aligned}$$

where $\rightharpoonup$ and $\rightharpoonup^*$ denote weak and weak* convergence, respectively.

As above (before Definition (WIP)) one concludes that

$$\operatorname{div} \sigma_\mu + f = 0, \quad \sigma_\mu n|_\Gamma + S_2 \gamma u_\mu = 0 \tag{50}$$

so that the weak limit (w, σ) satisfies the same relations: $\operatorname{div} \sigma + f = 0$, $\sigma n|_\Gamma + S_2 w = 0$. According to the definition of the subdifferential and because of $\chi \in B$ we have

$$\varphi_\mu(\sigma_\mu) \le \langle D\varphi_\mu(\sigma_\mu), \tau_\mu \rangle_L.$$

Because of (49), this gives $\int_0^T \varphi_\mu(\sigma_\mu)^2 \, dt \le c_1$. One proves $\sigma \in B$ as in the proof of Lemma 7.9; cf. (39).

In order to prove (47) let $\tau \in \Sigma \cap B$. As explained above, (u_μ, σ_μ) satisfies (43). By (50) this gives

$$0 \leq \langle A\sigma'_\mu, (\tau - \sigma_\mu)\rangle_L - \langle \gamma^* \tau n|_\Gamma + Su_\mu, u'_\mu\rangle + \int_\Omega u'_\mu(\operatorname{div} \tau - f)\, d\Omega.$$

By time integration and the weak convergence one proves that (47) holds for the weak limits.

Finally, the claimed uniqueness can be proved following the lines of the proof of Lemma 5.4 starting from (47); we omit the details. □

Remark 11. Note that, following the arguments of this section, similar results can be obtained in the more general case if hardening parameters occur which do not allow estimates of the form

$$\| \epsilon u' \|_{L^2(0,T;L)} \leq c \| (\sigma', q') \|_{L^2(0,T;L\times L')}$$

as in Lemma 7.2.

REFERENCES

[1] J. BIELAK AND R. C. MACCAMY, *An exterior interface problem in two-dimensional elastodynamics*, Quart. Appl. Math., 41 (1983), pp. 143–159.

[2] C. CARSTENSEN, *Nonlinear interface problems in solid mechanics -finite and boundary element coupling*, Habilitation thesis, Universität Hannover, Germany, 1993.

[3] ———, *Interface problem in holonomic elastoplasticity*, Math. Meth. Appl. Sci., 16 (1993), pp. 813–835.

[4] C. CARSTENSEN AND E. P. STEPHAN, *Interface problem in elasto-viscoplasticity*, Quart. Appl. Math. (1995).

[5] M. COSTABEL, *Symmetric methods for the coupling of finite elements and boundary elements*, in C.A. Brebia et al., eds., Boundary Elements IX, Vol. 1, pp. 411–420, Springer-Verlag, Berlin 1987.

[6] M. COSTABEL, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.

[7] M. COSTABEL AND E. P. STEPHAN, *Boundary integral equations for mixed boundary value problems in polygonal domains and Galerkin approximation*, Banach Center Publ., 15 (1985), pp. 175–251.

[8] ———, *A direct boundary integral equation method for transmission problems*, J. Math. Anal. Appl., 106 (1985), pp. 367–413.

[9] ———, *Integral equations for transmission problems in linear elasticity*, J. Integr. Eq. Appl., 2 (1990), pp. 211–223.

[10] ———, *Coupling of finite and boundary element methods for an elastoplastic interface problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1212–1226.

[11] G. DUVAUT AND J. L. LIONS, *Inequalities in Mechanics and Physics*, Springer-Verlag, New York, 1976.

[12] G. N. GATICA AND G.C. HSIAO, *The coupling of boundary element and finite element methods for a nonlinear exterior boundary value problem*, Z. Anal. Anw., 8 (1989), pp. 377–387.

[13] ———, *On the coupled* BEM *and* FEM *for a nonlinear exterior Dirichlet problem in* R^2. Numer. Math., 61 (1992), pp. 171–214.

[14] H. HAN, *A new class of variational formulations for the coupling of finite and boundary element methods*, J. Comput. Math., 8 (1990), pp. 223–232.

[15] A. IONESCU TULCEA AND C. IONESCU TULCEA, *Topics in the Theory of Liftings*, Springer-Verlag, Berlin, 1969.

[16] C. JOHNSON, *Existence theorems for plasticity problems*, J. Math. Pures et Appl., 55 (1976), pp. 431–444.

[17] ———, *On plasticity with hardening*, J. Math. Anal. Appl., 62 (1978), pp. 325–336.

[18] C. JOHNSON AND J. C. NEDELEC, *On the coupling of boundary integral and finite element methods*, Math. Comp., 35 (1980), pp. 1063–1079.

[19] V. D. KUPRADZE, ET AL., *Three-dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity*. North-Holland, Amsterdam, 1979.

[20] J. C. SIMO, *Nonlinear stability of the time-discrete variational problem of evolution in nonlinear heat conduction, plasticity and viscoplasticity*, Comput. Methods Appl. Mech. Engrg., 88 (1991), pp. 111–131.

[21] P.-M. SUQUET, *Sur les équations de la plasticité: existence et régularité des solutions*, J. de Mécanique, 20 (1981), pp. 3–39.

[22] ———, *Discontinuities and plasticity*, in Nonsmooth Mechanics and Applications, CISM Courses and Lectures 302, J. J. Moreau and P.D. Panagiotopoulos, eds., Springer-Verlag, New York, 1988, pp. 279–341.

[23] P. LE TALLEC, *Numerical Analysis of Viscoelastic Problems*, RMA 15, Springer-Verlag, New York, 1990.

[24] R. TEMAM, *Mathematical Problems in Plasticity*, Gauthier-Villars, Paris, 1985.

[25] E. ZEIDLER, *Nonlinear Functional Analysis and its Applications* III, Springer-Verlag, New York 1985.

[26] ———, *Nonlinear Functional Analysis and its Applications* IV, Springer-Verlag, New York 1988.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1488–1503, November 1994

003

FINITE ENERGY SOLUTIONS OF NONLINEAR SCHRÖDINGER EQUATIONS OF DERIVATIVE TYPE*

NAKAO HAYASHI† AND TOHRU OZAWA‡

Abstract. This paper is concerned with the initial value problem for nonlinear Schrödinger equations of the form

$$(\dagger) \qquad \begin{cases} i\partial_t\psi + \partial\psi = i\lambda\partial(|\psi|^2\psi) + \lambda_1|\psi|^{p_1-1}\psi + \lambda_2|\psi|^{p_2-1}\psi, & (t,x) \in \mathbb{R}\times\mathbb{R}, \\ \psi(0,x) = \phi(x), & x \in \mathbb{R}, \end{cases}$$

where $\partial = \partial_x = \partial/\partial x, \lambda, \lambda_1, \lambda_2 \in \mathbb{R}$ and $2 \le p_1 < p_2 < 5$. It is shown that if $\phi \in H^1(\mathbb{R})$ and $\|\phi\|_2^2 < 2\pi/|\lambda|$, then there exists a unique global solution ψ of (†) such that $\psi \in C(\mathbb{R}; H^1(\mathbb{R}))$. This paper introduces a new method to obtain the result.

Key words. derivative nonlinear Schrödinger equations, gauge transformation

AMS subject classifications. 35Q55, 35Q60

1. Introduction. In this paper we study the Cauchy problem for nonlinear Schrödinger equations of the form

$$(1.1) \qquad \begin{cases} i\partial_t\psi + \partial^2\psi = i\lambda\partial(|\psi|^2\psi) + F(\psi), & (t,x) \in \mathbb{R}\times\mathbb{R}, \\ \psi(0,x) = \phi(x), & x \in \mathbb{R}, \end{cases}$$

where ψ is a complex valued function of $(t,x) \in \mathbb{R}\times\mathbb{R}$, $\partial = \partial/\partial x$, $\lambda \in \mathbb{R}$ and $F \in C(\mathbb{R}^2;\mathbb{R}^2)$ satisfies the gauge condition $F(e^{i\theta}\zeta) = e^{i\theta}F(\zeta)$, $\theta \in \mathbb{R}$.

We assume that F can be written as $F = F_1 + F_2$ with $F_j \in C^2(\mathbb{R}^2 \setminus \{0\};\mathbb{R}^2)$, $j = 1,2$, satisfying $F_j'' \in L^\infty_{\mathrm{loc}}(\mathbb{R}^2)$, $F_j(0) = F_j'(0) = 0$ and that there exist positive constants C_j, D_j such that

$$(1.2) \qquad \begin{cases} |F_j(\zeta) - F_j(\zeta')| \leqq C_j(|\zeta|^{p_j-1} + |\zeta'|^{p_j-1})|\zeta - \zeta'|, \\ |F_j'(\zeta) - F_j'(\zeta')| \leqq D_j(|\zeta|^{p_j-2} + |\zeta'|^{p_j-2})|\zeta - \zeta'|, \end{cases}$$

where $2 \leqq p_1 < p_2 < 5$. Furthermore we assume that there exists a function $H \in C^1(\mathbb{R}^2;\mathbb{R})$ satisfying $H(0) = 0$, $F = \partial H/\partial\bar{\zeta}$ and

$$(1.3) \qquad \int_{\mathbb{R}} H(f)dx \geq -M(\|f\|_2) - \mu\|f\|_6^6 \quad \text{for all } f \in H^1(\mathbb{R}),$$

where $\mu \in \mathbb{R}^+$, $M \in C(\mathbb{R}^+;\mathbb{R}^+)$. Under the above conditions we prove the following theorem.

THEOREM 1. *We assume that* (1.2) *and* (1.3) *are satisfied,* $\phi \in H^1(\mathbb{R})$, *and*

$$\|\phi\|_2^2 < \frac{2\pi}{\sqrt{\lambda^2 + 16\mu^2}}.$$

* Received by the editors March 26, 1993; accepted for publication (in revised form) August 10, 1993.

† Department of Mathematics, Faculty of Engineering, Gunma University, Kiryu 376, Japan (nhayashi@eg.gunma-u.ac.jp).

‡ Department of Mathematics, Hokkaido University, Sapporo 060, Japan.

Then there exists a unique global solution ψ *of* (1.1) *such that*

$$\psi \in C(\mathbb{R}; H^1(\mathbb{R})) \cap L^4_{\mathrm{loc}}(\mathbb{R}; W^{1,\infty}(\mathbb{R})).$$

Moreover, the map $\phi \mapsto \psi$ *is continuous from* $\{\phi \in H^1(\mathbb{R}); \|\phi\|_2^2 < 2\pi/\sqrt{\lambda^2 + 16\mu^2}\}$ *with topology induced from* $H^1(\mathbb{R})$ *to* $C(\mathbb{R}; H^1(\mathbb{R})) \cap L^4_{\mathrm{loc}}(\mathbb{R}; W^{1,\infty}(\mathbb{R}))$.

We prove that our result applies to the example

$$F(\psi) = \lambda_1 |\psi|^{p_1 - 1}\psi + \lambda_2 |\psi|^{p_2 - 1}\psi, \tag{1.4}$$

where $2 \le p_1 < p_2 < 5$, $\lambda_1, \lambda_2 \in \mathbb{R}$. First, it is clear that (1.4) satisfies (1.2) (see, e.g., [6]). Second, we have

$$\int_{\mathbb{R}} H(\psi)dx = \frac{2\lambda_1}{p_1 + 1}\|\psi\|_{p_1+1}^{p_1+1} + \frac{2\lambda_2}{p_2 + 1}\|\psi\|_{p_2+1}^{p_2+1}. \tag{1.5}$$

We deduce (1.3) from (1.5). By Hölder's inequality we have for $1 \le p < 5$

$$\|f\|_{p+1}^{p+1} \le \|f\|_6^{\frac{3(p-1)}{2}} \|f\|_2^{\frac{5-p}{2}},$$

from which we see that for any $\varepsilon > 0$, there exists a positive constant C_ε such that

$$\|f\|_{p+1}^{p+1} \le \varepsilon \|f\|_6^6 + C_\varepsilon \|f\|_2^2.$$

Using this and (1.5) we obtain

$$\int_{\mathbb{R}} H(\psi)dx \ge -C_{\varepsilon'}\|f\|_2^2 - \varepsilon' \|f\|_6^6,$$

where

$$\varepsilon' = \sum_{j=1}^{2} \frac{2|\lambda_j|}{p_j + 1}\varepsilon, \quad C_{\varepsilon'} = \sum_{j=1}^{2} \frac{2|\lambda_j|}{p_j + 1} C_\varepsilon,$$

which implies (1.3) with $\mu = \varepsilon'$. Hence we have the following corollary to Theorem 1.

COROLLARY 1. *We let* $F(\psi) = \sum_{j=1}^{2} \lambda_j |\psi|^{p_j - 1}\psi$, *where* $2 \le p_1 < p_2 < 5$. *We assume that* $\phi \in H^1(\mathbb{R})$ *and* $\|\phi\|_2^2 < \frac{2\pi}{|\lambda|}$. *Then the result of Theorem* 1 *is valid.*

In the case of $F(\psi) \equiv 0$, Theorem 1 was proved in [4] (see also [3]) using the result of [9, Thm. 3], which ensures the existence of global smooth solutions of (1.1) with $F(\psi) \equiv 0$. When $F(\psi) \not\equiv 0$, the result [9, Thm. 3] is not applicable, and therefore we need a different method from the previous ones (see [3], [4]). We state our strategy of the proof of Theorem 1 for the convenience of the reader. To obtain the result we first consider the system of nonlinear Schrödinger equations

$$\begin{cases} Lu = -i\lambda u^2 \bar{v} + F(u), \\ Lv = i\lambda v^2 \bar{u} + \partial_u F(u) \cdot v + \partial_{\bar{u}} F(u) \cdot \bar{v}, \\ u(0) = u_0, \ v(0) = v_0, \end{cases} \tag{1.6}$$

with the constraint

$$v_0 = \partial u_0 + i\frac{\lambda}{2}|u_0|^2 u_0. \tag{1.7}$$

The initial value problem of (1.6) without constraint (1.7) fits in the framework of the previous methods as in [7], [8], [10], [11] by making use of the space–time estimates of solutions to the linear Schrödinger equation. We first give the basic results about (1.6) in §2 (see Propositions 2.1–2.3). The proof of the local existence of solutions

to (1.6) requires the differentiability of nonlinear terms with respect to (u, v), which in turn requires the condition $p_1 \geq 2$. Second, we prove the existence of solutions to (1.6) with constraint (1.7) in Propositions 2.4 and 2.5. In Proposition 2.4, we prove the unique existence of $H^2 \times H^2$ solutions to (1.6) and the invariance of the constraint $v = \partial u + i(\lambda/2)|u|^2 u$. The proof of the invariance requires the gauge condition of F. In Proposition 2.5, we prove that the unique existence of solutions to (1.6) still holds for the data with minimal regularity assumption $(u_0, v_0) \in H^1 \times L^2$ under constraint (1.7), which is to be invariant under the time evolution. The proof of Proposition 2.5 requires the smallness condition of u_0 in the L^2-norm.

In Theorem 2 we show that Propositon 2.5 implies the existence of a unique local solution to (1.1) in H^1 through the relation

$$u_0 = \exp\left(-i\lambda \int_{-\infty}^{x} |\phi|^2 dy\right) \phi.$$

Finally, we prove Theorem 1 by a priori estimates of local solutions to (1.1).

We state the results of our related problem. In [5], we studied nonlinear Schrödinger equations of the form

$$\begin{cases} i\partial_t u + \partial^2 u = F(u, \partial u, \bar{u}, \partial \bar{u}), \\ u(0) = u_0, \end{cases} \tag{1.8}$$

where $F : \mathbb{C}^4 \to \mathbb{C}$ is a polynomial having neither constant nor linear terms. We showed the existence of a unique local solution of (1.8) when the data satisfy conditions such as $u_0 \in H^3$ and $xu_0 \in H^2$. The mixed nonlinear Schrödinger equation

$$i\partial_t u + \partial^2 u = i\beta u^2 \partial \bar{u} + i\gamma |u|^2 \partial u + g(|u|^2)u$$

was studied in [1] using energy methods, where $\beta, \gamma \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ is a function satisfying some regularity conditions. The authors established existence and uniqueness of global smooth solutions in H^s for $s \geq 3$ under some smallness condition on the data.

In the case where the underlying space is a bounded domain $\Omega = (0, \ell)$ with $\ell > 0$, the global existence of $H_0^1(\Omega)$ solutions of the equation

$$i\partial_t u + \partial^2 u = i\lambda \partial(|u|^2 u) + \alpha |u|^\rho u, \quad t > 0, \ x \in \Omega,$$

with Dirichlet zero condition was shown in [2] under the smallness condition of the $H_0^1(\Omega)$ norm of the data, where $\lambda, \alpha \in \mathbb{R}$, and $\rho \geq 2$.

Recently in [12], a result similar to ours was obtained. We note here that our method is different from that in [12] and our argument in the proof of Proposition 2.4 is useful to prove the existence of the modified wave operators of the derivative nonlinear Schrödinger equation in [13].

We conclude this section by giving notation. We abbreviate $\partial/\partial u$ to ∂_u. By $\bar{u}$ we denote the complex conjugate of u. We let $L^p = \{f; f \text{ is measurable on } \mathbb{R}, ||f||_p < \infty\}$, where $||f||_p^p = \int_{\mathbb{R}} |f(x)|^p dx$ if $1 \leq p < \infty$ and $||f||_\infty = \text{ess. sup}\{|f(x)|; x \in \mathbb{R}\}$ if $p = \infty$, and we let $W^{m,p} = \{f \in L^p; ||f||_{W^{m,p}} = \sum_{j=0}^m ||\partial^j f||_p\}$. For simplicity we put $W^{m,2} = H^m$. We denote by $(\cdot, \cdot)$ the inner product in L^2. For any interval I of $\mathbb{R}$ and a Banach space B with norm $||\cdot||_B$, we let $C(I; B)$ be the space of continuous functions from I to B and $L^p(I; B)$ be the space consisting of strongly measurable B-valued functions $u(\cdot)$ defined on I such that $\int_I ||u(t)||_B^p dt < \infty$. Different positive constants will be denoted by the same letter C. If necessary, by $C(*, \ldots, *)$ we denote constants depending on the quantities appearing in parentheses.

2. Proof of Theorem 1. We consider the system of nonlinear Schrödinger equations of the form

$$\begin{cases} Lu = -i\lambda u^2\bar{v} + F(u), \\ Lv = i\lambda v^2\bar{u} + \partial_u F(u)\cdot v + \partial_{\bar{u}} F(u)\cdot \bar{v}, \\ u(0) = u_0,\ v(0) = v_0, \end{cases} \tag{2.1}$$

where $L = i\partial_t + \partial_x^2$. In what follows we assume that (1.2) and (1.3) are satisfied. For simplicity we restrict our attention to positive times since the problem is treated analogously for negative times.

To obtain Theorem 2 stated below we need the following propositions.

PROPOSITION 2.1. *We assume that $u_0 \in L^2$ and $v_0 \in L^2$. Then there exist unique solutions u, v of* (2.1) *and a positive constant T such that*

$$u, v \in C([0,T]; L^2) \cap L^4(0,T; L^\infty).$$

For the proof, see [8, Thm. I].

PROPOSITION 2.2. *We assume that $u_0 \in H^2$ and $v_0 \in H^2$. Then there exist unique solutions u, v of* (2.1) *such that*

$$u, v \in C([0,T]; H^2) \cap L^4(0,T; W^{2,\infty})$$

for the same T as that given in Proposition 2.1.

For the proof, see [8, Thm. V].

PROPOSITON 2.3. *We assume that $u_0^{(n)}$, $v_0^{(n)} \in H^2$, u_0, $v_0 \in L^2$, and $||u_0^{(n)} - u_0||_2 + ||v_0^{(n)} - v_0||_2 \to 0$ as $n \to \infty$. We let $u^{(n)}$ and $v^{(n)}$ be the solutions constructed in Proposition* 2.2 *with data $u_0^{(n)}$ and $v_0^{(n)}$, respectively, and we let u and v be the solutions constructed in Proposition* 2.1 *with data u_0 and v_0, respectively. Then we have*

$$\begin{aligned} &\sup_{0\le t\le T} ||u^{(n)}(t) - u(t)||_2 + \sup_{0\le t\le T} ||v^{(n)}(t) - v(t)||_2 \\ &\quad + \left(\int_0^T ||u^{(n)} - u(t)||_\infty^4 dt\right)^{1/4} + \left(\int_0^T ||v^{(n)}(t) - v(t)||_\infty^4 dt\right)^{1/4} \\ &\quad \to 0 \quad \text{as} \quad n \to \infty, \end{aligned}$$

where T is the same as that given in Proposition 2.1.

For the proof, see [8, Thm. I′].

Using Proposition 2.2 we now prove the following proposition.

PROPOSITION 2.4. *We assume that $u_0 \in H^3$, $v_0 \in H^2$ satisfy the condition*

$$v_0 = \partial u_0 + i\frac{\lambda}{2}|u_0|^2 u_0.$$

Then there exist unique solutions u, v of (2.1) *such that*

$$u, v \in C([0,T]; H^2) \cap L^4(0,T; W^{2,\infty})$$

and

$$v = \partial u + i\frac{\lambda}{2}|u|^2 u \quad \text{in} \quad C([0,T]; L^2) \tag{2.2}$$

for the same time T as that given in Propositon 2.1.

Proof. By Proposition 2.2 it is sufficient to prove that (2.2) holds, namely

$$\sup_{0\le t\le T} ||v(t) - (\partial u(t) + i\frac{\lambda}{2}|u|^2u(t))||_2 = 0. \tag{2.3}$$

A direct calculation gives

$$L(\partial u) = \partial Lu = -2i\lambda u\bar{v}\partial u - i\lambda u^2\partial\bar{v} + \partial_u F(u)\cdot\partial u + \partial_{\bar{u}}F(u)\cdot\partial\bar{u}. \tag{2.4}$$

$$\begin{aligned} L(|u|^2u) &= 2|u|^2Lu + 2(\partial u)^2\bar{u} + 4u|\partial u|^2 + u^2(-\overline{Lu} + 2\partial^2\bar{u}) \\ &= 2|u|^2(-i\lambda u^2\bar{v} + F(u)) + 2(\partial u)^2\bar{u} + 4u|\partial u|^2 \\ &\quad + 2u^2\partial^2\bar{u} - u^2(i\lambda\bar{u}^2v + \overline{F(u)}) \qquad \text{(by (2.1))} \\ &= -2i\lambda|u|^2u^2\bar{v} + 2(\partial u)^2\bar{u} + 4u|\partial u|^2 + 2u^2\partial^2\bar{u} \\ &\quad - i\lambda|u|^4v + (2|u|^2F(u) - u^2\overline{F(u)}). \end{aligned} \tag{2.5}$$

We put $w = \partial u + \frac{i\lambda}{2}|u|^2u$, then we have by (2.4) and (2.5)

$$\begin{aligned} Lw = &-2i\lambda u\bar{v}\partial u - i\lambda u^2\partial\bar{v} + \lambda^2|u|^2u^2\bar{v} + i\lambda(\partial u)^2\bar{u} \\ &+ 2i\lambda u|\partial u|^2 + i\lambda u^2\partial^2\bar{u} + \frac{\lambda^2}{2}|u|^4v \\ &+ \left[\frac{i}{2}\lambda(2|u|^2F - u^2\bar{F}) + \partial_u F\cdot\partial u + \partial_{\bar{u}}F\cdot\partial\bar{u}\right]. \end{aligned}$$

Hence by the second identity of (2.1) we get

$$\begin{aligned} L(w-v) = &-2i\lambda u\bar{v}\partial u - i\lambda u^2\partial\bar{v} + \lambda^2|u|^2u^2\bar{v} + i\lambda(\partial u)^2\bar{u} \\ &+ 2i\lambda u|\partial u|^2 + i\lambda u^2\partial^2\bar{u} + \frac{\lambda^2}{2}|u|^4v - i\lambda v^2\bar{u} \\ &+ \left[\frac{i}{2}\lambda(2|u|^2F - u^2\bar{F}) + \partial_u F\cdot(\partial u - v) + \partial_{\bar{u}}F\cdot(\partial\bar{u} - \bar{v})\right]. \end{aligned} \tag{2.6}$$

We denote the jth term of the right-hand side of (2.6) by I_j. Using the definition $w = \partial u + \frac{i\lambda}{2}|u|^2u$ we obtain

$$\begin{aligned} I_1 &= -2i\lambda u\bar{v}\partial u \\ &= 2i\lambda u(\bar{w} - \bar{v})\partial u - 2i\lambda u\left(\overline{\partial u + \frac{i\lambda}{2}|u|^2u}\right)\partial u \\ &= 2i\lambda u(\bar{w} - \bar{v})\partial u - I_5 - \lambda^2|u|^4\partial u. \end{aligned}$$

This implies

$$I_1 + I_5 = 2i\lambda u(\bar{w} - \bar{v})\partial u - \lambda^2|u|^4\partial u. \tag{2.7}$$

We have

$$\begin{aligned} I_2 &= -i\lambda u^2\partial\bar{v} \\ &= i\lambda u^2(\partial\bar{w} - \partial\bar{v}) - i\lambda u^2\partial\left(\overline{\partial u + \frac{i\lambda}{2}|u|^2u}\right) \\ &= i\lambda u^2\partial(\bar{w} - \bar{v}) - I_6 - \frac{\lambda^2}{2}u^2\partial(|u|^2\bar{u}). \end{aligned} \tag{2.8}$$

Since

$$
\begin{aligned}
-\frac{\lambda^2}{2}u^2\partial(|u|^2\bar{u}) &= -\frac{\lambda^2}{2}|u|^4\partial u - \lambda^2|u|^2u^2\partial\bar{u} \\
&= -\frac{\lambda^2}{2}|u|^4(w-v) + \frac{\lambda^2}{2}|u|^4\left(\frac{i\lambda}{2}|u|^2u - v\right) - \lambda^2|u|^2u^2\partial\bar{u} \\
&= -\frac{\lambda^2}{2}|u|^2(w-v) + \frac{i\lambda^3}{4}|u|^6u - \lambda^2|u|^2u^2\partial\bar{u} - I_7,
\end{aligned}
$$

we obtain from (2.8)

$$
I_2 + I_6 + I_7 = i\lambda u^2\partial(\bar{w}-\bar{u}) - \frac{\lambda^2}{2}|u|^2(w-v) + \frac{i\lambda^3}{4}|u|^6u - \lambda^2|u|^2u^2\partial\bar{u}. \tag{2.9}
$$

We next consider the contributions of the terms I_3, I_4, and I_8, the last term of the right-hand side of (2.7), and the last two terms of the right-hand side of (2.9). We have

(2.10)

$$
\begin{aligned}
&\lambda^2|u|^2u^2\bar{v} + i\lambda(\partial u)^2\bar{u} - i\lambda v^2\bar{u} - \lambda^2|u|^4\partial u + \frac{i\lambda^3}{4}|u|^6u - \lambda^2|u|^2u^2\partial\bar{u} \\
&= i\lambda((\partial u)^2 + i\lambda|u|^2u\partial u - v^2)\bar{u} - \lambda^2|u|^2u^2(\partial\bar{u}-\bar{v}) + \frac{i\lambda^3}{4}|u|^6u \\
&= i\lambda\left(\left(\partial u + \frac{i\lambda}{2}|u|^2u\right)^2 - v^2\right)\bar{u} - \lambda^2|u|^2u^2\left(\overline{\partial u + \frac{i\lambda}{2}|u|^2u} - \bar{v}\right) \\
&= i\lambda(w^2-v^2)\bar{u} - \lambda^2|u|^2u^2(\bar{w}-\bar{v}).
\end{aligned}
$$

By the gauge condition $F(e^{i\theta}\psi) = e^{i\theta}F(\psi)$, $\theta \in \mathbb{R}$, we see that $F(|\psi|) = F(\frac{|\psi|}{\psi}\psi) = \frac{|\psi|}{\psi}F(\psi)$. Hence $F(\psi)$ is written as

$$
F(\psi) = G(|\psi|^2)\psi \tag{2.11}
$$

if we put

$$
G(s^2) = \begin{cases} F(s)/s & \text{for} \quad s > 0, \\ 0 & \text{for} \quad s = 0. \end{cases}
$$

Using (2.11), the last term of the right-hand side of (2.6), I_9, is rewritten as

$$
\begin{aligned}
I_9 = \frac{i}{2}\lambda&(2G(|u|^2)|u|^2u - \bar{G}(|u|^2)|u|^2u) \\
&+ (G'(|u|^2)u\cdot\bar{u} + G(|u|^2))(\partial u - v) + G'(|u|^2)u\cdot u(\partial\bar{u}-\bar{v}),
\end{aligned}
$$

where $G'(|u|^2) = \partial_{|u|^2}G(|u|^2)$. From the condition that there exists a function $H \in C^1(\mathbb{R}^2;\mathbb{R})$ satisfying $F(\zeta) = \partial H/\partial\bar{\zeta}$, it follows that $G = \bar{G}$. Hence

$$
\begin{aligned}
I_9 &= G'(|u|^2)|u|^2\left(\partial u + \frac{i}{2}\lambda|u|^2u - v\right) \\
&\quad + G(|u|^2)\left(\partial u + \frac{i}{2}\lambda|u|^2u - v\right) + G'(|u|^2)u^2\left(\overline{\partial u + \frac{i}{2}\lambda|u|^2u} - \bar{v}\right) \qquad (2.12) \\
&= G'(|u|^2)|u|^2(w-v) + G(|u|^2)(w-v) + G'(|u|^2)u^2(\bar{w}-\bar{v}) \\
&= \partial_uF\cdot(w-v) + \partial_{\bar{u}}F\cdot(\bar{w}-\bar{v}).
\end{aligned}
$$

By (2.6), (2.7), (2.9), (2.10), and (2.12)

$$
\begin{aligned}
L(w-v) &= \sum_{j=1}^{9} I_j \\
&= 2i\lambda u\partial u(\bar{w}-\bar{v}) + i\lambda u^2\partial(\bar{w}-\bar{v}) - \frac{\lambda^2}{2}|u|^2(w-v) + i\lambda(w^2-v^2)\bar{u} \\
&\quad - \lambda^2|u|^2u^2(\bar{w}-\bar{v}) + \partial_u F\cdot(w-v) + \partial_{\bar{u}}F\cdot(\bar{w}-\bar{v}).
\end{aligned} \tag{2.13}
$$

Multiplying both sides of (2.13) by $\bar{w}-\bar{v}$, taking the imaginary part, and integrating over space with integration by parts for the second term of the right-hand side of (2.13), we obtain

$$
\begin{aligned}
\frac{d}{dt}||w-v||_2^2 \le (&2|\lambda|\ ||u||_\infty||\partial u||_\infty + 2|\lambda|\ ||w+v||_\infty||u||_\infty \\
&+ 2\lambda^2||u||_\infty^4 + ||\partial_u F||_\infty + ||\partial_{\bar{u}}F||_\infty)||w-v||_2^2,
\end{aligned}
$$

from which we have

$$
\begin{aligned}
||w-v||_2^2 &\le ||w(0)-v(0)||_2^2 \\
&\quad\times \exp\left(C\int_0^t (||u||_\infty||\partial u||_\infty + ||u||_\infty||v||_\infty + ||u||_\infty^4 + ||u||_\infty^{p_1-1} + ||u||_\infty^{p_2-1})ds\right).
\end{aligned}
$$

This implies the desired identity (2.3). Therefore we have the proposition.

We next prove the following proposition.

PROPOSITION 2.5. *We assume that $u_0 \in H^1$, $v_0 \in L^2$ satisfy the conditions*

$$||u_0||_2^2 < 2\pi/|\lambda|$$

and

$$v_0 = \partial u_0 + i\frac{\lambda}{2}|u_0|^2u_0.$$

Then there exist unique solutions u, v of (2.1) *such that*

$$
\begin{cases}
u \in C([0,T];H^1)\cap L^4(0,T;W^{1,\infty}), \\
v \in C([0,T];L^2)\cap L^4(0,T;L^\infty), \\
v = \partial u + i\frac{\lambda}{2}|u|^2u \quad \textit{in} \quad C([0,T];L^2),
\end{cases} \tag{2.14}
$$

for the same time T as that given in Proposition 2.1. *Moreover, the map $u_0 \mapsto u$ is continuous from $\{u_0 \in H^1; ||u_0||_2^2 < 2\pi/|\lambda|\}$ with topology induced from H^1 to $C([0,T];H^1)\cap L^4(0,T;W^{1,\infty})$.*

Proof. We let $u_0^{(n)} \in H^2$, $v_0^{(n)} \in H^2$ satisfy $||u_0^{(n)}||_2 = ||u_0||_2$ for any $n \in \mathbb{N}$ and $||u_0^{(n)}-u_0||_2 + ||v_0^{(n)}-v_0||_2 \to 0$ as $n\to\infty$. Then by Proposition 2.4 we see that there exist unique solutions $u^{(n)}$, $v^{(n)}$ of (2.1) with the data $u^{(n)}(0) = u_0^{(n)}$, $v^{(n)}(0) = v_0^{(n)}$ satisfying

$$
\begin{cases}
u^{(n)}, v^{(n)} \in C([0,T];H^2)\cap L^4(0,T;W^{2,\infty}), \\
v^{(n)} = \partial u^{(n)} + i\frac{\lambda}{2}|u^{(n)}|^2u^{(n)} \quad \text{in} \quad C([0,T];L^2).
\end{cases} \tag{2.15}
$$

From Proposition 2.3 it follows that

$$\lim_{n\to\infty} u^{(n)} = u \text{ and } \lim_{n\to\infty} v^{(n)} = v \text{ strongly in } C([0,T];L^2)\cap L^4(0,T;L^\infty), \tag{2.16}$$

where u, v are the unique solutions to (2.1) with the data $u(0) = u_0$, $v(0) = v_0$. Hence Proposition 2.5 is obtained by showing

$$\begin{cases} u \in C([0,T]; H^1) \cap L^4(0,T; W^{1,\infty}), \\ v = \partial u + i\frac{\lambda}{2}|u|^2 u \quad \text{in} \quad C([0,T]; L^2). \end{cases} \tag{2.17}$$

We now prove (2.17).

By the second line of (2.15) and the first identity of (2.1) we have

$$Lu^{(n)} = -i\lambda(u^{(n)})^2 \partial \bar{u}^{(n)} - \frac{\lambda^2}{2}|u^{(n)}|^4 u^{(n)} + F(u^{(n)}). \tag{2.18}$$

Multiplying both sides of (2.18) by $\bar{u}^{(n)}$, taking the imaginary part, and integrating in x, we obtain

$$||u^{(n)}(t)||_2 = ||u_0^{(n)}||_2 = ||u_0||_2, \tag{2.19}$$

where we have used the condition (1.3).

We put

$$\psi_j^{(n)} = G_n^{-(6-j)/4} u^{(n)}, \quad j = 2, 3, \dots, 6$$

with

$$G_n = \exp\left(-i\frac{\lambda}{2}\int_{-\infty}^{x} |u^{(n)}|^2 dy\right),$$

then it is clear that

$$\psi_j^{(n)} = G_n^{-1/4} \psi_{j+1}^{(n)}.$$

Hence

$$\partial \psi_j^{(n)} = G_n^{-1/4}\left(i\frac{\lambda}{8}|u^{(n)}|^2 \psi_{j+1}^{(n)} + \partial \psi_{j+1}^{(n)}\right), \tag{2.20}$$

from which it follows that

$$\begin{aligned} ||\partial \psi_j^{(n)}||_2^2 &= ||\partial \psi_{j+1}^{(n)}||_2^2 + \frac{\lambda}{4}\operatorname{Im}(|u^{(n)}|^2 \psi_{j+1}^{(n)}, \partial \psi_{j+1}^{(n)}) \\ &\quad + \frac{\lambda^2}{4}||\,|u^{(n)}|^2 \psi_{j+1}^{(n)}||_2^2 \\ &\geq ||\partial \psi_{j+1}^{(n)}||_2^2 - \frac{|\lambda|}{4}||\psi_{j+1}^{(n)}||_6^3 ||\partial \psi_{j+1}^{(n)}||_2, \end{aligned} \tag{2.21}$$

since

$$|u^{(n)}| = |\psi_j^{(n)}| = |\psi_{j+1}^{(n)}|. \tag{2.22}$$

We apply the Gagliardo–Nirenberg inequality of the form

$$||f||_6^6 \leq \frac{4}{\pi^2}||\partial f||_2^2 ||f||_2^4 \tag{2.23}$$

to (2.21) to obtain

$$||\partial \psi_j^{(n)}||_2^2 \geq \left(1 - \frac{|\lambda|}{2\pi}||\psi_{j+1}^{(n)}||_2^2\right)||\partial \psi_{j+1}^{(n)}||_2^2.$$

We use (2.19) and (2.21) to see that

$$||\partial \psi_j^{(n)}||_2^2 \geq \left(1 - \frac{|\lambda|}{2\pi}||u_0||_2^2\right)||\partial \psi_{j+1}^{(n)}||_2^2,$$

and from this and the condition $||u_0||_2^2 < 2\pi/|\lambda|$ it follows that

$$||\partial\psi_{j+1}^{(n)}||_2^2 \le \left(1 - \frac{|\lambda|}{2\pi}||u_0||_2^2\right)^{-1} ||\partial\psi_j^{(n)}||_2^2. \tag{2.24}$$

The definition of $\psi_j^{(n)}$ implies

$$\partial\psi_2^{(n)} = G_n^{-1}v^{(n)} \quad \text{and} \quad \psi_6^{(n)} = u^{(n)}. \tag{2.25}$$

By (2.24) and (2.25)

$$||\partial\psi_{j+1}^{(n)}||_2 \le \left(1 - \frac{|\lambda|}{2\pi}||u_0||_2^2\right)^{-(j-1)/2} ||v^{(n)}||_2 \tag{2.26}$$

for $j = 1, 2, \dots, 5$.

We now prove $\{u^{(n)}\}$ is a Cauchy sequence in $C([0,T]; H^1)$. Using (2.20) and (2.22) we write

$$\begin{aligned}\partial\psi_j^{(n)} - \partial\psi_j^{(m)} = G_n^{-1/4}&\left(i\frac{\lambda}{8}(|\psi_{j+1}^{(n)}|^2\psi_{j+1}^{(n)} - |\psi_{j+1}^{(m)}|^2\psi_{j+1}^{(m)}) + \partial\psi_{j+1}^{(n)} - \partial\psi_{j+1}^{(m)}\right)\\ &+ (G_n^{-1/4} - G_m^{-1/4})\left(i\frac{\lambda}{8}|\psi_{j+1}^{(m)}|^2\psi_{j+1}^{(m)} + \partial\psi_{j+1}^{(m)}\right).\end{aligned}$$

From the mean value theorem and the Schwarz inequality we have

$$\begin{aligned}||\partial\psi_{j+1}^{(n)} - \partial\psi_{j+1}^{(m)}||_2 &\le ||\partial\psi_j^{(n)} - \partial\psi_j^{(m)}||_2\\ &+ C(||\psi_{j+1}^{(n)}||_\infty^2 + ||\psi_{j+1}^{(m)}||_\infty^2)||\psi_{j+1}^{(n)} - \psi_{j+1}^{(m)}||_2\\ &+ C(||\psi_{j+1}^{(m)}||_6^3 + ||\partial\psi_{j+1}^{(m)}||_2)(||u^{(m)}||_2 + ||u^{(m)}||_2)||u^{(n)} - u^{(m)}||_2.\end{aligned} \tag{2.27}$$

We apply the Gagliardo–Nirenberg inequality, (2.26), and (2.22) to (2.27) to obtain

$$\begin{aligned}||\partial\psi_{j+1}^{(n)} - \partial\psi_{j+1}^{(m)}||_2 &\le ||\partial\psi_j^{(n)} - \partial\psi_j^{(m)}||_2\\ &+ C(||v^{(n)}||_2||u^{(n)}||_2 + ||v^{(m)}||_2||u^{(m)}||_2)||\psi_{j+1}^{(n)} - \psi_{j+1}^{(m)}||_2\\ &+ C(||v^{(m)}||_2||u^{(m)}||_2^2 + ||v^{(m)}||_2)(||u^{(n)}||_2 + ||u^{(m)}||_2)||u^{(n)} - u^{(m)}||_2.\end{aligned} \tag{2.28}$$

On the other hand, we have from

$$\psi_j^{(n)} - \psi_j^{(m)} = G_n^{-(6-j)/4}(u^{(n)} - u^{(m)}) + (G_n^{-(6-j)/4} - G_m^{-(6-j)/4})u^{(m)}$$

that

$$\begin{aligned}||\psi_j^{(n)} - \psi_j^{(m)}||_2 &\le ||u^{(n)} - u^{(m)}||_2\\ &+ C||u^{(m)}||_2(||u^{(n)}||_2 + ||u^{(m)}||_2)||u^{(n)} - u^{(m)}||_2\end{aligned} \tag{2.29}$$

for $j = 2, 3, \dots, 6$.

Hence (2.16), (2.28), and (2.29) imply that there exists a positive constant C which does not depend on n, m, j and satisfies

$$||\partial\psi_{j+1}^{(n)} - \partial\psi_{j+1}^{(m)}||_2 \le ||\partial\psi_j^{(n)} - \partial\psi_j^{(m)}||_2 + C||u^{(n)} - u^{(m)}||_2$$

for $j = 2, 3, \dots, 5$, from which and (2.25) it follows that

$$||\partial u^{(n)} - \partial u^{(m)}||_2 \le ||\partial\psi_2^{(n)} - \partial\psi_2^{(m)}||_2 + C||u^{(n)} - u^{(m)}||_2. \tag{2.30}$$

In the same way as in the proof of (2.29) we have

$$||\partial\psi_2^{(n)} - \partial\psi_2^{(m)}||_2 \le ||v^{(n)} - v^{(m)}||_2 + C||v^{(n)}||_2(||u^{(n)}||_2 + ||u^{(m)}||_2)||u^{(n)} - u^{(m)}||_2. \tag{2.31}$$

Therefore we obtain by (2.16), (2.30), and (2.31)

$$||\partial u^{(n)} - \partial u^{(m)}||_2 \le ||v^{(n)} - v^{(m)}||_2 + C||u^{(n)} - u^{(m)}||_2. \tag{2.32}$$

This and (2.16) imply that $\{u^{(n)}\}$ is a Cauchy sequence in $C([0,T];H^1)$. Hence,

$$\lim_{n\to\infty} u^{(n)} = u \quad \text{strongly in} \quad C([0,T];H^1). \tag{2.33}$$

From (2.33) and the Gagliardo–Nirenberg inequality it is clear that

$$\lim_{n\to\infty} |u^{(n)}|^2 u^{(n)} = |u|^2 u \quad \text{strongly in} \quad C([0,T];L^2). \tag{2.34}$$

By (2.16), (2.33), and (2.34) we have the second line of (2.17). The proof of (2.17) is completed by showing

$$\partial u \in L^4(0,T;L^\infty). \tag{2.35}$$

Since $u \in C([0,T];H^1) \cap L^4(0,T;L^\infty)$, the Gagliardo–Nirenberg inequality gives $|u|^2 u \in L^4(0,T;L^\infty)$. Hence the second line of (2.17) and (2.16) yields (2.35). This completes the proof of Proposition 2.5.

We next prove the local existence of solutions to the original equation (1.1).

THEOREM 2. *We assume that $\phi \in H^1(\mathbb{R})$ and $||\phi||_2^2 < 2\pi/|\lambda|$. Then there exist a unique solution ψ of* (1.1) *and a positive constant T such that*

$$\psi \in C([0,T];H^1) \cap L^4(0,T;W^{1,\infty}).$$

Moreover, the map $\phi \mapsto \psi$ is continuous from $\{\phi \in H^1; ||\phi||_2^2 < 2\pi/|\lambda|\}$ with topology induced from H^1 to $C([0,T];H^1) \cap L^4(0,T;W^{1,\infty})$.

Proof. For any $\phi \in H^1(\mathbb{R})$ we define u_0 and v_0 as follows:

$$u_0 = \exp\left(-i\lambda \int_{-\infty}^{x} |\phi|^2 dy\right)\phi,$$

$$v_0 = \partial u_0 + \frac{i}{2}\lambda |u_0|^2 u_0.$$

Then it is clear that $u_0 \in H^1$, $v_0 \in L^2$, and $||u_0||_2^2 = ||\phi||_2^2 < 2\pi/|\lambda|$. Hence by Proposition 2.5 there exist unique solutions u, v of (2.1) satisfying (2.14). We define ψ by

$$\psi(t,x) = \exp\left(i\lambda \int_{-\infty}^{x} |u(t,y)|^2 dy\right) u(t,x). \tag{2.36}$$

Theorem 2 is established if we prove that ψ satisfies (1.1). By (2.14) and the first equation of (2.1) for u we obtain

$$Lu = -i\lambda u^2 \partial\bar{u} - \frac{\lambda^2}{2}|u|^4 u + F(u). \tag{2.37}$$

A direct calculation and (2.36) give

(2.38)
$$L\psi = \exp\left(i\lambda\int_{-\infty}^{x}|u|^2dy\right)\left(Lu + i\lambda\left(L\int_{-\infty}^{x}|u|^2dy\right)u + 2i\lambda|u|^2\partial u - \lambda^2|u|^4u\right).$$

By (2.37)

(2.39)
$$\begin{aligned}\partial_t|u|^2 &= 2\,\mathrm{Im}(\bar{u}(-\partial^2u - i\lambda u^2\partial\bar{u}))\\ &= -\partial(2\,\mathrm{Im}(\bar{u}\partial u)) - 2\lambda|u|^2\,\mathrm{Re}(u\partial\bar{u})\\ &= -\partial\left(2\,\mathrm{Im}(\bar{u}\partial u) + \frac{\lambda}{2}|u|^4\right),\end{aligned}$$

(2.40)
$$\begin{aligned}i\lambda L\int_{-\infty}^{x}|u|^2dy &= -\lambda\int_{-\infty}^{x}\partial_t|u|^2dy + i\lambda\partial|u|^2\\ &= 2\lambda\,\mathrm{Im}(\bar{u}\partial u) + \frac{\lambda^2}{2}|u|^4 + i\lambda\partial|u|^2 \quad \text{(by (2.37))}\\ &= -i\lambda(\bar{u}\partial u - u\partial\bar{u}) + \frac{\lambda^2}{2}|u|^4 + i\lambda(\bar{u}\partial u + u\partial\bar{u})\\ &= 2i\lambda u\partial\bar{u} + \frac{\lambda^2}{2}|u|^4.\end{aligned}$$

From (2.37)–(2.40) we have

(2.41)
$$\begin{aligned}L\psi = \exp(i\lambda\int_{-\infty}^{x}|u|^2dy)\Big(&-i\lambda u^2\partial\bar{u} - \frac{\lambda^2}{2}|u|^4u + F(u)\\ &+ 2i\lambda u^2\partial\bar{u} + \frac{\lambda^2}{2}|u|^4u + 2i\lambda|u|^2\partial u\\ &- \lambda^2|u|^4u\Big)\\ = \exp\left(i\lambda\int_{-\infty}^{x}|u|^2dy\right)&(i\lambda u^2\partial\bar{u} + 2i\lambda|u|^2\partial u - \lambda^2|u|^4u + F(u))\\ = \exp\left(i\lambda\int_{-\infty}^{x}|u|^2dy\right)&(2i\lambda|u|^2(\partial u + i\lambda|u|^2) + i\lambda u^2(\partial\bar{u} - i\lambda|u|^2\bar{u}) + F(u)).\end{aligned}$$

Since

$$\partial\psi = \exp(i\lambda\int_{-\infty}^{x}|u|^2dy)(\partial u + i\lambda|u|^2u),$$
$$|u|^2 = |\psi|^2 \quad \text{and} \quad u^2 = \psi^2\exp(-2i\lambda\int_{-\infty}^{x}|u|^2dy),$$

we have by (2.41)

$$L\psi = 2i\lambda|\psi|^2\partial\psi + i\lambda\psi^2\partial\bar{\psi} + \exp\left(i\lambda\int_{-\infty}^{x}|u|^2dy\right)F(u).$$

Gauge condition of F implies the desired identity

$$L\psi = i\lambda\partial(|\psi|^2\psi) + F(\psi).$$

The last equation makes sense in $C([0,T];H^{-1})$ and all the computations above are justified in $C([0,T];H^{-2})$. We could go though with the computations using approximate solutions $u^{(n)}$ as in the proof of Proposition 2.5 before limiting procedure at the final stage. This completes the proof of Theorem 2.

We are now in a position to prove Theorem 1.

Proof of Theorem 1. The iterative use of the same argument as that in the proof of Theorem 2 and a priori estimates of solution ψ of (1.1) in H^1 yield Theorem 1. Then it is sufficient to prove

$$\sup_{t\in[0,T]} ||\psi(t)||_{H^1} \le C(||\phi||_{H^1}). \tag{2.42}$$

We prove (2.42). After a long tedious calculation (see the Appendix) we arrive at

$$\frac{d}{dt}\left(||\partial\psi||_2^2 + \frac{3}{2}\lambda\,\mathrm{Im}(|\psi|^2\psi,\partial\psi) + \frac{1}{2}\lambda^2||\psi||_6^6 + \int H(\psi)dx\right) = 0,$$

and from this and (1.3) we have

(2.43)

$$\begin{aligned} ||\psi(t)||_2 &= ||\phi||_2 \\ E \equiv ||\partial\psi||_2^2 &+ \frac{3}{2}\lambda\,\mathrm{Im}(|\psi|^2\psi,\partial\psi) + \frac{1}{2}\lambda^2||\psi||_6^6 + \int H(\psi)dx \\ &= ||\partial\phi||_2^2 + \frac{3}{2}\lambda\,\mathrm{Im}(|\phi|^2\phi,\partial\phi) + \frac{1}{2}\lambda^2||\phi||_6^6 + \int H(\phi)dx. \end{aligned}$$

By the Gagliardo–Nirenberg inequality and (1.3) we get

$$E \le C(||\phi||_{H^1}). \tag{2.44}$$

We put

$$G = \exp\left(-i\frac{\lambda}{2}\int_{-\infty}^{x}|\psi(t,y)|^2dy\right).$$

Then

$$\begin{aligned} ||\partial\psi||_2^2 + \frac{3}{2}\lambda\,\mathrm{Im}(|\psi|^2\psi,\partial\psi) &= ||\partial\psi||_2^2 - \frac{3}{2}\lambda\,\mathrm{Re}(i|\psi|^2\psi,\partial\psi) \\ &= \left\|\partial\psi - \frac{3}{4}i\lambda|\psi|^2\psi\right\|_2^2 - \frac{9}{16}\lambda^2||\psi||_6^6 \\ &= ||\partial(G^{3/2}\psi)||_2^2 - \frac{9}{16}\lambda^2||\psi||_6^6. \end{aligned}$$

Hence

$$E = ||\partial(G^{3/2}\psi)||_2^2 - \frac{1}{16}\lambda^2||\psi||_6^6 + \int H(\psi)dx.$$

By (1.3) we see that

$$E \ge ||\partial(G^{3/2}\psi)||_2^2 - M(||\psi||_2) - \left(\frac{\lambda^2}{16}+\mu\right)||G^{3/2}\psi||_6^6.$$

We apply (2.23) and (2.43) to the above to obtain

$$E + M(||\phi||_2) \ge \left(1 - \frac{4}{\pi^2}\left(\frac{\lambda^2}{16}+\mu\right)||\phi||_2^4\right)||\partial(G^{3/2}\psi)||_2^2.$$

From this and (2.44) it follows that

$$||\partial(G^{3/2}\psi)||_2^2 \le C(||\phi||_{H^1}). \tag{2.45}$$

We let $\psi_j = G^{(6-j)/4}\psi$ for $0 \le j \le 6$. We have by (2.23) and (2.43)

$$\begin{aligned}
||\partial\psi_j||_2^2 &= ||\partial(G^{1/4}\psi_{j+1})||_2^2 \\
&= ||G^{1/4}\left(\partial\psi_{j+1} - \frac{1}{8}i\lambda|\psi|^2\psi_{j+1}\right)||_2^2 \\
&= ||\partial\psi_{j+1}||_2^2 - \frac{1}{4}\lambda\,\mathrm{Im}(\partial\psi_{j+1}, |\psi|^2\psi_{j+1}) + \frac{\lambda^2}{64}||\psi||_6^6 \\
&\ge \left(1 - \frac{|\lambda|}{2\pi}||\phi||_2^2\right)||\partial\psi_{j+1}||_2^2.
\end{aligned}$$

Therefore

$$||\partial\psi_0||_2^2 \ge \left(1 - \frac{|\lambda|}{2\pi}||\phi||_2^2\right)^6 ||\partial\psi_6||_2^2. \tag{2.46}$$

Since $\psi = \psi_6$, $G^{3/2}\psi = \psi_0$, by (2.45) and (2.46) we have the desired estimate (2.42). This completes the proof of Theorem 1.

Appendix.

LEMMA A. *Let ψ be the solution of* (1.1) *constructed in Theorem* 2. *Then we have*

(a) $$\frac{d}{dt}||\psi||_2^2 = 0,$$

(b) $$\frac{d}{dt}(||\partial\psi||_2^2 + \frac{3}{2}\lambda\,\mathrm{Im}(|\psi|^2\psi, \partial\psi) + \frac{1}{2}\lambda^2||\psi||_6^6 + \int H(\psi)dx) = 0.$$

Proof. We first note that the computation given below is rather formal, but it can be justified using H^2-solutions ψ_k with the data $\phi_k \in H^2$, which satisfy the following continuous dependence such that

$$\lim_{k\to\infty}||\phi_k - \phi||_{H^1} = 0 \quad \text{implies} \quad \lim_{k\to\infty}||\psi_k(t) - \psi(t)||_{H^1} = 0$$

(see Proposition 2.5). In what follows we use the integration by parts without particular comments. We have by (1.1)

$$\begin{aligned}
\partial_t|\psi|^2 &= 2\,\mathrm{Im}(\bar\psi i\partial_t\psi) \\
&= 2\,\mathrm{Im}(\bar\psi(-\partial^2\psi + i\lambda\partial(|\psi|^2\psi) + F(\psi))) \\
&= -\partial(2\,\mathrm{Im}(\bar\psi\partial\psi)) + 2\lambda\,\mathrm{Re}(\bar\psi\partial(|\psi|^2\partial si)).
\end{aligned} \tag{A.1}$$

Since $\mathrm{Re}(\bar\psi\partial(|\psi|^2\psi)) = \frac{3}{4}\partial|\psi|^4$ we have by (A.1)

$$\partial_t|\psi|^2 = \partial\left(-2\partial\,\mathrm{Im}(\bar\psi\partial\psi) + \frac{3}{2}\lambda|\psi|^4\right), \tag{A.2}$$

from which (a) follows.

We have

$$
\begin{aligned}
\frac{d}{dt}||\partial\psi||_2^2 &= 2\,\mathrm{Re}(\partial\psi, \partial_t\partial\psi) \\
&= -2\,\mathrm{Re}(\partial^2\psi, \partial_t\psi) \\
&= -2\,\mathrm{Re}(-i\partial_t\psi + i\lambda\partial(|\psi|^2\psi) + F(\psi), \partial_t\psi) \qquad \text{(by (1.1))} \\
&= 2\lambda\,\mathrm{Im}(\partial(|\psi|^2\psi), \partial_t\psi) - 2\,\mathrm{Re}(F(\psi), \partial_t\psi) \\
&= -2\lambda\,\mathrm{Im}(|\psi|^2\psi, \partial_t\partial_\psi) - \frac{d}{dt}\int H(\psi)dx \\
&\qquad \text{(by the condition of } F\text{)}.
\end{aligned} \tag{A.3}
$$

We next consider the first term of the right-hand side of (A.3). We have

$$
\begin{aligned}
-\mathrm{Im}(|\psi|^2\psi, \partial_t\partial\psi) = -\frac{d}{dt}(\mathrm{Im}(|\psi|^2\psi, \partial\psi)) \\
+ \mathrm{Im}((\partial_t|\psi|^2)\psi, \partial\psi) + \mathrm{Im}(|\psi|^2\partial_t\psi, \partial\psi).
\end{aligned} \tag{A.4}
$$

We apply (A.2) to the second term of the right-hand side of (A.4) to obtain

$$
\begin{aligned}
\mathrm{Im}((\partial_t|\psi|^2)\psi, \partial\psi) &= \mathrm{Im}\left(\left(\partial\left(-2\,\mathrm{Im}(\bar{\psi}\partial\psi) + \frac{3}{2}\lambda|\psi|^4\right)\right)\psi, \partial\psi\right) \\
&= 2(\partial(\mathrm{Im}\,\bar{\psi}\partial\psi), \mathrm{Im}\,\bar{\psi}\partial\psi) + \frac{3}{2}\lambda\,\mathrm{Im}((\partial|\psi|^4)\psi, \partial\psi) \\
&= -\frac{3}{2}\lambda\,\mathrm{Im}(|\psi|^4\psi, \partial^2\psi)
\end{aligned} \tag{A.5}
$$

By using (1.1) in the third term of the right-hand side of (A.4), we have

$$
\begin{aligned}
&\mathrm{Im}(|\psi|^2\partial_t\psi, \partial\psi) \\
&\quad = -\mathrm{Im}((\partial|\psi|^2)\partial_t\psi, \psi) - \mathrm{Im}(|\psi|^2\partial_t\partial\psi, \partial\psi) \\
&\quad = \mathrm{Re}(\partial|\psi|^2 \cdot (-\partial^2\psi + i\lambda\partial(|\psi|^2\psi) + F(\psi)), \psi) + \mathrm{Im}(|\psi|^2\psi, \partial_t\partial\psi) \\
&\quad = -\mathrm{Re}(\partial|\psi|^2 \cdot \partial^2\psi, \psi) - \lambda\,\mathrm{Im}(\partial|\psi|^2 \cdot \partial(|\psi|^2\psi), \psi) + \mathrm{Im}(|\psi|^2\psi, \partial_t\partial\psi).
\end{aligned} \tag{A.6}
$$

Since

$$
\begin{aligned}
&\mathrm{Re}(\partial(|\psi|^2\psi), \partial^2\psi) \\
&\quad = \mathrm{Re}(\partial|\psi|^2 \cdot \psi, \partial^2\psi) + \mathrm{Re}(|\psi|^2\partial\psi, \partial^2\psi) \\
&\quad = \mathrm{Re}(\partial|\psi|^2 \cdot \partial^2\psi, \psi) - \frac{1}{2}(\partial|\psi|^2, |\partial\psi|^2) \\
&\quad = \mathrm{Re}(\partial|\psi|^2 \cdot \partial^2\psi, \psi) - \frac{1}{2}(\partial|\psi|^2, \frac{1}{2}\partial^2|\psi|^2 - \mathrm{Re}(\bar{\psi}\partial^2\psi)) \\
&\quad = \frac{3}{2}\,\mathrm{Re}(\partial|\psi|^2 \cdot \partial^2\psi, \psi),
\end{aligned}
$$

the first term of the right-hand side of (A.6) is written as

$$
\begin{aligned}
(A.7)\qquad & -\operatorname{Re}(\partial|\psi|^2\cdot\partial^2\psi,\psi)\\
&=-\frac{2}{3}\operatorname{Re}(\partial(|\psi|^2\psi),\partial^2\psi)\\
&=-\frac{2}{3}\operatorname{Re}(\partial(|\psi|^2\psi),-i\partial_t\psi+i\lambda\partial(|\psi|^2\psi)+F(\psi))\quad\text{(by (1.1))}\\
&=\frac{2}{3}\operatorname{Im}(\partial(|\psi|^2\psi),\partial_t\psi)\quad\text{(by the condition of } F)\\
&=-\frac{2}{3}\operatorname{Im}(|\psi|^2\psi,\partial_t\partial\psi).
\end{aligned}
$$

By (A.4), (A.6), and (A.7)

$$
\begin{aligned}
-\operatorname{Im}(|\psi|^2\psi,\partial_t\partial\psi)&=-\frac{d}{dt}(\operatorname{Im}(|\psi|^2\psi,\partial\psi))-\frac{3}{2}\lambda\operatorname{Im}(|\psi|^4\psi,\partial^2\psi)\\
&\quad-\lambda\operatorname{Im}(\partial|\psi|^2\cdot\partial(|\psi|^2\psi),\psi)+\frac{1}{3}\operatorname{Im}(|\psi|^2\psi,\partial_t\partial\psi)\\
&=-\frac{d}{dt}(\operatorname{Im}(|\psi|^2\psi,\partial\psi))+2\lambda\operatorname{Im}(|\psi|^4\partial^2\psi,\psi)+\frac{1}{3}\operatorname{Im}(|\psi|^2\psi,\partial_t\partial\psi).
\end{aligned}
$$

From above it follows that

$$
(A.8)\qquad -\operatorname{Im}(|\psi|^2\psi,\partial_t\partial\psi)=-\frac{3}{4}\frac{d}{dt}(\operatorname{Im}(|\psi|^2\psi,\partial\psi))+\frac{3}{2}\lambda\operatorname{Im}(|\psi|^4\partial^2\psi,\psi).
$$

From (A.3) and (A.8) it follows that

$$
(A.9)\qquad \frac{d}{dt}\|\partial\psi\|_2^2=-\frac{3}{2}\lambda\frac{d}{dt}(\operatorname{Im}(|\psi|^2\psi,\partial\psi))+3\lambda^2\operatorname{Im}(|\psi|^4\partial^2\psi,\psi)+\frac{d}{dt}\int H(\psi)dx.
$$

On the other hand, we have by (1.1)

$$
\begin{aligned}
(A.10)\qquad \operatorname{Im}(|\psi|^4\partial^2\psi,\psi)&=\operatorname{Im}(|\psi|^4(-i\partial_t\psi+i\lambda\partial(|\psi|^2\psi)+F(\psi)),\psi)\\
&=-\operatorname{Re}(|\psi|^4\partial_t\psi,\psi)+\lambda\operatorname{Re}(|\psi|^4(\partial|\psi|^2\cdot\psi+|\psi|^2\partial\psi),\psi)\\
&=-\frac{1}{6}\frac{d}{dt}\|\psi\|_6^6.
\end{aligned}
$$

Hence (A.9) and (A.10) yield the desired identity (b).

Acknowledgments. This work was performed when one of the authors (N.H) was at Université de Paris Sud. N.H. is grateful to Professors J. Ginibre and M. Fontannaz for their kind hospitality at Laboratoire de Physique Théorique et Hautes Energy, Université de Paris Sud. The authors also would like to thank Professor Y. Tsutsumi for fruitful discussions on the derivative nonlinear Schrödinger equation.

REFERENCES

[1] G. Boling and T. Shaobin, *On smooth solutions to the initial value problem jor the mixed nonlinear Schrödinger equations*, Proc. Royal Soc. Edinburgh, Sect A, 119 (1991), pp. 31–45.

[2] Y.M. Chen, *The initial-boundary value problem for a class of nonlinear Schrödinger equations*, Acta Math. Sci., 6 (1986), pp. 405–418.

[3] N. HAYASHI, *The initial value problem for the derivative nonlinear Schrödinger equation*, Nonlinear Anal., 20 (1993), pp. 823–833.

[4] N. HAYASHI AND T. OZAWA, *On the derivative nonlinear Schrödinger equation*, Phys. D, 55 (1992), pp. 14–36.

[5] ———, *Remarks on nonlinear Schrödinger equations in one space dimension*, Differential Integral Equations, 7 (1994), pp. 453–461.

[6] N. HAYASHI AND Y. TSUTSUMI, *Remarks on the scattering problem for nonlinear Schrödinger equations*, in Lecture Notes in Math. 1285, Springer-Verlag, Berlin, 1987, pp. 162–168.

[7] T. KATO, *On nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré, Phys. Théor., 46 (1987), pp. 113–129.

[8] ———, *Nonlinear Schrödinger equations*, in Schrödinger Operators, Lecture Notes in Phys. 345, Springer-Verlag, Berlin, 1989.

[9] M. TSUTSUMI AND I. FUKUDA, *On solutions of the derivative nonlinear Schrödinger equation* II, Funkcial. Ekvac., 24 (1981), pp. 85–94.

[10] Y. TSUTSUMI, *L^2-solutions for nonlinear Schrödinger equations and nonlinear groups*, Funkcial. Ekvac., 30 (1987), pp. 115–125.

[11] ———, *Global strong solutions for nonlinear Schrödinger equations*, Nonlinear Anal., 11 (1987), pp. 1143–1154.

[12] T. SHAOBIN AND Z. LINGHAI, *On weak solutions of the mixed nonlinear Schrödinger equations*, IAPCM, pp. 92–03, preprint.

[13] N. HAYASHI AND T. OZAWA, *Modified wave operators for the derivative nonlinear Schrödinger equation*, Math. Ann., 298 (1994), pp. 557–576.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1504–1520, November 1994

004

PHASELOCKING IN A REACTION-DIFFUSION EQUATION WITH TWIST *

G. BARD ERMENTROUT† AND W. C. TROY†

Abstract. A generic reaction-diffusion equation near a Hopf bifurcation is analyzed by a two-dimensional topological shooting argument. Previous results of the authors are extended to the case where there is amplitude modulation of the frequency. The results are compared to a realistic chemical model.

Key words. oscillators, reaction-diffusion, phaselocking

AMS subject classifications. 92C20, 92C05, 34B15

1. Introduction.

The behavior of coupled nonlinear oscillators has been the subject of numerous theoretical and experimental studies [1]–[6]. The majority of these papers concern finitely many discretely coupled differential equations [1], [2], [4], [5]. The general method of analysis is to introduce some type of small parameter and then use the method of averaging to reduce the equations to a flow on the torus [7]. Alternative analyses exploit the special form of the oscillators near a Hopf bifurcation [8] or in the weakly nonlinear limit [9]. Little theoretical or numerical consideration has been given to continuous (as opposed to discrete) oscillatory media (see, e.g., [5] and [10]).

In two previous papers, we have analyzed continuous diffusion models of coupled oscillators for a special class of reaction-diffusion equations [11], [12]. These papers are motivated by some recent experimental work by Tam and Swinney et al., in which an oscillatory chemical reaction with a one-dimensional spatial gradient in some parameter is allowed to react and diffuse in a one-dimensional spatial domain. In this paper we continue this analysis for a system of more generic reaction-diffusion models. Consider the general system:

$$\begin{aligned} v_t &= (D(\sigma x)v_x)_x + F(\sigma x, v), \qquad 0 < x < 1, \\ v_x(0,t) &= v_x(1,t) = 0. \end{aligned} \tag{1.1}$$

We assume that the equation $v_t = F(\sigma x, v)$ has an asymptotically stable periodic solution for each x in the interval [0,1], but the period and shape of these oscillations is allowed to vary continuously in space. That is, we allow some spatial variation (parametrized by σ) in the one-dimensional medium. Tam and Swinney [13] consider both an experimental system and a model reaction and numerically study the behavior. In a companion paper [14], Vastano, Russo, and Swinney describe some plausible mechanisms for the interactions. In particular, for a large set of experimental and numerical parameters, they find *phaselocked* solutions in which the entire medium oscillates with the same period. Our goal is to determine conditions that guarantee the existence of phaselocked solutions to equations such as (1.1). This is generally an

*Received by the editors October 26, 1992; accepted for publication (in revised form) July 15, 1993. This work was supported in part by National Science Foundation grants DMS9303706 and DMS9002028.

†Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

impossible task, but under some circumstances formal and rigorous conditions can be obtained.

We will study the following system of reaction-diffusion equations:

$$\begin{aligned} v_{1t} &= (v_1 - qv_2)(1 - r^2) - (1 + \sigma x)v_2 + dv_{1xx}, \\ v_{2t} &= (v_2 + qv_1)(1 - r^2) + (1 + \sigma x)v_1 + dv_{2xx}, \end{aligned} \tag{1.2}$$

where $r^2 = v_1^2 + v_2^2$, and subscripts x and t are partial derivatives. This class of equations arises formally from a general reaction-diffusion equation whose dynamics occur near a Hopf bifurcation and where each reactant has the same diffusion coefficient. This latter assumption on the diffusion coefficients is reasonable for chemical reactions where all of the species are of similar size. In particular, this is the case for the commonly studied Field–Noyes equations [13], [14]. The assumption that the dynamics lie near a Hopf bifurcation is not so easily justified; numerical simulations away from the Hopf bifurcation on models such as the Brusselator (see §4) provide some evidence that the behavior is similar.

If the diffusivity D and the spatial gradient are small compared to the rate of attractions to the limit cycle, then perturbation methods can be used to derive a set of slowly varying phase equations. Neu [15] uses these methods to show that if the spatial variation and the diffusivity are of the same order of magnitude, then (1.1) reduces to a Burgers-type equation. More recently, Ermentrout [16] shows that if the diffusion is $O(\epsilon)$ and the spatial variation is $O(\epsilon^\nu)$ with $0 < \nu < 1$, then locking always occurs and (1.1) becomes a singularly perturbed nonlinear boundary-value problem. For a particularly simple model system, it is also shown that $O(1)$ variations can lead to loss of locking when the diffusion is small.

However, if the attraction to the oscillator and the spatial variations and diffusion are all of the same order, (1.1) becomes much more difficult. In [11] we consider a special case of (1.2), where $q = 0$. This form of (1.2) is amenable to rigorous analysis since it reduces to a third-order differential equation that is analyzed with a shooting argument. The appearance of the generically occuring parameter q, which is often called the "twist" of the oscillator, makes the analysis considerably more difficult. A formal perturbation of (1.1) near a Hopf bifurcation together with the assumption that $D(\sigma x) \equiv D$ and the spatial variation σ are both small leads to an equation of the form (1.2) with the addition of the "twist" term. (This type of calculation is performed in [8] for a pair of discrete diffusively coupled oscillators.) "Twist" plays a role in breaking the symmetry of coupled oscillators and, if sufficient, can even destroy the stability of a synchronous solutions to identical oscillators (see [8]). Kuramoto [3] has noted that in reaction-diffusion equations, this term leads to spatio-temporal chaos if it is too large. We have recently shown that the presence of twist in a reaction-diffusion equation on a disk can lead to rotating spiral waves (Paullet, Ermentrout, and Troy [17]); with $q = 0$ the spirals have straight arms.

If we convert (1.2) to polar coordinates using $v_1 = r\cos\theta$ and $v_2 = r\sin\theta$, we obtain the following problem:

$$\begin{aligned} r_t &= r(1 - r^2) + d(r_{xx} - r\theta_x^2), \\ \theta_t &= 1 + \sigma x + q(1 - r^2) + d\left(\frac{2r_x\theta_x}{r} + \theta_{xx}\right), \\ r_x(0,t) &= r_x(1,t) = 0, \qquad \theta_x(0,t) = \theta_x(1,t) = 0. \end{aligned} \tag{1.3}$$

The boundary conditions are equivalent to Neumann conditions for (1.2) on the unit interval. We seek time-periodic solutions to (1.3), which have the form $\theta(x,t) =$

$(\Omega+1)t+\int_0^x \phi(s)\,ds$ and $r(x,t)=\rho(x)$. Then, (1.3) becomes the third-order boundary value problem:

$$
\begin{aligned}
0 =& \rho(1-\rho^2)+d(\rho''-\rho\phi^2),\\
\Omega =& \sigma x+q(1-\rho^2)+d\left(\frac{2\rho'\phi}{\rho}+\phi'\right),\\
0 =& \rho'(0)=\rho'(1)=\phi(0)=\phi(1).
\end{aligned}
\tag{1.4}
$$

We note that the period of the solution is $T=2\pi/(\Omega+1)$, independent of $x\in[0,1]$. Thus, a solution of (1.4) is a periodic phaselocked solution of (1.3).

When $q=0$, (1.4) is identical to the problem solved in [11] by a shooting argument. The techniques in [11] can not be extended to the present case because when $q\neq 0$, reflection symmetry is lost. In [11], we were able to exploit this symmetry and thus show that $\Omega=\sigma/2$. Then, the resulting problem reduces to a one-parameter shooting argument. Here, we cannot determine Ω explicitly and must leave it as a free parameter. Furthermore, we must solve (1.4) over the entire domain $0\le x\le 1$ rather than up to $x=1/2$.

If we set $\sigma=0$ in (1.4), then the solution is trivially given as

$$\rho(x)=1,\quad \phi(x)=0,\qquad \Omega=0.$$

This is an asymptotically stable periodic solution of period 2π. Thus, one can use this as a starting point and explore the behavior of (1.4) for σ small. In §2, we use a regular perturbation method to find the qualitative form of solutions to (1.4) for σ small. Since σ is small, and the perturbation is regular, ρ stays close to 1. Large variations in σ are required to reduce the magnitude, $\rho(x)$.

Section 3, which contains the bulk of the mathematics of this paper, includes the shooting argument for σ "large." By "large," we mean that ρ stays bounded away from 1 (in fact, $0<\rho(0)<1/2$). The proof of existence depends on a two-parameter shooting method along with a topological theorem of McLeod and Serrin [18].

In the last section, we solve (1.4) numerically for a variety of values of q. We compare these solutions to the solutions of the full partial-differential equations; our numerical results indicate that the constructed solutions are asymptotically stable. We also compute the full "bifurcation" picture for a fixed value of d as σ increases. We show the phenomenon of oscillator death; i.e., the stabilization of the trivial state, $v_1=v_2=0$. The stabilization was proven in an earlier paper [12]. Finally, we provide some numerical evidence of a similar phenomenon occurring for the Brusselator model of chemical reactions.

2. Perturbation for small σ. For small σ we can solve (1.4) by a regular perturbation series. This allows us to see the qualitative behavior of the amplitude and the phase. In particular, it is seen that the magnitude of the oscillation decreases, with the minimum in the middle, and that the phase gradient ϕ is close to quadratic. As the numerics in §4 attest, this qualitative picture is preserved throughout the entire range of values of the parameters.

We seek solutions to (1.4) of the form

$$\phi(x,\sigma)=\sigma\phi_1(x)+\sigma^2\phi_2(x)+O(\sigma^3), \tag{2.1}$$

$$\rho(x,\sigma)=1+\sigma\rho_1(x)+\sigma^2\rho_2(x)+O(\sigma^3), \tag{2.2}$$

$$\Omega(\sigma) = \sigma\Omega_1 + \sigma^2\Omega_2 + O(\sigma^3), \tag{2.3}$$

We note that Ω is an unknown function of σ and is determined by satisfying the boundary conditions. Substitution of (2.1)–(2.3) into (1.4) yields

$$-2\rho_1 + d\rho_1'' \equiv L\rho_1 = 0.$$

The boundary conditions imply that $\rho_1 = 0$, so we must go to higher order to see the effect of σ on ρ. The equation for ϕ_1 satisfies

$$d\phi_1' = \Omega_1 - x.$$

This, along with the boundary conditions, yields

$$\Omega_1 = 1/2, \qquad \phi_1 = x(1-x)/(2d). \tag{2.4}$$

Thus, the locked frequency, Ω, is close to the mean frequency, $\sigma/2$. Next, we find that ρ_2 satisfies

$$L\rho_2 = x^2(1-x)^2/(4d),$$

which has a solution:

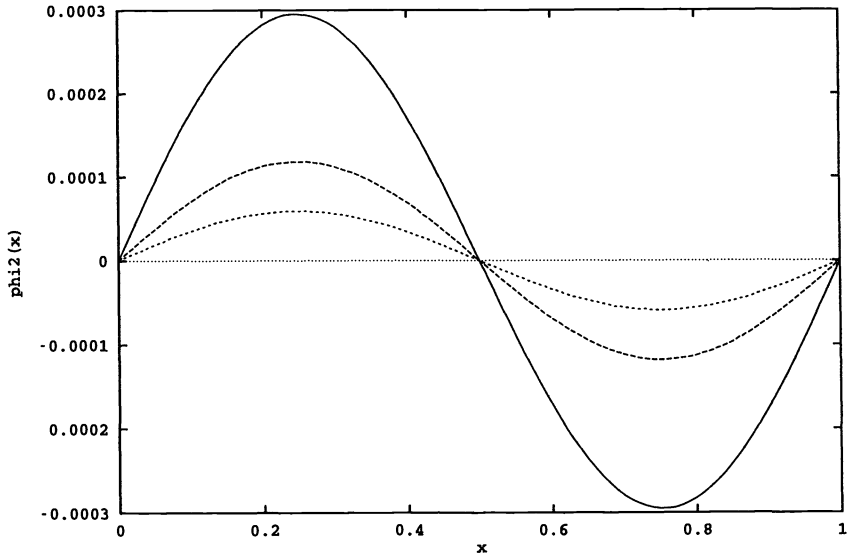
$$\rho_2(x) = \frac{x^3}{4\,d} - \frac{x^2}{8\,d} - \frac{x^4}{8\,d} - \frac{3\,d}{4} - \frac{1}{8} + \frac{3x}{4} - \frac{3x^2}{4} + \frac{3\sqrt{2}\sqrt{d}\cosh(\frac{\sqrt{2}(x-1/2)}{\sqrt{d}})}{8\sinh(\frac{\sqrt{2}}{2\sqrt{d}})}. \tag{2.5}$$

This can be used to obtain the solution to ϕ_2, which satisfies

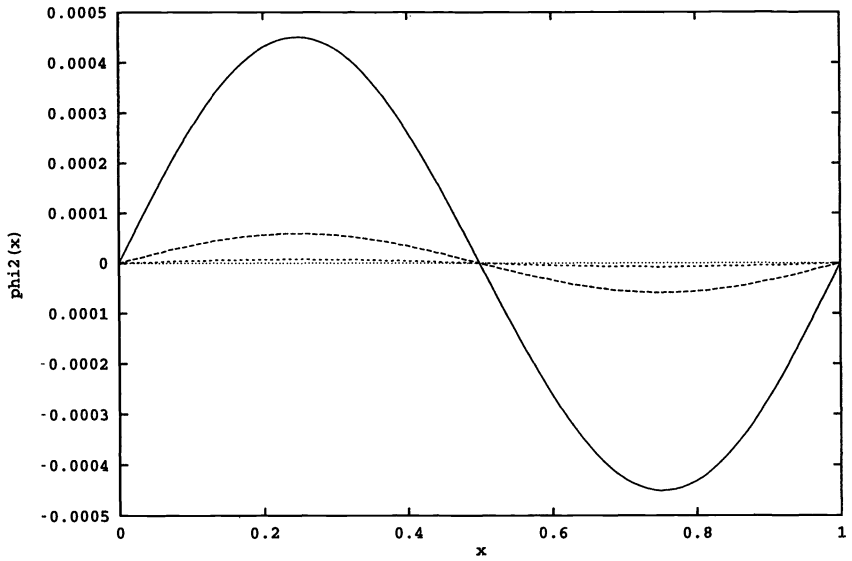
$$d\phi_2' = \Omega_2 + 2q\rho_2(x). \tag{2.6}$$

We do not explicitly give the value here; rather, we plot the results of the perturbation. The details are tedious and were symbolically solved with MAPLE.

There are several interesting results of the calculations. First, the role of q is to make the phase gradient ϕ asymmetric about the origin. This is seen by plotting $\phi_2(x)$, shown in Fig. 2.1, for various values of q and d. When this function is added to the $O(\sigma)$ part of ϕ the result is a slightly skewed phase gradient. Even for fairly large q or small d the effects of twist are very small and probably would not be seen numerically (see §4.) The phaselocked frequency is also skewed from the mean by an amount related to q and is found to be $\Omega_2 = q/120d$. $q > 0$ tends to lower the ensemble frequency and $q < 0$ increases it. The form of (2.6) shows that the frequency depends on q in a linear fashion; using MAPLE, one finds that the slope tends to zero for large d and tends to infinity for small d. Thus, we cannot expect this expansion to be good if d is very small; indeed, the case of small d is the subject of another paper [16]. Finally, in Fig. 2.2, we show the $O(\sigma^2)$ deviation of the amplitude $\rho_2(x)$ for various values of d. It is clear that for d close to 1, $\rho_2(x)$ is small, so that very large gradients in σ are required to effectively move $\rho(x)$ away from 1. One can use MAPLE to show that $\rho_2(x)$ is strictly negative in $[0,1]$ so that the effect of increasing σ is to always reduce the magnitude of $\rho(x)$. The next order expansion of $\rho(x)$ will introduce the skewing due to q, so that the apparent symmetry of $\rho(x)$ about the center is a consequence of our truncation. We finally note that each term of the series will involve only terms of the form x^l and $\exp(ax)$, thus we can find $\Omega_j, \rho_j(x)$, and $\phi_j(x)$ at each step.



(a)



(b)

FIG. 2.1. *$\phi_2(x)$ for q and d varying.* (a) *$d = 1$ is fixed and $q = 1, 2, 5$. Larger magnitude curves correspond to larger q.* (b) *$q = 1$ and $d = .5, 1, 2$. Larger magnitude curves correspond to smaller values of d.*

In the next section, we let σ get larger and show that the magnitude of the oscillations is bounded away from 1. Thus, we can "paste" together these two regimes to construct a full picture of the structure as σ increases. This is what is done in the last section.

3. Existence of solutions to (1.4). Following our earlier work, we introduce a new variable u defined as the logarithmic derivative of ρ, $u = \rho_x/\rho$. Furthermore, we

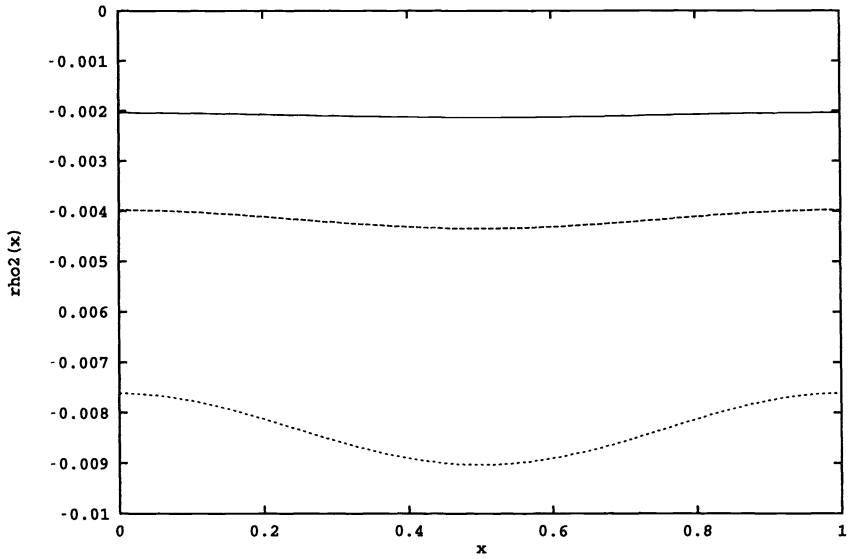


FIG. 2.2. *The amplitude correction, $\rho_2(x)$, for $d = .5, 1, 2$. More negative values correspond to larger values of d.*

note that if (ρ, u, ϕ) satisfies (1.4), then so does $(\rho, u, -\phi)$ for $-\sigma, -q$, and $-\Omega$. Thus, for simplicity, we find it convenient to replace ϕ with $-\phi$ and set $\Omega = b\sigma$, so that (1.4) becomes the three-dimensional boundary value problem:

$$\rho' = \rho u, \tag{3.1}$$

$$u' = (\rho^2 - 1)/d + \phi^2 - u^2, \tag{3.2}$$

$$\phi' = \sigma(x - b)/d - 2u\phi - q(\rho^2 - 1)/d, \tag{3.3}$$

where

$$0 < \rho(0) < 1, \qquad u(0) = \phi(0) = 0, \tag{3.4}$$

$$u(1) = \phi(1) = 0. \tag{3.5}$$

We assume that $d \geq 1$ is fixed. Furthermore, we assume that $\sigma > 0$ without loss of generality (for otherwise, we could reflect the medium about $x = 1/2$), $0 < b < 1$, and $|q| > 0$. We note that $q = 0$ has already been analyzed in [11]. We prove the following theorem.

THEOREM 3.1. *Let $q \in [-1/8, 0) \cup (0, 1/8]$ and $0 < \rho(0) < 1/5$. Then, there exist values $\bar{\sigma} > 0$, $\bar{b} \in (0, 1)$ such that the solution to* (3.1)–(3.4) *satisfies the boundary condition* (3.5).

Remark. One should note that the theorem states something that is slightly different from our objective. The two shooting parameters used are b and σ, rather than the ideal choices of Ω and $\rho(0)$. However, as is common in many nonlinear problems, the variables that one wishes to solve for are difficult, and a different parametrization is necessary. Thus, given a $\rho(0)$ we obtain a b and a σ solving (3.1)–(3.5). Clearly,

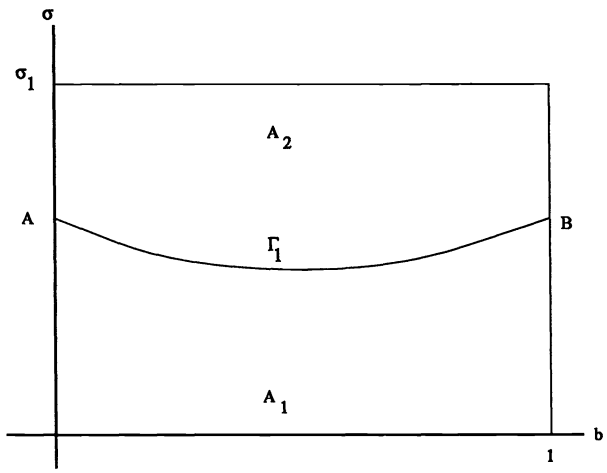


FIG. 3.1. *The shooting regime, σ vs b. At A, $\phi(1) > 0$ and at B, $\phi(1) \leq 0$.*

this yields Ω and, for the given $\rho(0)$, some value of σ. Since $\rho(0)$ is small, the value of σ will *not* be small (cf §2.) In order to get a complete picture of the two regimes, we resort to numerical solutions in §4.

Outline of proof. We employ a two-dimensional shooting argument. The parameters which are free to be varied are b and σ where $0 \leq b \leq 1$ and $\sigma \geq 0$. In Lemma 3.1 we show that $u(1) < 0$ along the line segment $\sigma = 0$, $0 \leq b \leq 1$. Lemma 3.2 proves that $u(1) > 0$ along a line segment $\sigma = \sigma_1 > 0$, $0 \leq b \leq 1$. We then refer to a topological argument of McLeod and Serrin [18] to conclude that there is a continuum Γ_1 joining the half-lines $b = 0$, $\sigma > 0$, and $b = 1$, $\sigma > 0$, and such that $u(1) = 0$ along Γ_1 (see Fig. 3.1.) Next, we turn our attention to the behavior of ϕ. We start with a point $(1, \sigma^*) \in \Gamma_1$ and show that $\phi(1) \leq 0$ (Lemma 3.6). Lemma 3.5 shows that $\phi(1) > 0$ on $\Gamma_1 \cap \{b = 0\}$. Also, standard theory shows that $\phi(1)$ is continuously dependent on $(b, \sigma) \in \Gamma_1$. Thus, since Γ_1 is a continuum, $\phi(1)$ must vanish at some point $(\bar{b}, \bar{\sigma}) \in \Gamma_1$.

Proof of Theorem 3.1. For notational simplicity and ease of exposition, we prove Theorem 3.1 for the case $d = 1$. The details for $d > 1$ are similar. For each $\sigma > 0$ and $\rho(0) \in [0, 1/2]$ we let $\tilde{x} \equiv \max\{\hat{x} \in (0,1) |$ the solution of (3.1)–(3.4) exists for all $x \in [0, \hat{x})\}$.

Our subsequent analysis requires that we obtain a useful lower bound on $u(x)$ for $0 \leq x \leq \tilde{x}$. From (3.2) we conclude that

$$u' \geq -(1 + u^2), \tag{3.6}$$

hence

$$u \geq -\tan(x) \geq -\tan(1) \quad \text{for } 0 \leq x < \tilde{x}. \tag{3.7}$$

In the first step of our shooting argument, we analyze the behavior of both u and ϕ for the special case $\sigma = 0$.

LEMMA 3.1. *Let $\sigma = 0$. If $0 < q \leq 1/8$ then $u(x) < 0$ and $\phi(x) > 0$ for all $x \in (0, 1]$. If $-1/8 \leq q < 0$ then $u(x) < 0$ and $\phi(x) < 0$ for all $x \in (0, 1]$.*

Proof. An integration of (3.3) gives

$$\phi(x) = q \int_0^x (1 - \rho^2) e^{-2\int_t^x u(s)ds} dt. \tag{3.8}$$

We distinguish two cases.

(i) $0 < q \leq 1/4$. Then (3.8) implies that $\phi > 0$ for $x \in (0, \tilde{x})$ as long as $\rho \in (0,1)$. Therefore, as long as $\phi^2 \leq 3/4$, then (3.1) and (3.2) show that $u' < 0, -\tan(x) < u < 0$, and so $0 < \rho < 1/2$. To prove that $\phi^2 \leq 3/4$ while $-\tan(x) \leq u \leq 0$, we combine (3.7) and (3.8) and obtain

$$\phi(x) \leq \frac{q}{\cos^2(x)} \int_0^1 \cos^2(t)dt < \sqrt{(.75)} \quad \text{for} \quad x \in [0, \tilde{x}).$$

We conclude that $\tilde{x} = 1$ since ϕ and u are bounded by $0 \leq \phi^2 \leq 3/4$, and $u < 0$ on $(0, \tilde{x})$. Also, $u(1) < 0$ since $u' < 0$ on $(0,1]$.

Next, we consider the second case.

(ii) $-1/4 \leq q < 0$. Then (3.8) implies that $\phi < 0$ for $x \in (0,1)$ as long as $0 < \rho < 1$. As in case (i) above it follows that $u' < 0, u < 0$, and $0 < \rho < 1/2$ as long as $\phi^2 \leq 3/4$. Since $-\tan(x) < u < 0$ while $\phi^2 \leq 3/4$, it follows from (3.8) that

$$0 \geq \phi(x) \geq \frac{q}{\cos^2(x)} \int_0^1 \cos^2(t)dt > -(.75)^{1/2} \quad \text{for} \quad 0 \leq x \leq 1,$$

and the proof is complete.

Next, we determine the behavior of u for σ large and $0 \leq b \leq 1$.

LEMMA 3.2. *There exists $\sigma_1 > 0$ such that if $\sigma \geq \sigma_1, 0 \leq b \leq 1$, and $|q| \leq 1/8$, then $u(\hat{x}) > \tan(1)$ for some $\hat{x} \in (0,1)$. Furthermore, $u(x) > 0$ for $\hat{x} \leq x < \tilde{x}$ and $\lim_{x \to \tilde{x}} u(x) > 0$.*

Proof. We assume that the lemma is false and obtain a contradiction. Thus, we assume that there exists a sequence $\{(\sigma_i, b_i)\}$ with $0 \leq b_i \leq 1$ for all $i, \lim_{i \to \infty} \sigma_i = \infty$ and such that for each $i, |u(x)| \leq \tan(1)$ for $x \in (0,1)$. Here $\tilde{x} = 1$ since u is bounded. Since $[0,1]$ is compact there exists $b \in [0,1]$ and a subsequence $\{b_{i_k}\}$ such that $i_k \longrightarrow \infty$ and $b_{i_k} \longrightarrow b$ as $k \longrightarrow \infty$. Thus we may set $b_i \cong b$, drop the dependence of σ_i on i, and consider large σ. We first consider the special case $b = 0$. Then (3.3) reduces to

$$\phi' = \sigma x - 2u\phi - q(\rho^2 - 1). \tag{3.9}$$

Since it is assumed that $u(x) \leq \tan(1)$ on $(0, \tilde{x})$, it follows from (3.1) and (3.7) that ρ is uniformly bounded on $(0, \tilde{x})$. Thus there is an $M > 0$, independent of σ, such that (3.9) reduces to

$$\phi' \geq \sigma x - M - 2u\phi. \tag{3.10}$$

Integrating (3.10), we obtain

$$\phi(x) \geq \int_0^x (\sigma t - M) e^{-2\int_t^x u(s)ds} dt. \tag{3.11}$$

If $0 \leq x \leq M/\sigma$, then (3.11) reduces to

$$\phi(x) \geq \int_0^x (\sigma t - M) e^{2\tan(1)(x-t)} dt \geq \left(\frac{\sigma x^2}{2} - Mx \right) e^{2\tan(1)}, \tag{3.12}$$

since $u > -\tan(1)$ on $(0, \tilde{x})$. In particular, from (3.12) we obtain

$$\phi(x) \geq -\frac{M^2 e^{2\tan(1)}}{2\sigma}, \qquad 0 \leq x \leq \frac{M}{\sigma}. \tag{3.13}$$

If $x > M/\sigma$ then

$$\begin{aligned}\phi(x) &\geq \int_0^{M/\sigma} (\sigma t - M) e^{-2\int_t^x u ds} dt + \int_{M/\sigma}^x (\sigma t - M) e^{-\int_t^x 2u ds} dt \\ &\geq -\frac{M^2 e^{2\tan(1)}}{2\sigma} + e^{-2\tan(1)} \left(\frac{\sigma x^2}{2} - Mx + \frac{M^2}{2\sigma} \right).\end{aligned} \tag{3.14}$$

Thus, for large $\sigma > 0$, we conclude from (3.14) that

$$\phi(x) \geq \frac{e^{-2\tan(1)} \sigma x^2}{4} \quad \text{for } x \in [1/2, 1]. \tag{3.15}$$

Substituting (3.15) into (3.2), we obtain

$$u' > -2\tan^2(1) + \frac{e^{-4\tan(1)} \sigma^2 x^4}{16} \quad \text{for } x \in [1/2, 1]. \tag{3.16}$$

An integration of (3.16) shows that $u(\hat{x}) > \tan(1)$ for some $\hat{x} \in (1/2, 1)$ and σ large, a contradiction of our earlier assumption.

A repetition of these steps shows that there is a $\hat{b} \in (0, 1)$ and a value $\hat{\sigma} > 0$ such that if $\sigma \geq \hat{\sigma}$ and $0 \leq b \leq \hat{b}$ then $u(\hat{x}) > \tan(1)$ for some $\hat{x} \in (0, 1)$. Next, let $b \in [\hat{b}, 1]$. Again, there exists an $N > 0$ such that (3.3) reduces to

$$\phi' \leq \sigma(x - \hat{b}) - 2u\phi + N \tag{3.17}$$

as long as $|u| \leq \tan(1)$. Under the assumption that $|u| \leq \tan(1)$ on $(0, \hat{b})$, we integrate (3.17) and find that $-\phi(x)$ is large and positive over $[\frac{\hat{b}}{2}, \frac{3\hat{b}}{4}]$. Again, an integration of (3.2) shows that $u(\hat{x}) > \tan(1)$ at some $\hat{x} \in (\frac{\hat{b}}{2}, \frac{3\hat{b}}{4})$. Finally, once we have $u(\hat{x}) > \tan(1)$ at some $\hat{x} \in (0, 1)$ we need to show that $u > 0$ for $x \in (\hat{x}, 1]$ as long as the solution exists. For this we return to (3.2) and see that $u' > -(1 + u^2)$. An integration of this inequality from $\hat{x}$ to x gives the result.

In order to complete the proof of Theorem 3.1 we will employ the following topological result.

THEOREM (McLeod and Serrin [16]). *Let I be the closed unit square $\{0 \leq x \leq 1, 0 \leq y \leq 1\}$ in the (x, y) plane, and let A_1 and A_2 be disjoint relatively open sets of I containing, respectively, the lines $y = 0$ and $y = 1$. Then the complement D of A_1 and A_2 in I contains a continuum, Γ_1, joining the lines $x = 0$ and $x = 1$.*

Remark. The McLeod–Serrin theorem applies equally well to any rectangle.

We are now prepared to define our first two shooting sets. We restrict (b, σ) to lie in the rectangle $S = \{(b, \sigma) | 0 \leq b \leq 1, 0 \leq \sigma \leq \sigma_1\}$, where σ_1 satisfies Lemma 3.2. Then,

$$A_1 = \{(b, \sigma) \in S | \tilde{x} = 1 \text{ and } u(1) < 0\}$$

and

$$A_2 = \{(b, \sigma) \in S | \text{ either } \tilde{x} < 1 \text{ or else } u(1) > 0\}.$$

LEMMA 3.3. *Let $q \in [-1/4, 0) \cup (0, 1/4]$. Then A_1 and A_2 are relatively open, nonempty subsets of S.*

Proof. Continuity forces both A_1 and A_2 to be relatively open sets. Lemmas 3.1 and 3.2 show that $A_1 \neq \phi$ and $A_2 \neq \phi$.

A_1 contains the line segment $\sigma = 0, 0 \leq b \leq 1$. A_2 contains the line segment $\sigma = \sigma_1, 0 \leq b \leq 1$.

The McLeod–Serrin theorem immediately guarantees the existence of a continuum $\Gamma_1 \subset S$ which joins the lines $b = 0$ and $b = 1$, and has the further property that $\Gamma_1 \cap A_1 = \Gamma_1 \cap A_2 = \phi$. One final property of Γ_1 is given in the following lemma.

LEMMA 3.4. *Let* $q \in [-1/8, 0) \cup (0, 1/8]$. *If* $(b, \sigma) \in \Gamma_1$ *then the solution exists for all* $x \in [0, 1]$, *and* $u(1) = 0$.

Proof. We assume the lemma is false and obtain a contradiction. Thus, we suppose that there is a point $(b, \sigma) \in \Gamma_1$ such that the solution becomes unbounded at some $\tilde{x} \in (0, 1]$. Then, one of ρ, u, and ϕ must become unbounded as $x \longrightarrow \tilde{x}^-$. Suppose that u becomes unbounded at $\tilde{x}$. Since $u > -\tan(1)$ on $[0, \tilde{x})$ it must be the case that $u > \tan(1)$ at some $\hat{x} \in (0, \tilde{x})$. But then $(b, \sigma) \in A_2$—a contradiction, since $\Gamma_1 \cap A_2 = \phi$. Therefore $-\tan(1) \leq u \leq \tan(1)$ for all $x \in [0, \tilde{x})$. It then follows from (3.1) and (3.3) that ρ and ϕ remain bounded over $[0, \tilde{x})$. Thus the solution must exist and remain bounded over $[0, 1]$. Finally, suppose that $u(1) \neq 0$. If $u(1) > 0$ then $(b, \sigma) \in A_2$—a contradiction, since $\Gamma_1 \cap A_2 \neq \phi$. If $u(1) < 0$ then $(b, \sigma) \in A_1$—again a contradiction since $\Gamma_1 \cap A_1 = \phi$. We must conclude that $u(1) = 0$, and the lemma is proved.

Having shown the existence of the continuum Γ_1 on which $u(1) = 0$, we have now completed the first half of our proof of Theorem 3.1. It remains to be proved that there exists a point $(\bar{b}, \bar{\sigma}) \in \Gamma_1$ for which $\phi(1) = 0$. We shall need the following result.

LEMMA 3.5. *Let* $b = 0$ *and* $q \in [-1/8, 0) \cup (0, 1/8]$. *If* $u(1) = 0$ *then* $\phi(1) > 0$.

Proof. The first case to be considered is that $0 < q \leq 1/8$. Again, we recall that if $u(1) = 0$, then it must be the case that

$$-\tan(1) < u(x) < \tan(1) \quad \text{for } 0 \leq x \leq 1. \tag{3.18}$$

An integration of (3.1) shows that

$$0 < \rho(x) \leq .95 \tag{3.19}$$

for $0 \leq x \leq 1$. It follows from (3.19) and (3.3) that

$$\phi' \geq \sigma x - 2u\phi \quad \text{for } 0 \leq x \leq 1. \tag{3.20}$$

An integration of (3.20) gives

$$\phi(x) \geq \sigma \int_0^x t e^{-2\int_t^x u(s)ds} dt > 0 \quad \text{for } 0 \leq x \leq 1.$$

In particular, $\phi(1) > 0$. Next, we consider the second case, that $-1/8 \leq q < 0$. We assume, for the sake of contradiction, that $\phi(1) \leq 0$. Since $u(0) = u(1) = 0$ and $u'(0) < 0$ there is a first $\bar{x} \in (0, 1)$ such that $u'(\bar{x}) = 0$. Then $0 < \rho(x) < \rho(0)$ on $(0, \bar{x}]$, so that (3.2) implies $\phi^2(\bar{x}) \geq 3/4$. Suppose that $\phi(\bar{x}) \geq \sqrt{3/4}$. Our assumption that $\phi(1) \leq 0$ implies that $\phi = 0$ at some first $x^* \in (\bar{x}, 1]$. It follows from (3.3) that

$$\phi' \geq q - 2u\phi. \tag{3.21}$$

An integration of (3.21), together with (3.18), gives

$$\begin{aligned} \phi(x) e^{2\int_{\bar{x}}^x u(s)ds} &\geq (.75)^{1/2} + q \int_{\bar{x}}^x e^{2\int_{\bar{x}}^t u(s)ds} dt \\ &\geq (.75)^{1/2} + 6.91q \\ &> 0 \end{aligned}$$

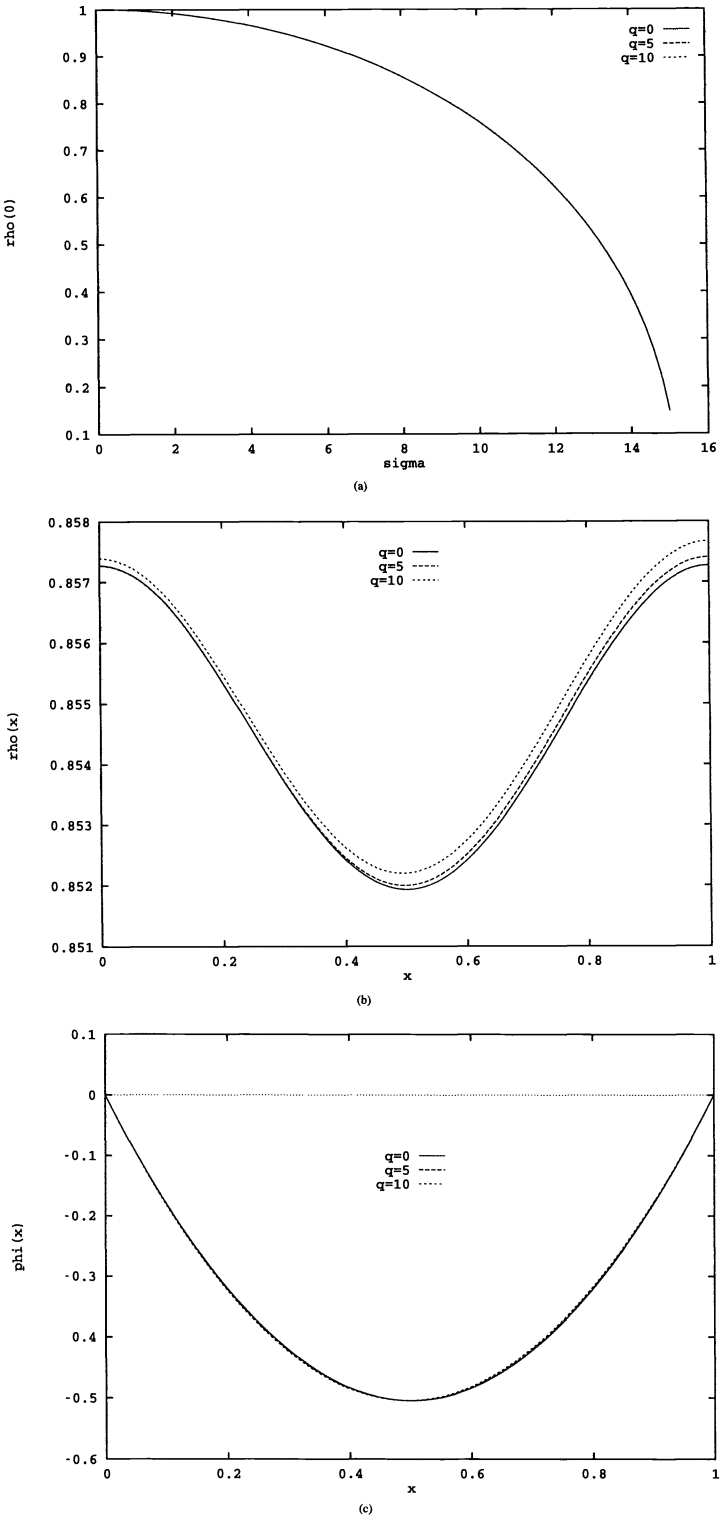


FIG. 4.1. *Solutions to* (3.1)–(3.5) *via a numerical shooting technique.* (a) *The magnitude at the origin,* $\rho(0)$, *as* σ *increases for* $q = 0, 5, 10$; (b) *magnitude versus* x *for* $\sigma = 8$ *and* $q = 0, 5, 10$; (c) *phase gradient versus* x *for same parameters as* (b).

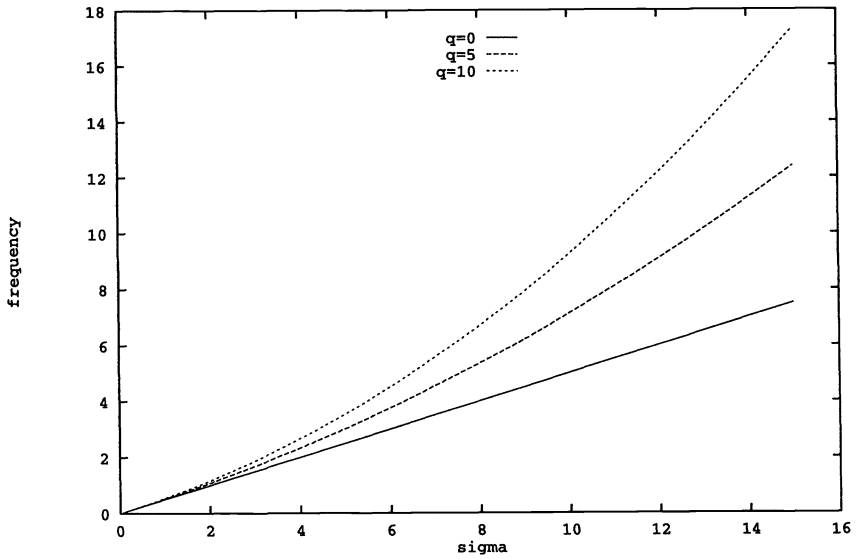


FIG. 4.2. *The ensemble frequency as a function of* σ *for* $q = 0, 5, 10$.

for $\bar{x} \leq x \leq 1$. Thus we have obtained a contradiction, at $x = 1$, to our assumption that $\phi(1) \leq 0$. It remains to consider the possibility that $\phi(\bar{x}) \leq -\sqrt{(.75)}$. For this case, we use (3.7), integrate (3.21), and obtain

$$\phi \geq \frac{q}{\cos^2(x)} > -\sqrt{(.75)} \quad \text{for } 0 \leq x \leq 1.$$

Thus $\phi(\bar{x}) > -\sqrt{(.75)}$, and we have arrived at a contradiction. This completes the proof of the lemma.

The next step in our shooting argument is to determine the behavior of solutions on the half-line $b = 1, \sigma > 0$, which we call L.

LEMMA 3.6. *Let* $q \in [-1/8, 0) \cup (0, 1/8]$. *If* $(1, \sigma^*) \in \Gamma_1 \cap L$ *then* $\phi(1) \leq 0$.

Proof. First we consider the case $q \in [-1/8, 0)$.

As in Lemma 3.5, since $u(1) = 0$, it must be the case that (3.18) and (3.19) hold. It follows from (3.19) and (3.3) that $\phi' \leq -2u\phi$. An integration yields $\phi(1) \leq 0$. Next, suppose that $q \in (0, 1/8]$. Since $u(0) = u(1) = 0$ and $u'(0) < 0$, there is a first $\bar{x} \in (0, 1)$ such that $u'(\bar{x}) = 0$, and therefore $\phi^2(\bar{x}) \geq \sqrt{(.75)}$. Since $\sigma(x - 1) \leq 0$, (3.3) reduces to

$$\phi' + 2u\phi \leq 1/4 \quad \text{for } 0 \leq x \leq 1. \tag{3.22}$$

From (3.7) and (3.22) an integration leads to $\phi \leq .25/\cos^2(x) < \sqrt{(.75)}$. Therefore, $\phi(\bar{x}) \leq -\sqrt{(.75)}$. Again, (3.22) holds over $[\bar{x}, 1]$. Integrating (3.22) from $\bar{x}$ to 1 and using (3.18), we obtain $\phi(1)\exp(\int_{\bar{x}}^1 2u(s)ds) \leq .125 - \sqrt{(.75)} < 0$. Thus $\phi(1) < 0$ and the lemma is proved.

We are now prepared to proceed with the final details of the proof of Theorem 3.1. Because the set Γ_1 and the interval $0 \leq x \leq 1$ are compact, it follows from standard theory that $\phi(1)$ is continuously dependent on $(b, \sigma) \in \Gamma_1$. Therefore, since $\phi(1) \geq 0$ on $\Gamma_1 \cap \{b = 0\}$, and $\phi(1) \leq 0$ on $\Gamma_1 \cap \{b = 1\}$, we conclude that there must be a point $(\bar{b}, \bar{\sigma}) \in \Gamma_1$ at which $\phi(1) = 0$. This concludes the proof.

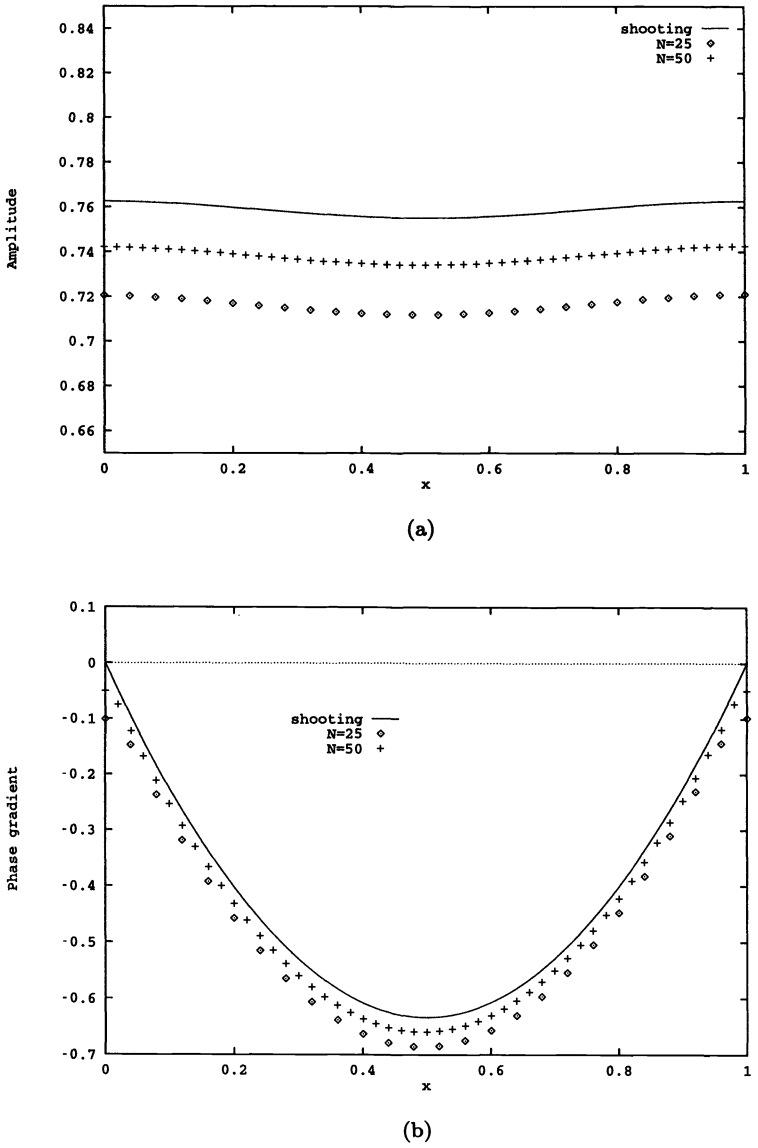


FIG. 4.3. *The solutions to the full partial differential equation compared to those obtained from the shooting argument. The solid line is the shooting curve. The remaining curves correspond to grids of* 50 *and* 25 *points, respectively.* ($q = 10, \sigma = 10, d = 2.$) (a) *The magnitude.* (b) *The phase gradient.*

4. Numerical solutions. In this section, we numerically investigate the solution to (1.1) and (1.2). We first solve the boundary value problem (3.1)–(3.5) and show the behavior of the solutions as a function of σ. Next, we compare these with the solutions obtained by solving the full partial differential equations (1.2). Finally, we solve (1.1) for a chemically realistic model and compare the results with the solution of (1.2). In particular, we show that the magnitude of the oscillators decays with steep

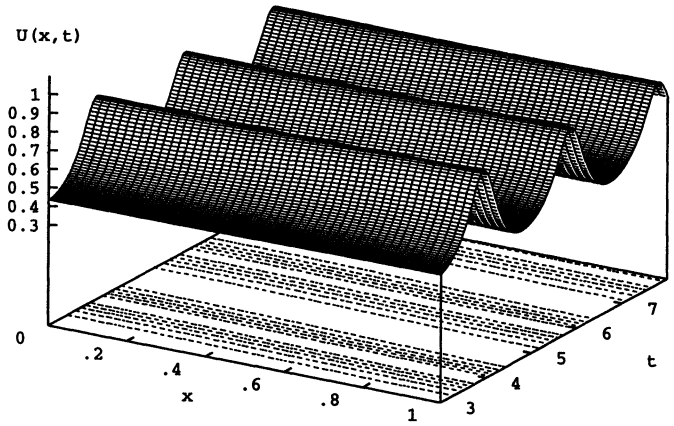


FIG. 4.4. *Space-time picture showing the first component of the Brusselator oscillator. The highest concentration is black and the lowest is white. Parameters are* $d = 1$, $a = .6$, $b = 1.5$, $\sigma = 9$. *The method of lines with Euler integration was used;* $dt = 0.0001$, $npts = 50$. *The initial values are on the homogeneous limit cycle. This figure depicts* $t = 3$ *to* $t = 8$ *in steps of* $t = 0.025$.

gradients in frequency.

We first numerically solve equations (3.1)–(3.5). There are two shooting parameters, the amplitude $\rho(0)$ and the bulk frequency Ω. We choose $\sigma = 0$ and thus obtain values $\rho(0) = 1, \Omega = 0$. Then, we can increase σ and follow this solution. We do it by integrating the initial value problem $\rho(0) = r_0, u(0) = 0, \phi(0) = 0$ to $x = 1$ to obtain values of $u(1) \equiv U(r_0, \Omega)$ and $\phi(1) \equiv \Phi(r_0, \Omega)$. If these are both zero, we are done. By integrating the variational equations obtained by differentiating equations (3.1)–(3.5) with respect to r_0 and Ω, respectively, we can numerically find the partial derivatives of the two functions $U(r_0, \Omega)$ and $\Phi(r_0, \Omega)$. Thus, we can use Newton's method to improve our guess of r_0, Ω. Since we start with a guess that is the final converged value for the previous value of σ, it typically takes only a few iterations to converge to a new value. If the convergence fails after 50 iterations, the program stops. Otherwise, we increase σ and continue the calculation. We use a fixed-stepsize Runge–Kutta integrator to solve the initial and variational problems.

In Fig. 4.1(a) we depict the magnitude of the solutions at $x = 0, r_0$ as σ increases for several values of q, d. There is virtually no difference between the curves as q changes; while the existence proof for $q \neq 0$ is far more difficult, the behavior is almost identical to the $q = 0$ case. It is clear that as σ increases, the magnitude of the oscillations decreases. Figures 4.1(b), (c) show the magnitude and phase gradient as a function of x for $\sigma = 8$ fixed and $q = 0, 5, 10$. The main consequence of the "twist" term is to induce a slight asymmetry on the amplitude. There is virtually no difference in the phase gradient. Numerical simulations later in this section and the results of [12] seem to imply that the trivial solution $\rho(x) \equiv 0$ is the only stable solution for $d >= 1$ and σ sufficiently large. Changing d has a more profound effect on the picture; with larger values of d the point at which the oscillation vanishes occurs at a larger value of σ. Again, citing our previous results [12], and noting that q appears only as a nonlinear effect, it is clear that this critical value of σ is independent of q.

The main effect of q is on the value of Ω, the ensemble frequency. Figure 4.2 shows Ω as a function of σ for q positive, negative, and zero. All curves pass through the origin; the $q = 0$ curve is given by $\Omega = \sigma/2$.

To ascertain if the solutions to the boundary value problem are stable solutions to the initial value problem (1.2), we integrate (1.2) starting with homogeneous initial

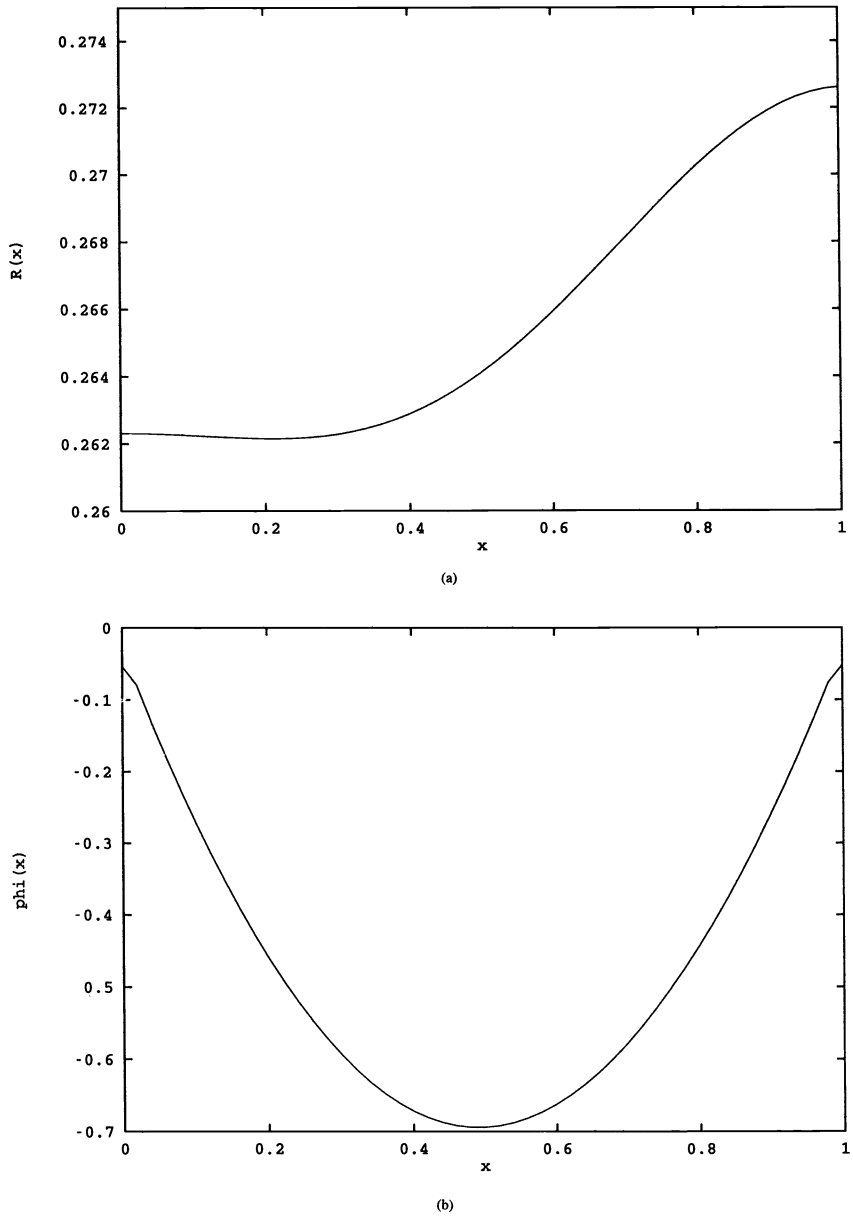


FIG. 4.5. *Magnitude and phase gradient from simulation in Fig.* 4.4. (a) *The average amplitude*; (b) *the average phase-gradient.*

conditions $v_1 = 1, v_2 = 0$. We discretize space into either 25 or 50 points, and use Euler integration with a small timestep to solve (1.2). In order to compare solutions to the boundary value problem, we compute $\rho(x) = \sqrt{v_1^2(x,t) + v_2^2(x,t)}$ at several times. Locked solutions will not vary. We also compute ϕ by numerically taking the spatial derivative and using the fact that $v_1 = \rho\cos\theta, v_2 = \rho\sin\theta$. In Fig. 4.3, we show ρ and ϕ as computed from (1.2) for 50 and 100 points, along with the solution to (3.1)–(3.5). There are differences which we believe to be primarily due to discretization error.

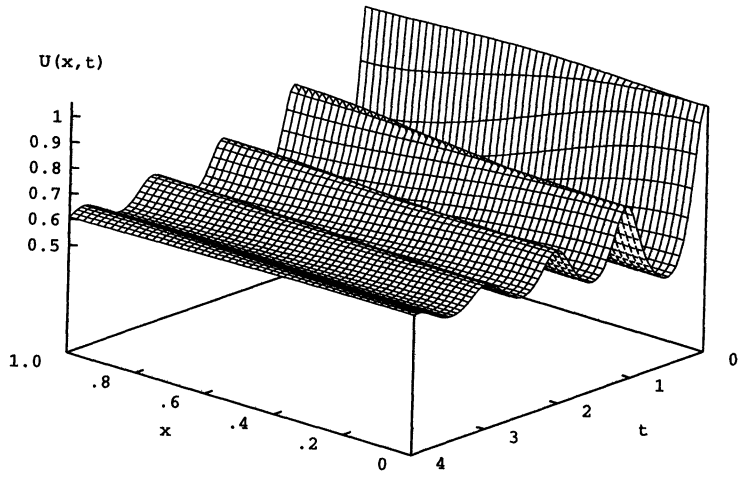


FIG. 4.6. *The first component of the Brusselator as a function of time and space when* $\sigma = 19$. *This shows the decay to rest. Numerical parameters as in Fig.* 4.4. *Time is from* $t = 0$ *to* $t = 5.0$ *in steps of* $t = .05$.

As a final example, we solve the following chemically motivated example (the well-studied Brussellator):

$$u_t = (1 + \sigma x)(a - (b+1)u + u^2 v) + du_{xx}, \tag{4.1}$$

$$v_t = (1 + \sigma x)(bu - u^2 v) + dv_{xx}, \tag{4.2}$$

with Neumann boundary conditions. We have artificially imposed a gradient in frequency by multiplying the kinetics by an x-dependent term. To compute the analogue of ρ and the phase gradient ϕ, we proceed as follows. We assume that at each spatial point, we can write

$$u(x,t) = \bar{u} + \rho(x,t)\cos(\theta(x,t)), \tag{4.3}$$

$$v(x,t) = \bar{v} + \rho(x,t)\cos(\theta(x,t)). \tag{4.4}$$

Here $(\bar{u}, \bar{v})$ is the equilibrium point of (4.1), (4.2), $(a, (b-1)/a)$. Thus, if $\rho = 0$ we are at equilibrium. To compute $\rho(x,t)$ we just calculate the root mean square deviation from the equilibrium. Differentiating (4.3) and (4.4) with respect to x, allows us to compute:

$$\theta_x(x,t) = \frac{(u(x,t) - \bar{u})v_x(x,t) - (v(x,t) - \bar{v})u_x(x,t)}{\rho^2(x,t)}.$$

Typically $\rho(x,t)$ and $\theta_x(x,t)$ will not be independent of time as in the example above, so we average these values over one complete cycle to get $r(x)$ and $\phi(x)$. In Fig. 4.4, we show the u component of the solution to (4.1), (4.2) in a space-time plot. Fig. 4.5 shows the averaged value of $r(x)$ and $\phi(x)$. We remark that for this fairly steep gradient the magnitude of the oscillators is about half the numerically computed magnitude for $\sigma = 0$. The qualitative picture of the phase gradient is virtually identical to that of the $\lambda - \omega$ system considered above. Finally, in Fig. 4.6, we show a space-time transient

plot of the u component for $\sigma = 19$. As can be seen, the magnitude of the oscillator damps out, and the equilibrium becomes asymptotically stable.

We expect that other examples of diffusively coupled oscillators with large gradients will behave in the same qualitative manner. There are two crucial assumptions that we have made for this behavior to occur: (i) the diffusion is strong, and (ii) the diffusion is scalar. Indeed, in a recent paper [17] Sherman and Rinzel show that for diffusively coupled membrane models (which have diffusion only of one of the variables), then complex and chaotic behavior can occur. In [8] we show that a pair of identical $\lambda - \omega$ oscillators can have a variety of complicated behaviors when the coupling is nonscalar.

For small diffusivity, the behavior of these systems as the gradient increases is quite different. Rather than a collapse to rest, phaselocking is lost and the medium breaks up into regions with different frequencies separated by quasi-periodic and chaotic regimes. We will consider some examples of this behavior and analyze this regime in a later paper.

REFERENCES

[1] N. KOPELL, *Towards a theory of modelling central pattern generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. H. Cohen, S. Rossignol, and S. Grillner, eds., John Wiley, New York, 1988.

[2] T. KIEMEL, *Three problems on coupled nonlinear oscillators,* Ph. D. thesis, Cornell University, Ithaca, NY, 1989.

[3] Y. KURAMOTO, *Chemical Oscillations, Waves, and Turbulence*, Springer-Verlag, New York, 1984.

[4] A. H. COHEN, P. J. HOLMES, AND R. H. RAND, *The nature of coupling between segmental oscillators of the lamprey spinal generator for locomotion: A mathematical model*, J. Math. Biol., 13 (1982), pp. 345–369.

[5] C. AMICK, *A problem in neural networks,* Proc. Roy. Soc., 118A (1991), pp. 1–12.

[6] S. STROGATZ AND R. MIROLLO, *Stability of incoherence in a population of coupled oscillators*, J. Stat. Phys. 63 (1991), pp. 613–635.

[7] G. B. ERMENTROUT AND N. KOPELL, *Frequency plateaus in a chain of weakly coupled oscillators* I, SIAM J. Math. Anal., 15 (1984), pp. 215–237.

[8] D. ARONSON, G. B. ERMENTROUT, AND N. KOPELL, *Amplitude response of coupled oscillators*, Phys. D, 41 (1990), pp. 403–449.

[9] T. CHAKRABORTY AND R.H. RAND, *The transition from phase-locking to drift in a system of two weakly coupled van der Pol oscillators,* Internat J. Non-linear Mech., 23 (1988), pp. 369–376.

[10] G. B. ERMENTROUT, *Stable periodic solutions to discrete and continuum arrays of weakly coupled nonlinear oscillators,* SIAM J. Appl. Math., 52 (1992), pp. 1665–1687.

[11] G. B. ERMENTROUT AND W. C. TROY, *Phaselocking in a reaction-diffusion system with a linear frequency gradient,* SIAM J. Appl. Math, 46 (1989), pp. 359–367.

[12] ———, *The uniqueness and stability of the rest state for strongly coupled oscillators,* SIAM J. Math. Anal., 20 (1989), pp. 1436–1446.

[13] W. Y. TAM AND H. L. SWINNEY, *Spatiotemporal patterns in a one-dimensional open reaction-diffusion system,* Phys. D, 46 (1990), pp. 10–22.

[14] J. A. VASTANO, T. RUSSO, AND H. L. SWINNEY, *Bifurcation to spatially induced chaos in a reaction-diffusion system,* Phys. D, 46 (1990), pp. 23–42.

[15] J. C. NEU, *Coupled chemical oscillators,* SIAM J. Appl. Math., 37 (1979), pp. 307–315.

[16] G. B. ERMENTROUT, in preparation.

[17] J. PAULLET, G. B. ERMENTROUT, AND W. TROY, *The existence of spiral waves in an oscillatory reaction-diffusion system*, SIAM J. Appl. Math., 54 (1994), pp. 1386–1401.

[18] J. B. MCLEOD AND A. SERRIN, *The existence of similarity solutions for some laminar boundary layer problems*, Arch. Rational Mech. Anal., 31 (1968), pp. 288–303.

[19] A. SHERMAN AND J. RINZEL, *Rhythmogenic effects of weak electrotonic coupling in neuronal models*, Proc. Nat. Acad. Sci. U.S.A., 89 (1992), pp. 2471–2474.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1521–1536, November 1994

005

SINGULAR PERTURBATIONS AND THE COUPLED/QUASI-STATIC APPROXIMATION IN LINEAR THERMOELASTICITY*

B. F. ESHAM† AND R. J. WEINACHT‡

Abstract. A uniform asymptotic expansion in the inertial constant ϵ is given for one-dimensional linear thermoelasticity by a two-timing method. The result shows that the usual coupled/quasi-static approximation ($\epsilon = 0$) is not uniformly asymptotically correct even to lowest order. The interaction between diffusion and wave propagation is clearly displayed. The proof of uniform asymptotic validity to order ϵ^2 is accomplished by energy estimates.

Key words. coupled/quasi-static linear thermoelasticity, singular perturbations, two-timing expansion, multiple scales

AMS subject classifications. 35B25, 73B30, 35K05, 73C02, 35Q72

1. Introduction.

In this paper we give a uniform asymptotic expansion with respect to the inertial constant for linearized one-dimensional (homogeneous, isotropic) thermoelasticity. The result shows that the usual coupled/quasi-static approximation (which corresponds to zero inertial constant) of the temperature, displacement, and stress is not uniformly asymptotically correct even to the lowest order. In contrast, this approximation of the entropy *is* correct to the lowest order.

The governing partial differential equations (PDEs) in question are (in dimensionless form)

$$\theta_t - \theta_{xx} + \gamma u_{xt} = 0, \tag{1.1}$$

$$\epsilon^2 u_{tt} - u_{xx} + \gamma \theta_x = 0, \tag{1.2}$$

where θ is the temperature (variation from a given reference temperature), u the displacement, ϵ the (square root of the) inertial constant, and γ the coupling constant, which is a measure of the thermal-mechanical interaction. Equation (1.1) is a statement of conservation of energy and (1.2) is a statement of balance of linear momentum (see, e.g., [1], [2], and [3] for a derivation of the equations).

The coupled/quasi-static approximation of (1.1)–(1.2) consists of setting $\epsilon = 0$ but retaining a nonzero coupling constant γ so that

$$\hat{\theta}_t - \hat{\theta}_{xx} + \gamma \hat{u}_{xt} = 0, \tag{1.3}$$

$$\gamma \hat{\theta}_x - \hat{u}_{xx} = 0. \tag{1.4}$$

If also $\gamma = 0$, the temperature $\tilde{\theta}$ is a solution of the classical heat equation

$$\tilde{\theta}_t - \tilde{\theta}_{xx} = 0, \tag{1.5}$$

*Received by the editors January 29, 1993; accepted for publication (in revised form) June 9, 1993.

†Department of Mathematics, State University of New York, Geneseo, New York.

‡Department of Mathematics, University of Delaware, Newark, Delaware 19716.

and under various boundary conditions the displacement is identically zero (rigid conductor).

The relationship of θ to $\hat{\theta}$ and $\tilde{\theta}$ has been studied extensively by Day (see [4]–[7]), who establishes the asymptotic equivalence of solutions for large t:

$$(\theta - \hat{\theta}) \to 0 \quad \text{and} \quad (\theta - \tilde{\theta}) \to 0 \quad \text{as } t \to \infty$$

for the temperature θ in (1.1)–(1.2), the coupled/quasi-static $\hat{\theta}$ in (1.3)–(1.4), and the classical $\tilde{\theta}$ in (1.5). Day's results are for the initial-boundary value problem (IBVP) with appropriate boundary conditions and are independent of the initial values of θ, u, u_t or $\hat{\theta}$ or $\tilde{\theta}$; moreover, the results are uniform in the space variable x. The cited results do not attempt to answer the question concerning the closeness of the approximations for finite t; indeed, the temperatures θ, $\hat{\theta}$, and $\tilde{\theta}$ need not be close initially. Our scaling for dimensionless quantities is the same as that used by Day in [7] but our notation differs in order to simplify the resulting formulas.

In this paper we focus attention on the asymptotic behavior in ϵ for small positive ϵ of solutions of an IBVP for (1.1)–(1.2). This is a singular perturbation problem. One might expect that there is an initial layer near $t = 0$ (as for hyperbolic-parabolic problems, see, e.g., [8] and [9]) and that one can obtain a uniform expansion by the addition of appropriate initial-layer correction terms. This is not the case, however, and there are wave propagation terms depending on ϵ, even at the lowest order. We obtain the asymptotic expansion by the two-timing technique of singular perturbation theory (see, e.g., [10]).

For definiteness we consider solutions of (1.1)–(1.2) subject to the boundary conditions

$$\theta(0,t;\epsilon) = 0, \quad \theta_x(1,t;\epsilon) = 0, \quad u_x(0,t;\epsilon) = 0, \quad u(1,t;\epsilon) = 0, \tag{1.6}$$

i.e., the temperature is maintained at zero on the stress-free end at $x = 0$, and the clamped end at $x = 1$ has perfect thermal insulation. We have treated other boundary conditions, including those considered by Day, but the resulting formulas here are simpler.

At $t = 0$ we specify the initial conditions

$$\theta(x,0;\epsilon) = \phi(x), \quad u(x,0;\epsilon) = f(x), \quad u_t(x,0;\epsilon) = g(x). \tag{1.7}$$

We will refer to the IBVP consisting of (1.1), (1.2), (1.6), and (1.7) as problem (P_ϵ). For smooth solutions of (P_ϵ) to exist there must be the usual compatibility between the initial and the boundary data, e.g.,

$$\phi(0) = \phi'(1) = f'(0) = f(1) = g'(0) = g(1) = 0$$

and further conditions for higher-order smoothness. These compatibility conditions will be assumed without further mention except in the statement of Theorem 1 in §5 on the validity of the asymptotic expansion. The theorem is proved by means of energy estimates.

2. The reduced problem and the coupled/quasi-static approximation. The coupled/quasi-static approximation consists of the system of PDEs

$$\hat{\theta}_t - \hat{\theta}_{xx} + \gamma \hat{u}_{xt} = 0, \tag{2.1}$$

$$\gamma \hat{\theta}_x - \hat{u}_{xx} = 0, \tag{2.2}$$

which is (1.1)–(1.2) with $\epsilon = 0$. The system (2.1)–(2.2) is to be supplemented by appropriate boundary and initial conditions. For definiteness we consider the boundary conditions (1.6)

$$\hat{\theta}(0,t;\epsilon) = 0, \quad \hat{\theta}_x(1,t;\epsilon) = 0, \quad \hat{u}_x(0,t;\epsilon) = 0, \quad \hat{u}(1,t;\epsilon) = 0, \tag{2.3}$$

but other boundary conditions, as mentioned above, are of interest. The system (2.1)–(2.2) is of second order like (1.1)–(1.2), but represents a change of type since (2.2) is just an ordinary differential equation (ODE) in x. The loss of the second time derivative of u when $\epsilon = 0$ implies that one can no longer impose the initial condition (1.7) on $\hat{u}_t$. In fact, it is not immediately evident how to correctly pose appropriate initial data. Because of (2.2) it is clear that $\hat{\theta}$ and $\hat{u}$ cannot be specified independently at $t = 0$.

The approach followed by Day [7], who is primarily interested in the evolution of the temperature, is to eliminate the displacement from the system (2.1)–(2.2). From (2.2) and the boundary conditions (2.3) we have

$$\gamma\hat{\theta}(x,t) = \hat{u}_x(x,t), \tag{2.4}$$

i.e., the displacement gradient is everywhere directly proportional to the temperature, so that (2.1) yields the classical heat equation

$$\hat{\theta}_t = (1+\gamma^2)^{-1}\hat{\theta}_{xx}. \tag{2.5}$$

With the boundary conditions (2.3) for $\hat{\theta}$ plus the initial condition $\hat{\theta}(x,0) = \phi(x)$, the IBVP for (2.5) is well posed with unique solution

$$\hat{\theta}(x,t) = \sum_{n=1}^{\infty} \phi_n e^{-\frac{\lambda_n^2}{c^2}t} \sin\lambda_n x, \tag{2.6}$$

where here and below $c := \sqrt{1+\gamma^2}$, $\lambda_n = \frac{(2n-1)\pi}{2}$ for $n = 1, 2, \ldots$, and ϕ_n is the nth Fourier coefficient of ϕ,

$$\phi_n = 2\int_0^1 \phi(x)\sin\lambda_n x \, dx.$$

In view of (2.4) it follows from (2.6) that

$$\hat{u}(x,t) = -\gamma\sum_{n=1}^{\infty} \frac{\phi_n}{\lambda_n} e^{-\frac{\lambda_n^2}{c^2}t} \cos\lambda_n x.$$

It is to be observed that $\sin\lambda_n x$ and $-(\frac{\gamma}{\lambda_n})\cos\lambda_n x$ appearing in the series for $\hat{\theta}$ and $\hat{u}$ are precisely the (unnormalized) eigenfunctions for the coupled/quasi-static problem (2.1)–(2.3).

In a similar manner, however, the temperature $\hat{\theta}$ can be eliminated from the system (2.1)–(2.2) yielding $\hat{u}_{tx} = (1+\gamma^2)^{-1}\hat{u}_{xxx}$, which, upon integration and imposition of the boundary conditions (2.3) on $\hat{u}$, yields the same heat equation, as in (2.5),

$$\hat{u}_t = (1+\gamma^2)^{-1}\hat{u}_{xx}. \tag{2.7}$$

The IBVP for (2.7) with boundary conditions (2.3) for $\hat{u}$ and initial condition $\hat{u}(x,0) = f(x)$ is well posed and has the unique solution given by

$$\hat{u}_1(x,t) = \sum_{n=1}^{\infty} f_n e^{-\frac{\lambda_n^2}{c^2}t} \cos \lambda_n x,$$

and from (2.4),

$$\hat{\theta}_1(x,t) = -\sum_{n=1}^{\infty} \frac{\lambda_n f_n}{\gamma} e^{-\frac{\lambda_n^2}{c^2}t} \sin \lambda_n x,$$

where f_n is the Fourier coefficient of f,

$$f_n = 2\int_0^1 f(x) \cos \lambda_n x \, dx.$$

The solutions $(\hat{\theta}, \hat{u})$ and $(\hat{\theta}_1, \hat{u}_1)$ do not coincide except in the special circumstance that $\gamma\phi(x) = f'(x)$. Thus, in addition to the loss of the initial condition on u_t, the independence of initial conditions on u and θ is also lost when $\epsilon = 0$. The crucial question is: What solution of (2.1)–(2.2) approximates the solution of (P_ϵ) up to $O(\epsilon)$ uniformly on $[0,1] \times [0,T]$? Following closely upon that question is the possibility of an asymptotic expansion in ϵ for higher-order approximations.

Before turning in the next section to the answer of the principal question just posed, it is enlightening to approach the situation from a different perspective. We note that (2.1) involves the time derivative of the entropy $\hat{\eta} := \hat{\theta} + \gamma \hat{u}_x$ and (2.2) the spatial derivative of the stress $\hat{\sigma} := \hat{u}_x - \gamma\hat{\theta}$. Introducing [7] the entropy and the stress as new dependent variables into the system (2.1)–(2.2) yields

$$\hat{\eta}_t - (1+\gamma^2)^{-1}\hat{\eta}_{xx} = 0, \tag{2.8}$$

$$\hat{\sigma}_x = 0. \tag{2.9}$$

The boundary conditions (2.3) together with (2.9) imply that $\hat{\sigma} \equiv 0$ and the boundary conditions (2.3) also imply

$$\hat{\eta}(0,t) = 0, \qquad \hat{\eta}_x(1,t) = 0. \tag{2.10}$$

In addition, the initial conditions (1.7) require that we take

$$\hat{\eta}(x,0) = \phi(x) + \gamma f'(x). \tag{2.11}$$

The parabolic IBVP consisting of (2.8), (2.10), and (2.11) is well posed with unique solution given by

$$\hat{\eta}(x,t) = \sum_{n=1}^{\infty} (\phi_n - \gamma\lambda_n f_n) e^{-\frac{\lambda_n^2}{c^2}t} \sin \lambda_n x,$$

from which it follows, using (2.4), that

$$\hat{\theta}_2(x,t) = \sum_{n=1}^{\infty} \frac{1}{c^2} (\phi_n - \gamma\lambda_n f_n) e^{-\frac{\lambda_n^2}{c^2}t} \sin \lambda_n x \tag{2.12}$$

and

$$\hat{u}_2(x,t) = \sum_{n=1}^{\infty} \frac{\gamma}{c^2\lambda_n}(\gamma\lambda_n f_n - \phi_n)e^{-\frac{\lambda_n^2}{c^2}t}\cos\lambda_n x. \tag{2.13}$$

The asymptotic analysis of §3 will confirm that the solution (θ, u) of (P_ϵ) is closely related to the solution $(\hat{\theta}_2, \hat{u}_2)$ of the coupled/quasi-static problem. In particular, this demonstrates the inadequacy of the solution (2.6) of the classical IBVP to describe uniformly the behavior of the temperature (even asymptotically to lowest order). The parabolic equations (2.5) and (2.7) clearly imply that in this approximation the evolution of the initial temperature and displacement distributions is governed solely by thermal diffusion. This is not the case for the solution of (P_ϵ) which, even to lowest order in ϵ, is subject to thermal diffusion and dissipative wave propagation.

One of the dilemmas regarding the specification of initial data when $\epsilon = 0$ can be resolved in the following way. For $\epsilon > 0$ we are free to specify the initial temperature and displacement independently of each other. As we will see, the solution of (P_ϵ) for small ϵ, if one neglects wave propagation, is approximated to lowest order in ϵ by the solution of (2.1)–(2.3) corresponding to initial values

$$\hat{\theta}(x,0) = \frac{\phi(x) + \gamma f'(x)}{1+\gamma^2}$$

and

$$\hat{u}(x,0) = -\frac{\gamma}{1+\gamma^2}\int_x^1 [\phi(\xi) + \gamma f'(\xi)]d\xi,$$

for which it is true that $\gamma\hat{\theta}(x,0) = \hat{u}_x(x,0)$. But if one considers all contributions to lowest order in ϵ including wave propagation, then $\theta(x,0;\epsilon) = \phi(x)$ and $u(x,0;\epsilon) = f(x)$.

3. Asymptotic representation of temperature and displacement. In this section we give an expansion of the solution of the full problem (P_ϵ) in powers of the inertial parameter ϵ for ϵ tending to zero. We will see that *part* of the lowest-order approximation is precisely the coupled/quasi-static approximation (2.12)–(2.13) given by the reduced problem. Our result is valid (as proved in §5) uniformly on $[0,1]\times[0,T]$ in the x,t-plane for any given $T > 0$. The method used to develop the expansion is the two-timing technique of singular perturbation theory (see, e.g., [10]).

We make an expansion with respect to the eigenfunctions of the coupled/quasi-static problem

$$\theta(x,t;\epsilon) \sim \sum_{n=1}^{\infty} \Theta_n(t;\epsilon)\sin\lambda_n x, \tag{3.1}$$

$$u(x,t;\epsilon) \sim \sum_{n=1}^{\infty} U_n(t;\epsilon)\cos\lambda_n x, \tag{3.2}$$

where $\lambda_n = \frac{(2n-1)\pi}{2}$ for $n = 1, 2, \ldots$ and λ_n^2/c^2 are the corresponding eigenvalues.

In view of the orthogonality of the eigenfunctions, substitution of (3.1), (3.2) into (P_ϵ) results in a coupled system of linear ODEs

$$\dot{\Theta}_n + \lambda_n^2\Theta_n - \gamma\lambda_n\dot{U}_n = 0, \tag{3.3}$$

$$\epsilon^2\ddot{U}_n + \lambda_n^2 U_n + \gamma\lambda_n\Theta_n = 0 \tag{3.4}$$

for each $n = 1, 2, \ldots$, together with initial conditions

$$\Theta_n(0;\epsilon) = \phi_n, \quad U_n(0;\epsilon) = f_n, \quad \dot{U}_n(0;\epsilon) = g_n, \tag{3.5}$$

where the ϕ_n and f_n are the Fourier coefficients of the initial data ϕ and f, as in §2, and similarly

$$g_n = 2\int_0^1 g(x)\cos\lambda_n x \, dx.$$

To study the asymptotics in ϵ of (3.3)–(3.5) we introduce two time scales for each mode,

$$s := t, \qquad \tau := \frac{t}{\epsilon}(1 + \beta_n\epsilon^2 + \gamma_n\epsilon^4 + \cdots),$$

with $\beta_n, \gamma_n, \ldots$, constants to be determined. This choice is motivated as follows: elimination of either Θ_n or U_n from (3.3)–(3.4) yields a third-order ODE with characteristic equation

$$\epsilon^2 r^3 + \epsilon^2\lambda_n^2 r^2 + (1+\gamma^2)\lambda_n^2 r + \lambda_n^4 = 0$$

whose roots r lead one to s and τ as given. One then makes the Ansatz

$$\begin{aligned}\Theta_n(t;\epsilon) &\sim \Phi_0^n(s,\tau) + \epsilon\Phi_1^n(s,\tau) + \epsilon^2\Phi_2^n(s,\tau) + \cdots,\\ U_n(t;\epsilon) &\sim F_0^n(s,\tau) + \epsilon F_1^n(s,\tau) + \epsilon^2 F_2^n(s,\tau) + \cdots,\end{aligned}$$

which upon substitution into (3.3)–(3.4) yields the following systems of PDEs:

$$L[F_0^n, \Phi_0^n] = 0, \qquad M[F_0^n, \Phi_0^n] = 0, \tag{3.6}$$

$$L[F_1^n, \Phi_1^n] = -2F_{0,s\tau}^n, \qquad M[F_1^n, \Phi_1^n] = -\mathcal{L}[F_0^n, \Phi_0^n], \tag{3.7}$$

$$\begin{aligned}L[F_2^n, \Phi_2^n] &= -2F_{1,s\tau}^n - F_{0,ss}^n - 2\beta_n F_{0,\tau\tau}^n,\\ M[F_2^n, \Phi_2^n] &= -\mathcal{L}[F_1^n, \Phi_1^n],\end{aligned} \tag{3.8}$$

$$\begin{aligned}L[F_3^n, \Phi_3^n] &= -2F_{2,s\tau}^n - F_{1,ss}^n - 2\beta_n F_{1,\tau\tau}^n - 2\beta_n F_{0,s\tau}^n,\\ M[F_3^n, \Phi_3^n] &= -\mathcal{L}[F_2^n, \Phi_2^n] + \beta_n\mathcal{L}[F_0^n, \Phi_0^n],\\ &\vdots\end{aligned} \tag{3.9}$$

where the differential operators L, M, and $\mathcal{L}$ are defined by

$$\begin{aligned}L[F,\Phi] &:= F_{\tau\tau} + \lambda_n^2 F + \gamma\lambda_n\Phi,\\ M[F,\Phi] &:= \Phi_\tau - \gamma\lambda_n F_\tau,\\ \mathcal{L}[F,\Phi] &:= \Phi_s + \lambda_n^2\Phi - \gamma\lambda_n F_s.\end{aligned} \tag{3.10}$$

Likewise, initial conditions are obtained from (3.5)

$$
\begin{aligned}
&\Phi_0^n(0,0) = \phi_n, \quad F_0^n(0,0) = f_n, \quad F_{0,\tau}^n(0,0) = 0,\\
&\Phi_j^n(0,0) = 0, \quad F_j^n(0,0) = 0, \quad j = 1,2,\ldots,\\
&F_{1,\tau}^n(0,0) = g_n - F_{0,s}^n(0,0),\\
&F_{2,\tau}^n(0,0) = -F_{1,s}^n(0,0).\\
&\qquad\vdots
\end{aligned} \tag{3.11}
$$

The temperature variable Φ can be eliminated from each of the systems (3.6)–(3.9) by using the operator identity

$$
\frac{\partial}{\partial\tau}L[F,\Phi] - \gamma\lambda_n M[F,\Phi] = \frac{\partial^3 F}{\partial\tau^3} + c^2\lambda_n^2\frac{\partial F}{\partial\tau}
$$

yielding the same third-order operator for $F_0^n, F_1^n, \ldots$. As above, c is the positive constant $\sqrt{1+\gamma^2}$. To avoid the presence of secular terms and subsequent deterioration of the approximation, we ask that the resulting equations be homogeneous. Thus we require the additional conditions:

$$
K[F_0^n, \Phi_0^n] = 0, \tag{3.12}
$$

$$
K[F_1^n, \Phi_1^n] = F_{0,\tau ss}^n + 2\beta_n F_{0,\tau\tau\tau}^n, \tag{3.13}
$$

$$
\begin{aligned}
K[F_2^n, \Phi_2^n] &= F_{1,\tau ss}^n + 2\beta_n F_{1,\tau\tau\tau}^n\\
&\quad - \gamma\lambda_n M[F_1^n, \Phi_1^n] + 2\beta_n F_{0,\tau\tau s}^n,\\
&\qquad\vdots
\end{aligned} \tag{3.14}
$$

where the operator K is defined by

$$
K[F,\Phi] = -2\frac{\partial}{\partial\tau}[F_{\tau s}] + \gamma\lambda_n\mathcal{L}[F,\Phi]. \tag{3.15}
$$

With the procedure for the expansion now delineated to any order, we find the solution of the problems for F_j^n and Φ_j^n, $j = 0,1,2$. Details appear in the Appendix.

Thus we are led to the following asymptotic representation which is valid uniformly on $[0,1]\times[0,T]$ for any given $T > 0$:

$$
\theta(x,t;\epsilon) = \theta_0(x,t;\epsilon) + \epsilon\theta_1(x,t;\epsilon) + \epsilon^2\theta_2(x,t;\epsilon) + O(\epsilon^3), \tag{3.16}
$$

$$
u(x,t;\epsilon) = u_0(x,t;\epsilon) + \epsilon u_1(x,t;\epsilon) + \epsilon^2 u_2(x,t;\epsilon) + O(\epsilon^3), \tag{3.17}
$$

where for $j = 0,1,2$

$$
\theta_j(x,t;\epsilon) = \sum_{n=1}^{\infty}\gamma\lambda_n\sin\lambda_n x \times \left[[C_j^n + \gamma_j^n t]e^{-\frac{\lambda_n^2}{c^2}t}\right. \tag{3.18}
$$

$$+ [(\tilde{A}_j^n + \alpha_j^n t)\cos\mu_n(\epsilon)t + (\tilde{B}_j^n + \beta_j^n t)\sin\mu_n(\epsilon)t]e^{-\frac{\gamma^2\lambda_n^2}{2c^2}t}\Bigg],$$

(3.19)

$$u_j(x,t;\epsilon) = \sum_{n=1}^{\infty}\cos\lambda_n x \times \Bigg[-\gamma^2[\tilde{C}_j^n + \gamma_j^n t]e^{-\frac{\lambda_n^2}{c^2}t} + [(A_j^n + \alpha_j^n t)\cos\mu_n(\epsilon)t + (B_j^n + \beta_j^n t)\sin\mu_n(\epsilon)t]\, e^{-\frac{\gamma^2\lambda_n^2}{2c^2}t}\Bigg]$$

for $\mu_n(\epsilon) = \lambda_n c(1+\beta_n\epsilon^2)/\epsilon$ and, in particular,

$$\theta_0(x,t;\epsilon) = \sum_{n=1}^{\infty}\frac{1}{c^2}(\phi_n - \gamma\lambda_n f_n)e^{-\frac{\lambda_n^2}{c^2}t}\sin\lambda_n x$$

(3.20)

$$+ \sum_{n=1}^{\infty} e^{-\frac{\gamma^2\lambda_n^2}{2c^2}t}\left\{\frac{\gamma}{c^2}(\lambda_n f_n + \gamma\phi_n)\cos\mu_n(\epsilon)t\right\}\sin\lambda_n x$$

and

$$u_0(x,t;\epsilon) = \sum_{n=1}^{\infty}\frac{\gamma}{c^2\lambda_n}(\gamma\lambda_n f_n - \phi_n)e^{-\frac{\lambda_n^2}{c^2}t}\cos\lambda_n x$$

(3.21)

$$+ \sum_{n=1}^{\infty} e^{-\frac{\gamma^2\lambda_n^2}{2c^2}t}\left\{\frac{1}{c^2}\left(f_n + \frac{\gamma}{\lambda_n}\phi_n\right)\cos\mu_n(\epsilon)t\right\}\cos\lambda_n x.$$

The coefficients C_j^n etc. for θ_1, θ_2, u_1, and u_2 are constants depending only on the data and are given explicitly at the end of the Appendix.

The first term in (3.20) and the first term in (3.21) are precisely the coupled/quasi-static approximations given in (2.12) and (2.13). Thus, even to lowest order, that approximation does not give a uniform approximation in ϵ of the full problem (P_ϵ) on $[0,1]\times[0,T]$. Such a uniform approximation is provided by (3.16) and (3.17) and to lowest order by (3.20) and (3.21). The correction terms here do not exhibit boundary layer behavior. Their influence decays exponentially but over a time interval independent of ϵ. Moreover, the interaction of the thermal diffusion and wave propagation is exhibited at the various orders in ϵ. Note that the damping of the waves is due to the coupling and, to the lowest order, is independent of the inertial constant ϵ. As the parameter ϵ gets smaller, the waves travel faster but do not damp out any more rapidly.

4. Entropy and stress. One can introduce [7] the entropy η and the stress σ,

$$\eta := \theta + \gamma u_x, \qquad \sigma := u_x - \gamma\theta, \tag{4.1}$$

in place of the temperature θ and the displacement u in order to study the thermoelastic problem considered here. Then from (1.1)–(1.2) one obtains the system of PDEs

$$(1+\gamma^2)\eta_t - \eta_{xx} + \gamma\sigma_{xx} = 0, \tag{4.2}$$

$$\epsilon^2(\sigma_{tt} + \gamma\eta_{tt}) - (1+\gamma^2)\sigma_{xx} = 0, \tag{4.3}$$

which is subject to the boundary conditions

$$\eta(0,t) = \sigma(0,t) = 0, \tag{4.4}$$

$$\eta_x(1,t) = \sigma_x(1,t) = 0, \tag{4.5}$$

and the initial conditions

$$\eta(x,0) = \chi(x), \tag{4.6}$$

$$\sigma(x,0) = \tau(x), \tag{4.7}$$

$$\sigma_t(x,0) + \gamma\eta_t(x,0) = \zeta(x), \tag{4.8}$$

where the relationships between these data and those for θ, u are

$$\chi(x) = \phi(x) + \gamma f'(x),$$
$$\tau(x) = f'(x) - \gamma\phi(x),$$
$$\zeta(x) = (1+\gamma^2)g'(x).$$

The initial-boundary value problem given by (4.2)–(4.8) offers an alternative approach to the thermoelastic problem. Moreover, there are certain simplifying features which accrue by using the entropy/stress variables. For example, it follows from (4.2)–(4.3) that the coupled/quasi-static approximation ($\epsilon = 0$) in these variables is governed by an uncoupled system

$$(1+\gamma^2)\hat{\eta}_t - \hat{\eta}_{xx} = 0,$$
$$\hat{\sigma}_{xx} = 0,$$

so that the stress in this approximation is (because of the boundary condition) identically zero and the entropy (in this approximation) is a solution of a heat equation.

The asymptotics in ϵ for the entropy and stress can be obtained immediately via (4.1) from the results in §3. The corresponding formulas are obvious (see, e.g., (4.9) and (4.10) below). Conversely, one can develop the asymptotic result for the entropy and stress directly from (4.2)–(4.8) by the same two-timing method of §3 (and the Appendix), and subsequently the asymptotics for the temperature and displacement follow via (4.1). It is a fact that there are simplifications in those details if one works with the entropy and stress. For example, the terms F_0^n (of equation (A.2)) for the displacement and Φ_0^n (of (A.3)) for the temperature have analogues

$$\chi_n e^{-\frac{\lambda_n^2}{c^2}s}$$

for the entropy (i.e., their contribution to the entropy exhibits pure diffusion) and

$$(\tau_n \cos\lambda_n c\tau)e^{-\frac{\gamma^2\lambda_n^2}{2c^2}s}$$

for the stress (i.e., this contribution to the stress exhibits pure damped wave motion). Here and below χ_n, τ_n, and ζ_n are the Fourier coefficients of the initial data in (4.6)–(4.8) so that $\chi_n = \phi_n - \gamma\lambda_n f_n$, $\tau_n = -\lambda_n f_n - \gamma\phi_n$, and $\zeta_n = -(1+\gamma^2)\lambda_n g_n$. In consequence,

$$\eta(x,t;\epsilon) = \sum_{n=1}^{\infty} \chi_n e^{-\frac{\lambda_n^2}{c^2}t}\sin\lambda_n x$$
$$+ \epsilon\sum_{n=1}^{\infty}\frac{\gamma}{c^3}e^{-\frac{\gamma^2\lambda_n^2}{2c^2}t}\{\lambda_n\tau_n\sin\mu_n(\epsilon)t\}\sin\lambda_n x + O(\epsilon^2) \tag{4.9}$$

and

$$\sigma(x,t;\epsilon) = \sum_{n=1}^{\infty} e^{-\frac{\gamma^2\lambda_n^2}{2c^2}t}(\tau_n \cos \mu_n(\epsilon)t) \sin \lambda_n x$$

$$+ \epsilon \sum_{n=1}^{\infty} \left[\frac{\zeta_n}{c\lambda_n} + \frac{\gamma\lambda_n}{2c^3}(2\chi_n - \gamma\tau_n)\right] e^{-\frac{\gamma^2\lambda_n^2}{2c^2}t} \sin \mu_n(\epsilon)t \sin \lambda_n x + O(\epsilon^2), \tag{4.10}$$

which shows how the diffusions and damped wave behavior interact at the lowest orders in ϵ. In particular, we note that to lowest order the coupled/quasi-static approximation correctly describes the entropy but not the stress.

5. Uniform asymptotic validity. In this section we prove the following.

THEOREM 1. *Let ϕ and g belong to the Sobolev space $H^8(0,1)$ and let f belong to $H^9(0,1)$ with*

$$\begin{aligned} &\phi(0) = \phi''(0) = \cdots = \phi^{(8)}(0) = 0, \\ &f'(0) = f'''(0) = \cdots = f^{(7)}(0) = 0, \qquad g'(0) = g'''(0) = \cdots = g^{(7)}(0) = 0, \\ &\phi'(1) = \phi'''(1) = \cdots = \phi^{(7)}(1) = 0, \\ &f(1) = f''(1) = \cdots = f^{(8)}(1) = 0, \qquad g(1) = g''(1) = \cdots = g^{(6)}(1) = 0. \end{aligned}$$

Then for any given $T > 0$ the solution (θ, u) of the initial-boundary value problem (P_ϵ) given by (1.1), (1.2), (1.6), *and* (1.7) *satisfies*

$$\begin{aligned} \theta(x,t;\epsilon) &= \theta_0(x,t;\epsilon) + \epsilon\theta_1(x,t;\epsilon) + O(\epsilon^2), \\ u(x,t;\epsilon) &= u_0(x,t;\epsilon) + \epsilon u_1(x,t;\epsilon) + O(\epsilon^2) \end{aligned}$$

uniformly as $\epsilon \to 0^+$ on $[0,1] \times [0,T]$. Here θ_0, θ_1 and u_0, u_1 are given in (3.18) *and* (3.19).

Remark. The smoothness hypotheses of the theorem are far more stringent than those required for the existence of smooth solutions in the closed region $[0,1] \times [0,T]$. They are dictated by the appearance of θ_2 and u_2 (see (3.18) and (3.19)) in the proof of the theorem, though not in its statement. The higher-order energy estimates considered in the proof place the greatest demands on the smoothness of the data.

We note that Day [6] also has rather stringent regularity hypotheses ($C^6(S)$ behavior on the strip $S = [0,1] \times (0,\infty)$) in examining the long time asymptotics of the coupled/quasi-static approximation for the boundary conditions $u(0,t) = u(1,t) = \theta(0,t) = 0$ and $\theta_x(1,t) = h(t)$. When this last boundary condition is replaced by $\theta(1,t) = f(t)$ Day [5] requires $\theta, u \in C^8([0,1] \times [0,\infty])$ and $f^{(n)} \in L^1(0,\infty) \cap L^2(0,\infty)$ for $2 \le n \le 6$ in examining the uncoupled/quasi-static approximation.

Proof. The proof uses energy estimates on the remainder terms P and R defined by

$$\begin{aligned} P(x,t;\epsilon) &:= \theta(x,t;\epsilon) - [\theta_0(x,t;\epsilon) + \epsilon\theta_1(x,t;\epsilon) + \epsilon^2\theta_2(x,t;\epsilon)], \\ R(x,t;\epsilon) &:= u(x,t;\epsilon) - [u_0(x,t;\epsilon) + \epsilon u_1(x,t;\epsilon) + \epsilon^2 u_2(x,t;\epsilon)]. \end{aligned}$$

Then (P,R) is a solution of the IBVP consisting of the system of PDEs

$$P_t - P_{xx} + \gamma R_{xt} = \epsilon^2\rho(x,t;\epsilon), \tag{5.1}$$

$$\epsilon^2 R_{tt} - R_{xx} + \gamma P_x = \epsilon^3 r(x,t;\epsilon) \tag{5.2}$$

with the initial conditions

$$P(x,0;\epsilon)=0, \quad R(x,0;\epsilon)=0, \quad R_t(x,0;\epsilon)=\epsilon^2 z(x;\epsilon)$$

and boundary conditions

$$P(0,t;\epsilon)=0, \quad P_x(1,t;\epsilon)=0, \quad R_x(0,t;\epsilon)=0, \quad R(1,t;\epsilon)=0.$$

The nonhomogeneous terms ρ and r are given by

$$\rho(x,t;\epsilon)=\sum_{n=1}^{\infty} e^{\frac{-\gamma^2\lambda_n^2}{2c^2}t}\,\rho_n\,\sin\lambda_n x$$

where

$$\begin{aligned}\rho_n=&\left(\cos\mu_n(\epsilon)t\left\{\left[-\frac{\gamma(\gamma^4+4\gamma^2+8)\lambda_n^5}{8c^6}\right]A_0^n+\left[\frac{\gamma(\gamma^2+2)\lambda_n^4}{2c^3}\right]B_1^n+[\gamma\lambda_n^3](A_2^n+\alpha_2^n t)\right\}\right.\\&\left.+\epsilon\sin\mu_n(\epsilon)t\left\{\left[-\frac{\gamma^3(\gamma^2+2)(\gamma^2+4)\lambda_n^6}{16c^9}\right]A_0^n+\left[\frac{\gamma^3(\gamma^2+4)\lambda_n^5}{8c^6}\right]B_1^n\right\}\right),\end{aligned}$$

and

$$\begin{aligned}r(x,t;\epsilon)=&\epsilon\sum_{n=1}^{\infty} e^{\frac{-\lambda n^2}{c^2}t}\cos\lambda_n x\cdot\left\{\frac{\gamma^2\lambda_n^2}{c^2}\left[\frac{\lambda_n^4}{c^6}C_0^n-\frac{\lambda_n^2}{c^2}(C_2^n+\gamma_2^n t)+2\gamma_2^n\right]\right\}\\&+\sum_{n=1}^{\infty} e^{\frac{-\gamma^2\lambda_n^2}{2c^2}t}\,r_n\,\cos\lambda_n x,\end{aligned}$$

where

$$\begin{aligned}r_n=&\epsilon\cos\mu_n(\epsilon)t\left(\left[-\frac{\gamma^4(\gamma^2+4)^2\lambda_n^6}{64c^{10}}\right]A_0^n+\left[\frac{\gamma^4(\gamma^2+4)\lambda_n^5}{8c^7}\right]B_1^n\right.\\&\left.+\frac{\gamma^2\lambda_n^4}{2c^4}\left[(\gamma^2+2)-\frac{\epsilon^2\gamma^2(\gamma^2+4)^2\lambda_n^2}{32c^6}\right](A_2^n+\alpha_2^n t)+\left[-\frac{\gamma^2\lambda_n^2}{c^2}\right]\alpha_2^n\right)\\&+\sin\mu_n(\epsilon)t\left(\left[-\frac{\gamma^4(\gamma^2+4)\lambda_n^5}{8c^7}\right]A_0^n+\frac{\gamma^2\lambda_n^4}{2c^4}\left[(\gamma^2+2)-\frac{\epsilon^2\gamma^2(\gamma^2+4)^2\lambda_n^2}{32c^6}\right]B_1^n\right.\\&+\frac{\gamma^2\lambda_n^3}{c}\left[1-\frac{\epsilon^2\gamma^2(\gamma^2+4)\lambda_n^2}{8c^6}\right](A_2^n+\alpha_2^n t)\\&\left.+\left[-2\lambda_n c+\frac{\epsilon^2\gamma^2(\gamma^2+4)\lambda_n^3}{4c^5}\right]\alpha_2^n\right),\end{aligned}$$

and the nonhomogeneous boundary datum z is

$$\begin{aligned}z(x;\epsilon)=&\sum_{n=1}^{\infty}\left\{-\frac{\gamma\lambda_n^3(\gamma^6+18\gamma^4-72\gamma^2+16)}{16c^{10}}\phi_n\right.\\&\left.+\frac{5\gamma^2\lambda_n^4(\gamma^4-12\gamma^2+8)}{16c^{10}}f_n+\frac{3\gamma^2\lambda_n^2(\gamma^2-4)}{8c^6}g_n\right\}\cos\lambda_n x.\end{aligned}$$

Now we obtain energy estimates for (P, R) in the usual way: multiply (5.1) by P and (5.2) by R_t; then integrate the sum over $[0,1] \times [0,t]$ for $0 < t \leq T$; integrate by parts using the boundary conditions to get the energy identity

$$E(t) - E(0) + \int_0^t \int_0^1 P_x^2 dx ds = \int_0^t \int_0^1 (\epsilon^2 \rho P + \epsilon^3 r R_t) dx ds, \tag{5.3}$$

where the energy at time t is

$$E(t) := \frac{1}{2} \int_0^1 (P^2 + \epsilon^2 R_t^2 + R_x^2) dx$$

so that

$$E(0) = \frac{\epsilon^6}{2} \int_0^1 z^2(x;\epsilon) dx.$$

Use of the Schwarz inequality and the arithmetic-geometric mean inequality gives the following estimate of the right-hand side of (5.3):

$$\int_0^t \int_0^1 (\epsilon^4 \rho^2 + P^2 + \epsilon^4 r^2 + \epsilon^2 R_t^2) dx ds$$

so that

$$E(t) + \int_0^t \int_0^1 P_x^2 dx ds \leq \left[E(0) + \epsilon^4 \int_0^t \int_0^1 (\rho^2 + r^2) dx ds \right] + \int_0^t E(s) ds,$$

and thus by the Gronwall Lemma,

$$E(t) + \int_0^t \int_0^1 P_x^2 dx ds \leq e^T \left\{ \frac{\epsilon^6}{2} \int_0^1 z^2(x;\epsilon) dx + \epsilon^4 \int_0^T \int_0^1 (\rho^2 + r^2) dx ds \right\}.$$

Therefore, we have the $L^2(0,1)$-estimates

$$\|P(\cdot, t; \epsilon)\|^2 \leq 2E(t) = O(\epsilon^4),$$
$$\|R_x(\cdot, t; \epsilon)\|^2 \leq 2E(t) = O(\epsilon^4),$$

and so from the Sobolev inequality

$$|R(x,t;\epsilon)|^2 \leq \int_0^1 R_x^2(\xi, t; \epsilon) d\xi,$$

a uniform estimate follows:

$$R(x,t;\epsilon) = O(\epsilon^2).$$

To obtain a uniform estimate on P we use a higher-order energy estimate as follows: differentiate (5.1) and (5.2) with respect to x, multiply the resulting equations by P_x and R_{xt}, respectively, integrate the sum over $[0,1] \times [0,t]$, integrate by parts using higher-order boundary conditions obtained from (5.1)–(5.2) and the original boundary conditions. This yields the energy identity

$$\hat{E}(t) - \hat{E}(0) + \int_0^t \int_0^1 P_{xx}^2 dx ds = \int_0^t \int_0^1 (\epsilon^2 \rho_x P_x + \epsilon^3 r_x R_{xt}) dx ds, \tag{5.4}$$

where

$$\hat{E}(t) := \frac{1}{2}\int_0^1 (P_x^2 + \epsilon^2 R_{xt}^2 + R_{xx}^2)dx.$$

It should be mentioned that it is this part of the argument that uses the stringent smoothness hypotheses of the theorem.

Estimating the right-hand side of (5.4) as was done for (5.3) yields $\hat{E}(t) = O(\epsilon^4)$ uniformly for $0 < t \leq T$ and so

$$|P(x,t;\epsilon)|^2 \leq \int_0^1 P_x^2 dx \leq 2\hat{E}(t) = O(\epsilon^4).$$

Thus we have the desired uniform estimate on P and the proof of the theorem is complete.

Appendix. Determination of F_j^n, Φ_j^n.

PROBLEM 0. The lowest-order terms F_0^n, Φ_0^n are determined by the system of PDEs

$$L[F_0^n, \Phi_0^n] = 0, \quad M[F_0^n, \Phi_0^n] = 0, \quad K[F_0^n, \Phi_0^n] = 0$$

subject to the initial conditions

$$\Phi_0^n(0,0) = \phi_n, \quad F_0^n(0,0) = f_n, \quad F_{0,\tau}^n(0,0) = 0,$$

where the operators L, M, K are defined in (3.10) and (3.15).

Elimination of Φ_0^n from $L[F_0^n, \Phi_0^n] = 0$, $M[F_0^n, \Phi_0^n] = 0$ yields

$$\frac{\partial^3 F_0^n}{\partial \tau^3} + c^2\gamma_n^2 \frac{\partial F_0^n}{\partial \tau} = 0$$

so that with $c = \sqrt{1+\gamma^2}$ as above,

$$F_0^n(s,\tau) = a_0^n(s)\cos\lambda_n c\tau + b_0^n(s)\sin\lambda_n c\tau + d_0^n(s),$$

and then from $M[F_0^n, \Phi_0^n] = 0$ we obtain

$$\Phi_0^n(s,\tau) = \gamma\lambda_n\{a_0^n(s)\cos\lambda_n c\tau + b_0^n(s)\sin\lambda_n c\tau + c_0^n(s)\}$$

with a_0^n, b_0^n, c_0^n, and d_0^n to be determined. From $L[F_0^n, \Phi_0^n] = 0$ one obtains the additional relationship

$$d_0^n(s) = -\gamma^2 c_0^n(s)$$

while application of the initial conditions implies

$$a_0^n(0) = \frac{1}{c^2}\left(f_n + \frac{\gamma}{\lambda_n}\phi_n\right), \quad b_0^n(0) = 0, \quad c_0^n(0) = \frac{1}{c^2}\left(\frac{\phi_n}{\gamma\lambda_n} - f_n\right).$$

Finally the equation $K[F_0^n, \Phi_0^n] = 0$ yields the first-order ODEs

$$\text{(A.1)} \qquad 2c^2\frac{da_0^n}{ds} + \gamma^2\lambda_n^2 a_0^n = 0, \quad 2c^2\frac{db_0^n}{ds} + \gamma^2\lambda_n^2 b_0^n = 0, \quad c^2\frac{dc_0^n}{ds} + \lambda_n^2 c_0^n = 0,$$

so that F_0^n and Φ_0^n are completely determined. The results are

$$F_0^n(s,\tau) = \frac{\gamma}{\lambda_n c^2}(\gamma\lambda_n f_n - \phi_n)e^{-\frac{\lambda_n^2}{c^2}s} + e^{-\frac{\gamma^2\lambda_n^2}{2c^2}s}\left\{\frac{1}{c^2}\left(f_n + \frac{\gamma}{\lambda_n}\phi_n\right)\cos\lambda_n c\tau\right\}, \tag{A.2}$$

$$\Phi_0^n(s,\tau) = \frac{1}{c^2}(\phi_n - \gamma\lambda_n f_n)e^{-\frac{\lambda_n^2}{c^2}s} + e^{-\frac{\gamma^2\lambda_n^2}{2c^2}s}\left\{\frac{\gamma}{c^2}(\lambda_n f_n + \gamma\phi_n)\cos\lambda_n c\tau\right\}. \tag{A.3}$$

PROBLEM 1. The functions F_1^n and Φ_1^n, are determined by the equations (3.7), (3.13) subject to the initial conditions (3.11) and thus their solution follows the same pattern as that for F_0^n and Φ_0^n as we now show. By virtue of the fact that $K[F_0^n, \Phi_0^n] = 0$ elimination of Φ_1^n from (3.7) yields for F_1^n the same third-order ODE as for F_0^n and so

$$F_1^n(s,\tau) = a_1^n(s)\cos\lambda_n c\tau + b_1^n(s)\sin\lambda_n c\tau + d_1^n(s).$$

Thus the "M-equation" in (3.7) yields

$$\Phi_1^n(s,\tau) = \gamma\lambda_n\left\{a_1^n(s)\cos\lambda_n c\tau + b_1^n(s)\sin\lambda_n c\tau + c_1^n(s) - \frac{\lambda_n}{c}a_0^n(s)\sin\lambda_n c\tau\right\},$$

while the "L-equation" in (3.7) implies

$$d_1^n(s) = -\gamma^2 c_1^n(s).$$

The initial conditions (3.11) imply that

$$a_1^n(0) = 0, \quad b_1^n(0) = \frac{g_n}{\lambda_n c} + \frac{\gamma}{2c^5}[3\gamma\lambda_n f_n + \phi_n(\gamma^2 - 2)], \quad c_1^n(0) = 0,$$

while the "K-equation" (3.13) yields the same first-order ODEs (A.1) for a_1^n, b_1^n, c_1^n as for a_0^n, b_0^n, c_0^n *provided* that

$$\beta_n = -\frac{\gamma^2\lambda_n^2(\gamma^2+4)}{8c^6}.$$

Hence

$$F_1^n(s,\tau) = e^{-\frac{\gamma^2\lambda_n^2}{2c^2}s}\left\{\frac{g_n}{\lambda_n c} + \frac{\gamma^2\lambda_n}{2c^5}\left[3f_n + \frac{\phi_n}{\gamma\lambda_n}(\gamma^2-2)\right]\right\}\sin\lambda_n c\tau$$

and

$$\Phi_1^n(s,\tau) = e^{-\frac{\gamma^2\lambda_n^2}{2c^2}s}\left\{\frac{\gamma g_n}{c} + \frac{\gamma\lambda_n^2(\gamma^2-2)f_n - \gamma^2\lambda_n(\gamma^2+4)\phi_n}{2c^5}\right\}\sin\lambda_n c\tau.$$

PROBLEM 2. The functions F_2^n and Φ_2^n are determined by (3.8), (3.14) subject to the initial conditions (3.11). Following the same method as above one finds

$$F_2^n(s,\tau) = e^{\frac{-\lambda_n^2}{c^2}s}\left\{\frac{\gamma^2}{c^4}g_n - \frac{\gamma^2\lambda_n^2}{c^8}(2-\gamma^2)f_n - \frac{\gamma\lambda_n}{c^8}(2\gamma^2-1)\phi_n \right.$$
$$\left. - s\left[\frac{\gamma^4\lambda_n^4}{c^{10}}\left(f_n - \frac{\phi_n}{\gamma\lambda n}\right)\right]\right\}$$
$$+ e^{\frac{-\gamma^2\lambda_n^2}{2c^2}s}\cos\lambda_n c\tau\left\{-\frac{\gamma^2}{c^4}g_n + \frac{\gamma^2\lambda_n^2}{c^8}(2-\gamma^2)f_n + \frac{\gamma\lambda_n}{c^8}(2\gamma^2-1)\phi_n \right.$$
$$\left. + s\left[\frac{\gamma^2\lambda_n^4}{2c^{10}}\left(f_n + \frac{\gamma\phi_n}{\lambda_n}\right)\right]\right\}$$

and

$$\Phi_2^n(s,\tau) = e^{-\frac{\lambda_n^2}{c^2}s}\left\{-\frac{\gamma\lambda_n}{c^4}g_n - \frac{\gamma\lambda_n^3(2\gamma^2-1)}{c^8}f_n + \frac{3\gamma^2\lambda_n^2\phi_n}{c^8} + s\left[\frac{\gamma^3\lambda_n^5}{c^{10}}\left(f_n - \frac{\phi_n}{\gamma\lambda_n}\right)\right]\right\}$$
$$+ e^{-\frac{\gamma^2\lambda_n^2}{2c^2}s}\cos\lambda_n c\tau\left\{\frac{\gamma\lambda_n}{c^4}g_n + \frac{\gamma\lambda_n^3(2\gamma^2-1)f_n - 3\gamma^2\lambda_n^2\phi_n}{c^8} \right.$$
$$\left. + s\left[\frac{\gamma^3\lambda_n^5}{2c^{10}}\left(f_n + \frac{\gamma\phi_n}{\lambda_n}\right)\right]\right\}.$$

Collecting the above results yields (3.18), (3.19) and hence (3.20), (3.21) with

$$\tilde{A}_0^n = A_0^n = \frac{(f_n + \frac{\gamma\phi_n}{\lambda_n})}{c^2}, \quad \tilde{B}_0^n = B_0^n = 0, \quad \tilde{C}_0^n = C_0^n = \frac{(\frac{\phi_n}{\gamma\lambda_n} - f_n)}{c^2},$$
$$\alpha_0^n = \beta_0^n = \gamma_0^n = 0,$$

$$\tilde{A}_1^n = A_1^n = 0, \quad B_1^n = \frac{g_n}{c\lambda_n} + \frac{\gamma^2\lambda_n}{2c^5}\left[3f_n + \frac{\phi_n}{\gamma\lambda_n}(\gamma^2-2)\right],$$
$$\tilde{B}_1^n = B_1^n - \frac{\lambda_n}{c}A_0^n, \quad \tilde{C}_1^n = C_1^n = 0, \quad \alpha_1^n = \beta_1^n = \gamma_1^n = 0,$$

$$A_2^n = -\frac{\gamma^2}{c^4}g_n + \frac{\gamma^2\lambda_n^2}{c^8}\left[(2-\gamma^2)f_n + \frac{\phi_n}{\gamma\lambda_n}(2\gamma^2-1)\right],$$
$$\tilde{A}_2^n = A_2^n - \frac{\lambda_n^2(2+\gamma^2)}{2c^4}A_0^n + \frac{\lambda_n}{c}B_1^n,$$
$$\tilde{B}_2^n = B_2^n = 0,$$

$$C_2^n = -\frac{g_n}{c^4} + \frac{\lambda_n}{c^8}[3\gamma\phi_n - (2\gamma^2-1)\lambda_n f_n], \qquad \tilde{C}_2^n = C_2^n - \frac{\lambda_n^2}{c^4}C_0^n,$$
$$\alpha_2^n = \frac{\gamma^2\lambda_n^3}{2c^{10}}(\lambda_n f_n + \gamma\phi_n), \quad \beta_2^n = 0, \quad \gamma_2^n = \frac{\gamma\lambda_n^3}{c^{10}}(\gamma\lambda_n f_n - \phi_n).$$

REFERENCES

[1] B. A. BOLEY AND J. H. WEINER, *Theory of Thermal Stresses*, Wiley, New York, 1960.

[2] D. E. CARLSON, *Linear Thermoelasticity*, in Handbuch der Physik, Volume VIa /2, Springer-Verlag, New York, 1972.

[3] P. CHADWICK, *Thermoelasticity: The dynamic theory*, in Progress in Solid Mechanics, North-Holland, Amsterdam, 1960.

[4] W. A. DAY, *Justification of the uncoupled and quasi-static approximation in a problem of dynamic thermoelasticity*, Arch. Rational Mech. Anal., 77 (1981), pp. 387–396.

[5] ———, *Further justification of the uncoupled and quasi-static approximations in thermoelasticity*, Arch. Rational Mech. Anal., 79 (1982), pp. 85–95.

[6] ———, *A comment on approximations to the temperature in dynamic linear thermoelasticity*, Arch. Rational Mech. Anal., 85 (1984), pp. 237–250.

[7] ———, *Heat conduction within linear thermoelasticity*, in Springer Tracts in Natural Philosophy, 30, Springer-Verlag, Berlin, 1985.

[8] B. F. ESHAM AND R. J. WEINACHT, *Hyperbolic-parabolic singular perturbations for quasi-linear equations*, SIAM J. Math. Anal., 20 (1989), pp. 1344–1365.

[9] G. C. HSIAO AND R. J. WEINACHT, *A singularly perturbed Cauchy problem*, J. Math. Anal. Appl., 71 (1979), pp. 242–250.

[10] J. KEVORKIAN AND J. D. COLE, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1537–1553, November 1994

006

ERROR BOUNDS FOR ASYMPTOTIC EXPANSIONS OF LAPLACE CONVOLUTIONS*

X. LI† AND R. WONG‡

Abstract. Asymptotic expansions are derived for the Laplace convolution $(f*g)(x)$ as $x \to \infty$, where f and g have asymptotic power series representation in descending powers of t. Bounds are also constructed for the error terms associated with these expansions. Similar results are given for the convolution integrals

$$\int_0^\infty f(t)g(x+t)dt \quad \text{and} \quad \int_0^\infty f(t)g(x-t)dt$$

as $x \to \infty$. These results can be used in the study of asymptotic solutions to the renewal equation and the Wiener–Hopf equations.

Key words. convolution integrals, asymptotic expansions, error bounds, integral equations

AMS subject classifications. 41A60, 45E10

1. Introduction.

Let $f(t)$ and $g(t)$ be locally integrable functions on $[0,\infty)$. The convolution integral

$$(f*g)(x) = \int_0^x f(x-t)g(t)dt, \qquad x > 0, \tag{1.1}$$

occurs frequently in Laplace transform theory [1], [10] and Volterra integral equations [6]. Asymptotic behavior of this integral, as $x \to \infty$, has been investigated by Riekstiņš [9] and Handelsman and Lew [2], under the condition that $f(t)$ and $g(t)$ have asymptotic series representations in descending powers of t near $t = \infty$. For simplicity, let us assume that

$$f(t) \sim \sum_{s=0}^{\infty} a_s t^{-s-\alpha}, \qquad 0 < \alpha \le 1, \tag{1.2}$$

and

$$g(t) \sim \sum_{s=0}^{\infty} b_s t^{-s-\beta}, \qquad 0 < \beta \le 1, \tag{1.3}$$

as $t \to \infty$. The results of Riekstiņš and Handelsman and Lew establish the existence of asymptotic expansions of $(f*g)(x)$ in certain forms, but these expansions involve exponents and coefficients that are not explicitly known. To be of any practical use, these quantities should be determined explicitly. The purpose of this paper is to provide formulas for these quantities, and to construct computable bounds for the

* Received by the editors March 22, 1993; accepted for publication (in revised form) August 10, 1993.

† Department of Applied Mathematics, Tsinghua University, Beijing, China.

‡ Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2. The research of this author was supported by Natural Sciences and Engineering Council of Canada grant A7359 and National Science Council of the Republic of China grant NSC 81-0208-M-001-15.

error terms associated with these expansions. For instance, in the case $0 < \alpha, \beta < 1$, it will be shown that

$$(f * g)(x) = \sum_{s=0}^{n-1} A_s x^{-\alpha-s} + \sum_{s=0}^{n-1} B_s x^{-\beta-s} + \sum_{s=0}^{n-1} C_s x^{1-\alpha-\beta-s} + R_n(x), \tag{1.4}$$

where the remainder satisfies

$$| R_n(x) | \leq Q_n x^{1-\alpha-\beta-n} \tag{1.5}$$

for $x > 1$. The coefficients A_s, B_s and C_s in (1.4) are given explicitly in (4.15)–(4.17) below, and the constant Q_n is also given; cf. (5.3).

Our approach is based on the distributional method introduced by McClure and Wong to obtain similar results for the Stieltjes transform [4] and the Riemann–Liouville fractional integral [5]. For convenience, we include in §2 a brief summary of some of the facts concerning distributions and convolutions given in these two references. In §3, we introduce the noncommutative convolution product

$$(f \odot g)(x) \equiv \int_0^{x/2} f(x-t)g(t)dt, \tag{1.6}$$

and explain why we need such a product. Section 4 contains the derivation of (1.4), while the proof of (1.5) is given in §5. The case when one of the exponents α and β in (1.2)–(1.3) equals 1 is treated in §6. Section 7 deals with the case $\alpha = \beta = 1$. In the final section, we illustrate how the results in [12] and the present paper can be used to derive the asymptotic expansions of the convolution integrals

$$G^+(x) = \int_0^{\infty} f(t)g(x+t)dt, \qquad x > 0 \tag{1.7}$$

and

$$G^-(x) = \int_0^{\infty} f(t)g(x-t)dt, \qquad x > 0. \tag{1.8}$$

In (1.8), it is assumed that $g(-t) = g(t)$ for every $t \neq 0$. Asymptotic behavior of these two integrals has been considered previously by Muki and Sternberg [7] in a study of an integral equation. Our investigation was motivated by a study of the asymptotic behavior of solutions to the renewal equation [11]

$$u(t) = g(t) + \int_0^t f(t-\tau)u(\tau)d\tau, \tag{1.9}$$

where $f(t)$ is a probability density function on $[0, \infty)$, and the Wiener–Hopf equation [8]

$$u(t) = g(t) + \int_0^{\infty} f(t-\tau)u(\tau)d\tau, \tag{1.10}$$

where $f(-t) = f(t)$. In both (1.9) and (1.10) f and g satisfy (1.2) and (1.3), respectively. We will defer this problem to another investigation.

2. Distributions and convolutions. Let I be an open interval (finite or infinite) in the real line $\mathbb{R}$, and let $\mathfrak{D}(I)$ be the test function space of all C^∞ functions $\varphi(t)$ with compact support in I. A *distribution* $\wedge$ on I is a continuous linear functional on $\mathfrak{D}(I)$. We write $\langle \wedge, \varphi \rangle$ for the action of a distribution $\wedge$ on a test function φ. Two

distributions $\wedge_1$ and $\wedge_2$, not necessarily with the same domain, are said to be *equal* on an interval I if $\langle \wedge_1, \varphi \rangle = \langle \wedge_2, \varphi \rangle$ for all $\varphi \in \mathfrak{D}(I)$.

Let $L_{loc}(I)$ denote the space of locally integrable functions on I. Each $f \in L_{loc}(I)$ defines a distribution f in I by

$$\langle f, \varphi \rangle = \int_I f(t)\varphi(t)dt, \qquad \varphi \in \mathfrak{D}(I). \tag{2.1}$$

If $\wedge$ is a distribution whose domain includes I, then we say that $\wedge$ belongs to $L_{loc}(I)$ if there is a function $f \in L_{loc}(I)$ such that $\langle \wedge, \varphi \rangle = \langle f, \varphi \rangle$ for all $\varphi \in \mathfrak{D}(I)$.

Let $\wedge$ be a distribution in I. Its derivative $D\wedge$ is defined by

$$\langle D\wedge, \varphi \rangle = -\langle \wedge, \varphi' \rangle, \qquad \varphi \in \mathfrak{D}(I),$$

where $\varphi'(t)$ denotes the ordinary derivative of φ. If $f \in L_{loc}(I)$, then we write Df for the distributional derivative of f. If f' also exists (in the usual sense) almost everywhere and is locally integrable in I, then f' also defines a distribution in the sense of (2.1). It is well known that the two derivatives are not always the same; see, e.g., [13, p. 250]. However, we have the following lemma [5]. (For a proof of this result, see [13, p. 251].) Recall that a function f is said to be *locally absolutely continuous* in I if it is absolutely continuous in every compact subinterval of I.

LEMMA 1. *Suppose $f \in L_{loc}(I)$. Then $Df \in L_{loc}(I)$ if and only if f is (equal a.e. to a) locally absolutely continuous (function) in I, and in that case $Df = f'$.*

Let $L^+_{loc}(\mathbb{R})$ denote the class of functions which are locally integrable on $\mathbb{R}$ and which vanish on $(-\infty, 0)$. The distributions associated with functions in $L^+_{loc}(\mathbb{R})$ all have support contained in $[0, \infty)$. From Lemma 1, we immediately have the following result; *see [13, p. 255].*

COROLLARY. *Suppose $g \in L^+_{loc}(\mathbb{R})$. Then $Dg \in L^+_{loc}(\mathbb{R})$ if and only if g is equal a.e. to a locally absolutely continuous function h in $\mathbb{R}$. In particular, $Dg \in L^+_{loc}(\mathbb{R})$ implies $h(0) = 0$.*

If f and g belong to $L^+_{loc}(\mathbb{R})$, then equation (1.1), i.e.,

$$(f * g)(x) = \int_0^x f(x-t)g(t)dt,$$

defines $f * g \in L^+_{loc}(\mathbb{R})$. We note in passing that $f * g$ is continuous at 0, and at every point of continuity of either f or g. This equation may also be regarded as the definition of the convolution of two distributions, each belonging to $L^+_{loc}(\mathbb{R})$. Now we consider distributions of the form $D^n f$, where n is a nonnegative integer and $f \in L^+_{loc}(\mathbb{R})$. Clearly, the Heaviside function H, the Dirac delta function $\delta = DH$, and $t^\mu H(t)$ are all of this form, where μ is any real (or complex) number. Throughout the remaining portion of the paper, we shall write t^μ for the last mentioned distribution, taking it as understood that the distribution vanishes on $(-\infty, 0)$. In [3, §6.5], Jones introduced the definition

$$D^n f * D^m g = D^{n+m}(f * g). \tag{2.2}$$

Since $f * g \in L^+_{loc}(\mathbb{R})$, the distributional derivative $D^{n+m}(f * g)$ on the right-hand side of (2.2) is well defined. It was this definition that led McClure and Wong [5] to the construction of an exact remainder for the asymptotic expansion of the Riemann–Liouville fractional integral.

3. The convolution $\odot$. Returning to (1.2) and (1.3), we write

$$f(t) = \sum_{s=0}^{n-1} a_s t^{-s-\alpha} + f_n(t) \tag{3.1}$$

and

$$g(t) = \sum_{s=0}^{n-1} b_s t^{-s-\beta} + g_n(t) \tag{3.2}$$

for each $n \geq 1$, and let $f_0(t) = f(t)$ and $g_0(t) = g(t)$. As in [4], we define inductively $f_{n,0}(t) = f_n(t)$ and

$$\begin{aligned} f_{n,j+1}(t) &= -\int_t^\infty f_{n,j}(\tau)d\tau \\ &= \frac{(-1)^{j+1}}{j!}\int_t^\infty (\tau - t)^j f_n(\tau)d\tau, \end{aligned} \tag{3.3}$$

$j = 0, 1, \ldots, n-1$. All these functions exist and are absolutely continuous on $[t, \infty)$ for each $t > 0$. Furthermore, if $0 < \alpha < 1$ then $f_{n,n}(t)$ is bounded on $[0, R]$ for any $R > 0$ and $O(t^{-\alpha})$ as $t \to \infty$. If $\alpha = 1$ then $f_{n,n}(t) = O(t^{-1})$ as $t \to \infty$ and $f_{n,n}(t) = O(|\log t|)$ as $t \to 0^+$.

Except for constant factors, each function $t^{-s-\alpha}$ in (3.1) defines a distribution, vanishing on $(-\infty, 0)$, as a derivative of $t^{-\alpha}$; f_n defines a distribution $D^n f_{n,n}$, also vanishing on $(-\infty, 0)$. In [4], it was proved that if $0 < \alpha < 1$ then these distributions are related by the equation

$$f = \sum_{s=0}^{n-1} a_s t^{-s-\alpha} - \sum_{s=1}^{n} c_s D^{s-1}\delta + f_n, \tag{3.4}$$

where

$$c_s = \frac{(-1)^s}{(s-1)!}\int_0^\infty t^{s-1} f_s(t)dt = \frac{(-1)^s}{(s-1)!} M[f; s], \tag{3.5}$$

$M[f; z]$ being the Mellin transform of f defined by

$$M[f; z] = \int_0^\infty t^{z-1} f(t)dt \tag{3.6}$$

or its analytic continuation. For the second equality in (3.5), see Lemma 7 in [13, p.173]. If $0 < \beta < 1$, then the corresponding result for (3.2) is

$$g = \sum_{s=0}^{n-1} b_s t^{-s-\beta} - \sum_{s=1}^{n} d_s D^{s-1}\delta + g_n, \tag{3.7}$$

where

$$d_s = \frac{(-1)^s}{(s-1)!}\int_0^\infty t^{s-1} g_s(t)dt = \frac{(-1)^s}{(s-1)!} M[g; s]. \tag{3.8}$$

Since the convolution product $*$ in (2.2) is distributive, taking the convolution of f and g given in (3.4) and (3.7) leads to the distribution $f_n * g_n$ By definition (2.2),

$$f_n * g_n = D^{2n}(f_{n,n} * g_{n,n}). \tag{3.9}$$

Now $f_{n,n} * g_{n,n}$ is a distribution in $(0,\infty)$ given by the locally integrable function

$$(f_{n,n} * g_{n,n})(x) = \int_0^x f_{n,n}(x-t)g_{n,n}(t)dt = x\int_0^1 f_{n,n}[x(1-u)]g_{n,n}(xu)du. \tag{3.10}$$

To differentiate this function under the integral sign, we recall the following result stated in [5]. (For a proof of this result, see [13, p. 327].)

LEMMA 2. *Suppose that $f(t,x)$ is integrable in a compact rectangle $[a,b]\times[c,d]$, and absolutely continuous as a function of x, for each $t\in[a,b]$. Suppose also that $\partial f(t,x)/\partial x$ is integrable in $[a,b]\times[c,d]$. Then the function $F(x)$ defined on $[c,d]$ by $F(x)=\int_a^b f(t,x)dt$ is absolutely continuous in $[c,d]$, and*

$$F'(x) = \int_a^b \frac{\partial}{\partial x} f(t,x)dt.$$

Since $f_{n,j}$ and $g_{n,j}$ are locally absolutely continuous in $(0,\infty)$ for $j=1,\dots,n$, the function $(f_{n,n}*g_{n,n})(x)$ in (3.10) can be differentiated n (and only n) times under integral signs. But equation (3.9) requires $2n$ times differentiation. Therefore, a straightforward extension of the method given in [5] does not work, and we shall make the following necessary modification.

The integral in (1.1) can be written as

$$(f*g)(x) = \int_0^{x/2} f(x-t)g(t)dt + \int_0^{x/2} g(x-t)f(t)dt. \tag{3.11}$$

This device is due to Riekstiņš [9]. In terms of the convolution product $\odot$, (3.11) becomes

$$(f*g)(x) = (f\odot g)(x) + (g\odot f)(x). \tag{3.12}$$

This formula holds for any f and g in $L^+_{loc}(\mathbb{R})$. The advantage of the convolution $\odot$ over the convolution $*$ is that the function $f(t)$ in (1.6) need be locally integrable only on the open interval $(0,\infty)$, and not on the half-closed interval $[0,\infty)$ as required in (1.1). Thus, equation (1.6) may be used to also define the convolution $f\odot g$, where f is a distribution in $L_{loc}(0,\infty)$ and g is a distribution in $L^+_{loc}(\mathbb{R})$. We wish to extend this to a definition of convolutions of the form $f\odot D^m g$, where m is a nonnegative integer.

Let f and g be two $(m-1)$-times differentiable functions in $(0,\infty)$, and let $g\in L^+_{loc}(\mathbb{R})$. For $x>0$, we define

$$f\odot D^m g = D^m(f\odot g) + \frac{1}{2}\sum_{\ell=0}^{m-1}\left[f\left(\frac{x}{2}\right)g^{(\ell)}\left(\frac{x}{2}\right)\right]^{(m-1-\ell)} \tag{3.13}$$

The motivation for this definition, and the answer to the obvious consistency question, are provided by the following lemma.

LEMMA 3. *If f,g and Dg all belong to $L^+_{loc}(\mathbb{R})$, then $f\odot g$ and $g\odot f$ are both locally absolutely continuous, and we have*

$$D(f\odot g) = f\odot Dg - \frac{1}{2}f\left(\frac{x}{2}\right)g\left(\frac{x}{2}\right) \tag{3.14}$$

and

$$D(g \odot f) = Dg \odot f + \frac{1}{2} f\left(\frac{x}{2}\right) g\left(\frac{x}{2}\right). \tag{3.15}$$

Proof. Since $Dg \in L^+_{loc}(\mathbb{R})$, by the corollary to Lemma 1 there exists a locally absolutely continuous function $h(x)$ such that $g(x) = h(x)$ almost everywhere in $\mathbb{R}$, and $h(0) = 0$. By Lemma 1, we also have $Dg = g' = h'$, and $h' \in L^+_{loc}(\mathbb{R})$. Thus, for $x \geq 0$ we have

$$\begin{aligned}\int_0^x (f \odot h')(t)dt &= \int_0^{x/2} h'(\tau) \int_{2\tau}^x f(t-\tau)dtd\tau \\ &= \int_0^{x/2} [f(x-\tau) + f(\tau)]g(\tau)d\tau.\end{aligned}$$

The last equality is obtained by integration by parts. Equation (1.6) then gives

$$\int_0^x (f \odot h')(t)dt = (f \odot g)(x) + \int_0^{x/2} f(\tau)g(\tau)d\tau. \tag{3.16}$$

Since h (and hence g) is locally absolutely continuous, g is locally bounded. Furthermore, since f is locally integrable, the integral on the right exists. Equation (3.16) implies that $f \odot g$ is locally absolutely continuous, and that

$$f \odot g' = f \odot h' = (f \odot g)' + \frac{1}{2} f\left(\frac{x}{2}\right) g\left(\frac{x}{2}\right).$$

By Lemma 1, $D(f \odot g) = (f \odot g)'$ and $Dg = g'$. This proves equation (3.14).

Using a similar argument, it can be shown that $g \odot f$ is locally absolutely continuous. Since $Dg \in L^+_{loc}(\mathbb{R})$, replacing f and g by Dg and f in (3.12), respectively, gives

$$Dg * f = Dg \odot f + f \odot Dg,$$

which, in view of the definition of $*$ in (2.2), is equivalent to

$$D(g * f) = Dg \odot f + f \odot Dg. \tag{3.17}$$

Coupling (3.14) and (3.17), we obtain

$$\begin{aligned}Dg \odot f &= D(g * f) - f \odot Dg \\ &= D(g * f) - D(f \odot g) - \frac{1}{2} f\left(\frac{x}{2}\right) g\left(\frac{x}{2}\right) \\ &= D(g \odot f) - \frac{1}{2} f\left(\frac{x}{2}\right) g\left(\frac{x}{2}\right).\end{aligned}$$

The last equality follows from (3.12). This proves equation (3.15). □

Using the fact that $\delta = DH$, it can be shown directly from (3.13) that

$$f \odot D^m \delta = D^m f \tag{3.18}$$

for any $f(x)$ which vanishes on $(-\infty, 0)$ and has a locally absolutely continuous mth derivative on $(0, \infty)$. In particular, we have

$$t^{-s-\alpha} \odot D^j \delta = (-1)^j (\alpha + s)_j t^{-s-j-\alpha}, \tag{3.19}$$

where s and j are nonnegative integers and $0 < \alpha \leq 1$. In (3.19) we have used the Pochhammer notation

$$(\gamma)_0 = 1, \qquad (\gamma)_n = \gamma(\gamma+1)\cdots(\gamma+n-1), \quad n = 1, 2, \ldots .$$

The incomplete Beta integral is defined by

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt,$$

where $a > 0, b > 0$ if $x \geq 1$ and b can be negative or zero if $x < 1$. In terms of this integral, it is evident that for $0 < \alpha \leq 1, 0 < \beta < 1$ and s any nonnegative integer, we have

$$t^{-\alpha-s} \odot t^{-\beta} = B_{\frac{1}{2}}(1-\beta, 1-\alpha-s)t^{1-\alpha-\beta-s}.$$

Using the fact that

$$t^{-\beta-j} = \frac{(-1)^j}{(\beta)_j} D^j t^{-\beta},$$

it can be shown from the definition (3.13) that

$$t^{-\alpha-s} \odot t^{-\beta-j} = e_{s,j}(\alpha,\beta)t^{1-\alpha-\beta-s-j} \tag{3.20}$$

in $(0,\infty)$, where

$$\begin{aligned} e_{s,j}(\alpha,\beta) = {} & \frac{(\alpha+\beta+s-1)_j}{(\beta)_j} B_{\frac{1}{2}}(1-\beta, 1-\alpha-s) \\ & - \sum_{\ell=0}^{j-1} 2^{\alpha+\beta+s+\ell-1} \frac{(\alpha+\beta+s+\ell)_{j-1-\ell}}{(\beta+\ell)_{j-\ell}}. \end{aligned} \tag{3.21}$$

If $\beta = 1$, then we use the facts that

$$t^{-1-j} = \frac{(-1)^j}{j!} D^{j+1} \log t$$

and

$$t^{-\alpha-s} \odot \log t = [\xi_s(\alpha)\log t + \eta_s(\alpha)]t^{1-\alpha-s},$$

where $\xi_0(1) = \log 2$,

$$\xi_s(\alpha) = \frac{1}{\alpha-1+s}(2^{\alpha-1+s} - 1) \tag{3.22}$$

for $0 < \alpha < 1, s = 0, 1, \ldots$, or $\alpha = 1, s = 1, 2, \ldots$, and

$$\eta_s(\alpha) = \int_0^{\frac{1}{2}} (1-u)^{-\alpha-s} \log u du = \frac{d}{da} B_{\frac{1}{2}}(a, 1-\alpha-s)\Big|_{a=1}, \tag{3.23}$$

for $0 < \alpha \leq 1$ and $s = 0, 1, \ldots$. The definition (3.13) then gives

$$t^{-\alpha-s} \odot t^{-1-j} = [h_{s,j}(\alpha)\log t + k_{s,j}(\alpha)]t^{-\alpha-s-j} \tag{3.24}$$

in $(0,\infty)$, where

$$h_{s,j}(\alpha) = \frac{(\alpha+s)_j}{j!} 2^{\alpha-1+s} - \frac{(\alpha-1+s)_{j+1}}{j!} \xi_s(\alpha) \tag{3.25}$$

and

(3.26)

$$k_{s,j}(\alpha) = (j+1)\frac{(\alpha-1+s)_j}{j!}\xi_s(\alpha) - \frac{(\alpha-1+s)_{j+1}}{j!}\eta_s(\alpha) - \frac{(\alpha+s)_j}{j!}2^{\alpha-1+s}\log 2 + \sum_{\ell=0}^{j-1}\left[\binom{j+1}{\ell}\frac{(\alpha-1+s)_\ell}{(j+1-\ell)_\ell}\xi_s(\alpha) - \frac{(\alpha+1+s+\ell)_{j-\ell-1}}{(\ell+1)_{j-\ell}}2^{\alpha+s+\ell} - \binom{j}{\ell}\frac{(\alpha+s)_\ell}{(j-\ell)_{\ell+1}}2^{\alpha-1+s}\right], \qquad 0 < \alpha \le 1.$$

To derive the expansion in (1.4), we need one further preliminary result.

LEMMA 4. *Let g_1 and g_2 be in $L^+_{loc}(\mathbb{R})$, and suppose that g_1 is n-times differentiable and g_2 is m-times differentiable in $(0,\infty)$. Let $N = \max\{n,m\}$, and suppose that f is N-times differentiable in $(0,\infty)$. Then we have*

$$f \odot (D^n g_1 + D^m g_2) = f \odot D^n g_1 + f \odot D^m g_2 \tag{3.27}$$

in $(0,\infty)$, i.e., the convolution $\odot$ is left-distributive.

Proof. Without loss of generality, we assume that $n \ge m$, and write $n = m+k$, $k \ge 0$. Then

$$D^n g_1 + D^m g_2 = D^n(g_1 + g_2^{(-k)}),$$

where $g_2^{(-k)}(x)$ is the kth iterated integral of $g_2(x)$, i.e.,

$$g_2^{(-k)}(x) = \frac{1}{(k-1)!}\int_0^x (x-t)^{k-1} g_2(t)\,dt.$$

Let $h = g_1 + g_2^{(-k)}$. Then $f \odot (D^n g_1 + D^m g_2) = f \odot D^n h$, and $h \in L^+_{loc}(\mathbb{R})$. By definition (3.13),

$$f \odot (D^n g_1 + D^m g_2) = D^n(f \odot h) + \frac{1}{2}\sum_{\ell=0}^{n-1}\left[f\left(\frac{x}{2}\right)h^{(\ell)}\left(\frac{x}{2}\right)\right]^{(n-1-\ell)}.$$

In terms of g_1 and $g_2^{(-k)}$, this becomes

(3.28)

$$f \odot (D^n g_1 + D^m g_2) = D^n(f \odot g_1) + D^n(f \odot g_2^{(-k)}) + \frac{1}{2}\sum_{\ell=0}^{n-1}\left[f\left(\frac{x}{2}\right)g_1^{(\ell)}\left(\frac{x}{2}\right)\right]^{(n-1-\ell)} + \frac{1}{2}\sum_{\ell=0}^{n-1}\left[f\left(\frac{x}{2}\right)g_2^{(\ell-k)}\left(\frac{x}{2}\right)\right]^{(n-1-\ell)}.$$

By re-indexing, we have

$$\sum_{\ell=0}^{n-1}\left[f\left(\frac{x}{2}\right)g_2^{(\ell-k)}\left(\frac{x}{2}\right)\right]^{(n-1-\ell)} = \sum_{j=0}^{m-1}\left[f\left(\frac{x}{2}\right)g_2^{(j)}\left(\frac{x}{2}\right)\right]^{(m-1-j)} + \sum_{i=1}^{k}\left[f\left(\frac{x}{2}\right)g_2^{(-i)}\left(\frac{x}{2}\right)\right]^{(m-1+i)}. \tag{3.29}$$

The Leibniz rule gives

$$D^n(f \odot g_2^{(-k)}) = D^m(f \odot g_2) - \frac{1}{2}\sum_{i=1}^{k}\left[f\left(\frac{x}{2}\right)g_2^{(-i)}\left(\frac{x}{2}\right)\right]^{(m+i-1)}. \tag{3.30}$$

In (3.29) and (3.30), empty sums are understood to be zero. Inserting (3.29) and (3.30) in (3.28), the final result (3.27) now follows from (3.13). □

4. Derivation of (1.4). From (3.1), we have

$$(f \odot g)(x) = \sum_{s=0}^{n-1} a_s (t^{-s-\alpha} \odot g)(x) + (f_n \odot g)(x) \tag{4.1}$$

for $0 < x < \infty$. Each of the convolutions in (4.1) exists as an ordinary integral in the form (1.6), and defines a locally integrable function in $(0, \infty)$. By Lemma 4, the convolution product $\odot$ in (3.13) is left-distributive. Hence for each $s = 0, 1, \dots, n-1$, we have from (3.7)

$$t^{-s-\alpha} \odot g = \sum_{j=0}^{n-1} b_j t^{-s-\alpha} \odot t^{-\beta-j} - \sum_{j=0}^{n-1} d_{j+1} t^{-s-\alpha} \odot D^j \delta + t^{-s-\alpha} \odot g_n, \tag{4.2}$$

which, combined with (3.19) and (3.20), gives

$$\begin{aligned} t^{-s-\alpha} \odot g = &\sum_{j=0}^{n-1} b_j e_{s,j}(\alpha, \beta) t^{1-\alpha-\beta-s-j} \\ &+ \sum_{j=0}^{n-1} (-1)^{j+1} d_{j+1} (\alpha + s)_j t^{-s-j-\alpha} + t^{-s-\alpha} \odot g_n. \end{aligned} \tag{4.3}$$

Since the distribution defined by g_n is $D^n g_{n,n}$, by (3.13)

$$t^{-s-\alpha} \odot g_n = D^n (t^{-s-\alpha} \odot g_{n,n}) + \frac{1}{2} \sum_{\ell=0}^{n-1} \left[\left(\frac{t}{2} \right)^{-s-\alpha} g_{n,n-\ell} \left(\frac{t}{2} \right) \right]^{(n-1-\ell)}. \tag{4.4}$$

Note that $g_{n,n}$ is locally integrable on $[0, \infty)$. Thus, from (1.6) we have

$$(t^{-s-\alpha} \odot g_{n,n})(x) = x^{1-\alpha-s} \int_0^{\frac{1}{2}} (1-u)^{-\alpha-s} g_{n,n}(xu) du. \tag{4.5}$$

Since $g_{n,j}$ is locally absolutely continuous in $(0, \infty)$ for $j = 1, \dots, n$, by Lemma 2 we can differentiate (4.5) under the integral sign n times to obtain

$$\begin{aligned} &(t^{-s-\alpha} \odot g_{n,n})^{(n)}(x) \\ &= \sum_{j=0}^{n} \binom{n}{j} (-1)^j (\alpha - 1 + s)_j x^{1-\alpha-s-j} \int_0^{\frac{1}{2}} (1-u)^{-\alpha-s} u^{n-j} g_{n,j}(xu) du \\ &\equiv x^{-n} \varepsilon_s^{(1)}(x), \end{aligned} \tag{4.6}$$

almost everywhere in $(0, \infty)$. We have written the last expression in the form $x^{-n}\varepsilon_s^{(1)}(x)$ in order to indicate that the rate of decay of $\varepsilon_s^{(1)}(x)$ is independent of n; see §5. Furthermore, since each function that has been differentiated is locally absolutely continuous, by Lemma 1

$$D^n (t^{-s-\alpha} \odot g_{n,n}) = (t^{-s-\alpha} \odot g_{n,n})^{(n)}(x) = x^{-n} \varepsilon_s^{(1)}(x). \tag{4.7}$$

Now each distribution in (4.3) is determined by a locally integrable function in $(0, \infty)$. By replacing these distributions by their corresponding functions, we obtain from (4.4)

and (4.7)

$$
\begin{aligned}
(t^{-s-\alpha} \odot g)(x) &= \sum_{j=0}^{n-1} b_j e_{s,j}(\alpha,\beta) x^{1-\alpha-\beta-s-j} \\
&\quad + \sum_{j=0}^{n-1} (-1)^{j+1} (\alpha+s)_j d_{j+1} x^{-s-j-\alpha} \\
&\quad + \frac{1}{2} \sum_{j=0}^{n-1} \left[\left(\frac{x}{2}\right)^{-s-\alpha} g_{n,n-j}\left(\frac{x}{2}\right) \right]^{(n-1-j)} + x^{-n} \varepsilon_s^{(1)}(x)
\end{aligned} \tag{4.8}
$$

holding in the sense of distributions in $(0,\infty)$. It is straightforward to show that this equation in fact holds pointwise almost everywhere in $(0,\infty)$. Since every convolution in (4.8) and (4.6) has at least one factor which is continuous in $(0,\infty)$, we conclude that (4.8) holds for each x in $(0,\infty)$. Substituting (4.8) in (4.1) gives

$$
\begin{aligned}
(f \odot g)(x) &= \sum_{s=0}^{n-1} \sum_{j=0}^{n-1} a_s b_j e_{s,j}(\alpha,\beta) x^{1-\alpha-\beta-s-j} \\
&\quad + \sum_{s=0}^{n-1} \sum_{j=0}^{n-1} (-1)^{j+1} (\alpha+s)_j a_s d_{j+1} x^{-s-j-\alpha} \\
&\quad + \frac{1}{2} \sum_{s=0}^{n-1} \sum_{j=0}^{n-1} a_s \left[\left(\frac{x}{2}\right)^{-\alpha-s} g_{n,n-j}\left(\frac{x}{2}\right) \right]^{(n-1-j)} \\
&\quad + \sum_{s=0}^{n-1} a_s \varepsilon_s^{(1)}(x) x^{-n} + (f_n \odot g)(x)
\end{aligned} \tag{4.9}
$$

for $0 < \alpha \le 1$ and $0 < \beta < 1$. Reversing the roles of f and g, we obtain a corresponding result for $(g \odot f)(x)$. Adding the two results together yields

$$
\begin{aligned}
(f * g)(x) &= \sum_{s=0}^{n-1} \sum_{j=0}^{n-1} a_s b_j [e_{s,j}(\alpha,\beta) + e_{j,s}(\beta,\alpha)] x^{1-\alpha-\beta-s-j} \\
&\quad + \sum_{s=0}^{n-1} \sum_{j=0}^{n-1} (-1)^{j+1} (\alpha+s)_j a_s d_{j+1} x^{-s-j-\alpha} \\
&\quad + \sum_{s=0}^{n-1} \sum_{j=0}^{n-1} (-1)^{j+1} (\beta+s)_j b_s c_{j+1} x^{-s-j-\beta} \\
&\quad + R_{n,4}(x) + R_{n,5}(x) + R_{n,6}(x),
\end{aligned} \tag{4.10}
$$

where

$$
R_{n,4}(x) = \frac{1}{2} \sum_{s=0}^{n-1} \sum_{j=0}^{n-1} \left[a_s \left(\frac{x}{2}\right)^{-\alpha-s} g_{n,n-j}\left(\frac{x}{2}\right) + b_s \left(\frac{x}{2}\right)^{-\beta-s} f_{n,n-j}\left(\frac{x}{2}\right) \right]^{(n-1-j)}, \tag{4.11}
$$

$$
R_{n,5}(x) = \sum_{s=0}^{n-1} [a_s \varepsilon_s^{(1)}(x) + b_s \varepsilon_s^{(2)}(x)] x^{-n}, \tag{4.12}
$$

$$R_{n,6}(x) = (f_n \odot g)(x) + (g_n \odot f)(x), \tag{4.13}$$

and $\varepsilon_n^{(2)}(x)$ is given by

(4.14)

$$x^{-n}\varepsilon_s^{(2)}(x) = \sum_{j=0}^{n} \binom{n}{j} (-1)^j (\beta - 1 + s)_j x^{1-\beta-s-j} \int_0^{\frac{1}{2}} (1-u)^{-\beta-s} u^{n-j} f_{n,j}(xu)du.$$

Each of the three double sums in (4.10) can be rearranged to give a single sum, and truncating each of these single sums after n terms gives the expansion (1.4) with

$$A_s = \sum_{j=0}^{s} (-1)^{j+1} (\alpha + s - j)_j a_{s-j} d_{j+1}, \tag{4.15}$$

$$B_s = \sum_{j=0}^{s} (-1)^{j+1} (\beta + s - j)_j b_{s-j} c_{j+1}, \tag{4.16}$$

and

$$C_s = \sum_{j=0}^{s} a_{s-j} b_j [e_{s-j,j}(\alpha, \beta) + e_{j,s-j}(\beta, \alpha)]. \tag{4.17}$$

The remainder $R_n(x)$ in (1.4) can be written as

$$R_n(x) = \sum_{i=1}^{6} R_{n,i}(x), \tag{4.18}$$

where

$$R_{n,1}(x) = \sum_{s=n}^{2n-2} \sum_{j=s-n+1}^{n-1} a_{s-j} b_j [e_{s-j,j}(\alpha, \beta) + e_{j,s-j}(\beta, \alpha)] x^{1-\alpha-\beta-s}, \tag{4.19}$$

$$R_{n,2}(x) = \sum_{s=n}^{2n-2} \sum_{j=s-n+1}^{n-1} (-1)^{j+1} (\alpha + s - j)_j a_{s-j} d_{j+1} x^{-s-\alpha}, \tag{4.20}$$

and

$$R_{n,3}(x) = \sum_{s=n}^{2n-2} \sum_{j=s-n+1}^{n-1} (-1)^{j+1} (\beta + s - j)_j b_{s-j} c_{j+1} x^{-s-\beta}. \tag{4.21}$$

The last three terms on the right-hand side of (4.18) are given in (4.11)–(4.13).

5. Proof of (1.5). First we estimate the terms $R_{n,1}(x)$, $R_{n,2}(x)$ and $R_{n,3}(x)$. By re-indexing, we have from (4.19)

$$R_{n,1}(x) = \sum_{s=0}^{n-2} \sum_{j=0}^{n-2-s} a_{n-j-1} b_{s+j+1} [e_{n-j-1,s+j+1}(\alpha, \beta) + e_{s+j+1,n-j-1}(\beta, \alpha)] x^{1-\alpha-\beta-n-s}.$$

Hence

$$| R_{n,1}(x) | \leq \gamma_{n,1} x^{1-\alpha-\beta-n},$$

where

$$\gamma_{n,1}(x) = \sum_{s=0}^{n-2} \sum_{j=0}^{n-2-s} \mid a_{n-j-1} b_{s+j+1} [e_{n-j-1,s+j+1}(\alpha, \beta) + e_{s+j+1,n-j-1}(\beta, \alpha)] \mid .$$

Similarly, we have

$$\mid R_{n,2}(x) \mid \leq \bar{A}_n x^{-\alpha-n}$$

and

$$\mid R_{n,3}(x) \mid \leq \bar{B}_n x^{-\beta-n},$$

where

$$\bar{A}_n = \sum_{s=0}^{n-2} \sum_{j=0}^{n-2-s} \mid a_{n-j-1} d_{s+j+2} (\alpha + n - j - 1)_{s+j+1} \mid$$

and

$$\bar{B}_n = \sum_{s=0}^{n-2} \sum_{j=0}^{n-2-s} \mid b_{n-j-1} c_{s+j+2} (\beta + n - j - 1)_{s+j+1} \mid .$$

Next we consider the terms $R_{n,4}(x)$, $R_{n,5}(x)$ and $R_{n,6}(x)$. As in [5], we *assume*

$$M_n(\alpha) = \sup_{(0,\infty)} \{t^{n+\alpha} \mid f_n(t) \mid\} < +\infty$$

and

$$N_n(\beta) = \sup_{(0,\infty)} \{t^{n+\beta} \mid g_n(t) \mid\} < +\infty.$$

These conditions do not follow from (1.2), (1.3), and the local integrability of f and g, but will be true in most applications. From (3.3), it is easily seen that for $j = 0, \ldots, n$,

$$\mid f_{n,j}(t) \mid \leq \frac{M_n(\alpha)}{(\alpha + n - j)_j} t^{-\alpha-n+j}. \tag{5.1}$$

Similarly, for $j = 0, \ldots, n$,

$$\mid g_{n,j}(t) \mid \leq \frac{N_n(\beta)}{(\beta + n - j)_j} t^{-\beta-n+j}. \tag{5.2}$$

Hence, from (4.6) and (4.14), we have

$$\mid \varepsilon_s^{(1)}(x) \mid \leq N_n(\beta) B_{\frac{1}{2}}(1 - \beta, 1 - \alpha - s) \left\{ \sum_{j=0}^{n} \binom{n}{j} \frac{\mid (\alpha + s - 1)_j \mid}{\mid (\beta + n - j)_j \mid} \right\} x^{1-\alpha-\beta-s}$$

and

$$\mid \varepsilon_s^{(2)}(x) \mid \leq M_n(\alpha) B_{\frac{1}{2}}(1 - \alpha, 1 - \beta - s) \left\{ \sum_{j=0}^{n} \binom{n}{j} \frac{\mid (\beta + s - 1)_j \mid}{\mid (\alpha + n - j)_j \mid} \right\} x^{1-\alpha-\beta-s}.$$

Coupling the last two estimates gives

$$\mid R_{n,5}(x) \mid \leq \gamma_{n,5} x^{1-\alpha-\beta-n}$$

for $x > 1$, where

$$\gamma_{n,5} = \sum_{s=0}^{n-1}\sum_{j=0}^{n}\binom{n}{j}\Big\{N_n(\beta) \mid a_s \mid \frac{\mid (\alpha+s-1)_j \mid}{\mid (\beta+n-j)_j \mid} B_{\frac{1}{2}}(1-\beta, 1-\alpha-s) + M_n(\alpha) \mid b_s \mid \frac{\mid (\beta+s-1)_j \mid}{\mid (\alpha+n-j)_j \mid} B_{\frac{1}{2}}(1-\alpha, 1-\beta-s)\Big\}.$$

To estimate $R_{n,4}(x)$, we carry out the differentiation in (4.11). The result is

$$R_{n,4}(x) = \sum_{s=0}^{n-1}\Big\{\sum_{\ell=0}^{n-1}\sum_{j=0}^{n-1-\ell}(-1)^j 2^{-n+\ell}\binom{n-\ell-1}{j}\Big[a_s(\alpha+s)_j\Big(\frac{x}{2}\Big)^{-\alpha-j} g_{n,j+1}\Big(\frac{x}{2}\Big) + b_s(\beta+s)_j\Big(\frac{x}{2}\Big)^{-\beta-j} f_{n,j+1}\Big(\frac{x}{2}\Big)\Big]\Big\}\Big(\frac{x}{2}\Big)^{-s}.$$

From (5.1) and (5.2), it follows that

$$\mid R_{n,4}(x) \mid \leq \gamma_{n,4} x^{1-\alpha-\beta-n},$$

where

$$\gamma_{n,4} = \sum_{s=0}^{n-1}\Big\{\sum_{\ell=0}^{n-1}\sum_{j=0}^{n-1-\ell} 2^{\alpha+\beta+\ell+s-1}\binom{n-\ell-1}{j} \times \Big[\frac{\mid a_s \mid (\alpha+s)_j}{(\beta+n-j-1)_{j+1}} N_n(\beta) + \frac{\mid b_s \mid (\beta+s)_j}{(\alpha+n-j-1)_{j+1}} M_n(\alpha)\Big]\Big\}.$$

Since the convolutions $f_n \odot g$ and $g_n \odot f$ exist as ordinary integrals, their estimations are rather easy, and we have from (4.13)

$$\mid R_{n,6}(x) \mid \leq \gamma_{n,6} x^{1-\alpha-\beta-n},$$

where

$$\gamma_{n,6} = M_n(\alpha) N_0(\beta) B_{\frac{1}{2}}(1-\beta, 1-\alpha-n) + M_0(\alpha) N_n(\beta) B_{\frac{1}{2}}(1-\alpha, 1-\beta-n).$$

A combination of the estimates for $R_{n,i}(x), i = 1, \ldots, 6$, gives

$$\mid R_n(x) \mid \leq \bar{A}_n x^{-\alpha-n} + \bar{B}_n x^{-\beta-n} + \bar{C}_n x^{1-\alpha-\beta-n}, \tag{5.3}$$

where

$$\bar{C}_n = \gamma_{n,1} + \gamma_{n,4} + \gamma_{n,5} + \gamma_{n,6}.$$

The result (1.5) now follows from (5.3) by letting $Q_n = \bar{A}_n + \bar{B}_n + \bar{C}_n$.

6. The case $0 < \alpha < 1$ and $\beta = 1$. If one of the exponents α and β in (1.2) and (1.3) is equal to 1, then we may assume, without loss of generality, that $0 < \alpha < 1$ and $\beta = 1$. In this case, the incomplete Beta integral in (3.21) does not exist, and hence the expansion in (1.4) no longer holds. To derive the asymptotic expansion of $f * g$ in this case, we shall make use of (3.24), instead of (3.20). Since the argument here is similar to that in §4, we shall keep our discussion brief. The analogue of (3.7) is

$$g = \sum_{s=0}^{n-1} b_s t^{-s-1} - \sum_{s=1}^{n} d_s^* D^{s-1}\delta + g_n,$$

where

$$d_s^* = \lim_{t\to 0}\left[g_{s,s}(t) + \frac{(-1)^{s-1}}{(s-1)!}b_{s-1}\log t\right];$$

see [4]. (For a more tractable expression of d_s^*, see [13, p. 300, eqs. (2.32) and (2.34)].) From (3.24) and (3.19), it follows that

$$t^{-s-\alpha}\odot g = \sum_{j=0}^{n-1} b_j[h_{s,j}(\alpha)\log t + k_{s,j}(\alpha)]t^{-s-j-\alpha} + \sum_{j=0}^{n-1}(-1)^{j+1}(\alpha+s)_j d_{j+1}^* t^{-s-j-\alpha} + t^{-s-\alpha}\odot g_n.$$

Since the results (4.4) and (4.7) continue to hold when $\beta = 1$, by the same argument used for (4.9) we obtain

$$\begin{aligned}(f\odot g)(x) = {} & \sum_{s=0}^{n-1}\sum_{j=0}^{n-1} a_s b_j[h_{s,j}(\alpha)\log x + k_{s,j}(\alpha)]x^{-s-j-\alpha} \\ & + \sum_{s=0}^{n-1}\sum_{j=0}^{n-1}(-1)^{j+1}(\alpha+s)_j a_s d_{j+1}^* x^{-s-j-\alpha} \\ & + \frac{1}{2}\sum_{s=0}^{n-1}\sum_{j=0}^{n-1} a_s\left[\left(\frac{x}{2}\right)^{-\alpha-s} g_{n,n-j}\left(\frac{x}{2}\right)\right]^{(n-1-j)} \\ & + \sum_{s=0}^{n-1} a_s\varepsilon_s^{(1)}(x)x^{-n} + (f_n\odot g)(x)\end{aligned} \tag{6.1}$$

for all x in $(0,\infty)$. Furthermore, since (4.9) holds even when $\alpha = 1$, reversing the roles of f and g (and hence α and β) in this equation gives an expansion for $(g\odot g)(x)$. We now add the expansions for $f\odot g$ and $g\odot f$ together, and rearrange the terms as in §4. The final result is

$$(f * g)(x) = \sum_{s=0}^{n-1}(D_s\log x + E_s)x^{-\alpha-s} + \sum_{s=0}^{n-1} F_s x^{-1-s} + R_n(x), \tag{6.2}$$

where

$$D_s = \sum_{j=0}^{s} a_{s-j}b_j h_{s-j,j}(\alpha), \tag{6.3}$$

$$E_s = \sum_{j=0}^{s} a_{s-j}\{b_j[k_{s-j,j}(\alpha) + e_{j,s-j}(1,\alpha)] + (-1)^{j+1}(\alpha+s-j)_j d_{j+1}^*\}, \tag{6.4}$$

and

$$F_s = \sum_{j=0}^{s}(-1)^{j+1}(1+s-j)_j b_{s-j}c_{j+1}. \tag{6.5}$$

To estimate the remainder $R_n(x)$, we *assume*, again as in [5], that there exists $\beta' \in (0,1)$ such that

$$N_n(\beta') = \sup_{(0,\infty)} \{t^{n+\beta'} \mid g_n(t) \mid\} < +\infty.$$

For $x > 1$, it can be shown as before that there are computable constants P_n and Q_n such that

$$\mid R_n(x) \mid \leq (P_n \log x + Q_n) x^{1-\beta'-\alpha-n}. \tag{6.6}$$

Note that the actual error $R_n(x)$ in the expansion (6.2) is $O(x^{-\alpha-n} \log x)$. Hence the estimate (6.6) is slightly short of the actual result.

7. The case $\alpha = \beta = 1$. This case turns out to be simpler than one might have expected. This is mainly due to the split of the convolution integral $(f * g)(x)$ in (1.1) into two convolution integrals $(f \odot g)(x)$ and $(g \odot f)(x)$ of the form (1.6). The expansion for $(f \odot g)(x)$ given in (6.1) allows the possibility of $\alpha = \beta = 1$; cf. (3.24)–(3.26). By reversing the roles of f and g, the corresponding expansion for $(g \odot f)(x)$ also allows this possibility. For convenience, we set

$$h_{s,j} \equiv h_{s,j}(1) \quad \text{and} \quad k_{s,j} \equiv k_{s,j}(1).$$

Now let $\alpha = 1$ in (6.1), and write down the corresponding expansion for $(g \odot f)(x)$. Adding up the two expansions and rearranging the terms as before, we obtain

$$(f * g)(x) = \sum_{s=0}^{n-1} (G_s \log x + H_s) x^{-s-1} + R_n(x), \tag{7.1}$$

where

$$G_s = \sum_{j=0}^{s} a_{s-j} b_j (h_{s-j,j} + h_{j,s-j}), \tag{7.2}$$

$$H_s = \sum_{j=0}^{s} [a_{s-j} b_j (k_{s-j,j} + k_{j,s-j}) + (-1)^{j+1} (1 + s - j)_j (a_{s-j} d^*_{j+1} + b_{s-j} c^*_{j+1})], \tag{7.3}$$

and c^*_{j+1} has the same meaning as d^*_{j+1} given at the beginning of this section except that $g_{s,s}(t)$ and b_{s-1} are now replaced by $f_{s,s}(t)$ and a_{s-1}.

To obtain an error bound, one may assume that there exists $\rho \in (0,1)$ such that

$$\sup_{(0,\infty)} \{t^{n+\rho} \mid f_n(t) \mid\} < +\infty$$

and

$$\sup_{(0,\infty)} \{t^{n+\rho} \mid g_n(t) \mid\} < +\infty.$$

Under these conditions, it can again be shown that there are computable constants P_n and Q_n such that

$$\mid R_n(x) \mid \leq (P_n \log x + Q_n) x^{1-2\rho-n}$$

for $x > 1$.

8. The integrals in (1.7) and (1.8). In [7], Muki and Sternberg have constructed one-term asymptotic approximation for the integrals $G^+(x)$ and $G^-(x)$ under the conditions

$$g(t) = \frac{a}{t^m} + O\left(\frac{1}{t^{m+1}}\right), \quad a \neq 0, \quad 1 < m < \infty, \tag{8.1}$$

and

$$f(t) = \frac{b}{t^\mu} + o\left(\frac{1}{t^\mu}\right), \quad b \neq 0, \quad 1 \leq \mu < \infty. \tag{8.2}$$

For the case of $G^-(x)$, they also assumed that $g(-t) = g(t)$ for every $t \neq 0$. Now if f and g have the asymptotic expansions (1.2) and (1.3), then it is natural to expect that infinite asymptotic expansions can be derived for these integrals. This is indeed the case, and we shall show in this section how this can be done.

The asymptotic expansions of $G^+(x)$ can be easily obtained from the corresponding result for the generalized Stieltjes transform

$$S_f(x) = \int_0^\infty \frac{f(t)}{(t+x)^\rho} dt. \tag{8.3}$$

If $f(t)$ is locally absolutely integrable on $[0, \infty)$ and satisfies (1.2) then the integral (8.3) converges absolutely as long as $\alpha + \rho > 1$. Asymptotic expansions for $S_f(x)$ have been derived in [12].

We suppose that f and g satisfy (1.2) and (1.3), respectively. When $0 < \alpha < 1$, we can derive the asymptotic expansion

$$G^+(x) \sim \sum_{s=0}^\infty A_s^* x^{1-s-\alpha-\beta} + \sum_{s=0}^\infty B_s^* x^{-s-\beta}, \qquad x \to \infty, \tag{8.4}$$

where

$$A_s^* = \Gamma(s+\alpha+\beta-1) \sum_{i=0}^s a_i b_{s-i} \frac{\Gamma(1-i-\alpha)}{\Gamma(s-i+\beta)} \tag{8.5}$$

and

$$B_s^* = \Gamma(s+\beta) \sum_{i=0}^s (-1)^i \frac{b_{s-i}}{\Gamma(s-i+\beta)} M[f; i+1], \tag{8.6}$$

where $M[f; z]$ is the Mellin transform defined in (3.6).

If $\alpha = 1$ then we have

$$G^+(x) \sim \sum_{s=0}^\infty (C_s^* \log x + D_s^*) x^{-s-\beta}, \tag{8.7}$$

where

$$C_s^* = \Gamma(s+\beta) \sum_{i=0}^s \frac{(-1)^i}{i! \Gamma(s-i+\beta)} a_i b_{s-i} \tag{8.8}$$

and

$$D_s^* = \sum_{i=0}^s c_i (s-i+\rho) b_{s-i}. \tag{8.9}$$

The constant $c_i(\rho)$ is given by

$$c_s(\rho) = \frac{(-1)^s}{s!}\frac{\Gamma(\rho+s)}{\Gamma(\rho)}\{[\psi(s+1)-\psi(s+\rho)]a_s + a_s^*(\rho)\}, \tag{8.10}$$

$\psi(z) = \Gamma'(z)/\Gamma(z)$, and

$$a_s^*(\rho) = \lim_{z\to s+\rho}\left\{M[f; z-\rho+1] + \frac{a_s}{z-s-\rho}\right\}. \tag{8.11}$$

Although no explicit bounds are given for the error terms associated with the asymptotic expansions (8.4) and (8.7), it is evident from our analysis such bounds can indeed be constructed.

In view of the condition $g(t) = g(-t)$ for every $t \neq 0$, the integral $G^-(x)$ in (1.8) can be written as

$$G^-(x) = (f * g)(x) + \int_0^\infty g(t)f(x+t)dt. \tag{8.12}$$

The second term on the right-hand side is an integral of the form (1.7) which has just been considered. Hence, the asymptotic expansion of $G^-(x)$ can be obtained by adding up the corresponding results for the two integrals on the right of (8.12).

Acknowledgment. This research was completed while R. Wong was a visiting member of the Institute of Mathematics, Academia Sinica, Taiwan. He gratefully acknowledges the hospitality extended to him during his visit, and in particular thanks Professor Chen Ming-Po for making that visit possible.

REFERENCES

[1] G. DOETSCH, *Introduction to the Theory and Application of the Laplace Transformation*, Springer-Verlag, Berlin, 1974.

[2] R. A. HANDELSMAN AND J. S. LEW, *Asymptotic expansion of Laplace convolutions for large argument and tail densities for certain sums of random variables*, SIAM J. Math. Anal., 5 (1974), pp. 425–451.

[3] D. S. JONES, *The Theory of Generalized Functions*, Cambridge University Press, Cambridge, U.K., 1982.

[4] J. P. MCCLURE AND R. WONG, *Explicit error terms for asymptotic expansions of Stieltjes transforms*, J. Inst. Math. Appl., 22 (1978), pp. 129–145.

[5] ———, *Exact remainders for asymptotic expansions of fractional integrals*, J. Inst. Math. Appl., 24 (1979), pp. 139–147.

[6] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.

[7] R. MUKI AND E. STERNBERG, *Note on an asymptotic property of solutions to a class of Fredholm integral equations*, Quart. Appl. Math., 28 (1970), pp. 277–281.

[8] A. C. PIPKIN, *A Course in Integral Equations*, Springer-Verlag, New York, 1991.

[9] E. RIEKSTIŅŠ, *Asymptotic representation of certain types of convolution integrals*, Latvian Math. Year-Book, 8 (1970), pp. 223–239.

[10] I. H. SNEDDON, *The Use of Integral Transforms*, McGraw-Hill, New York, 1972.

[11] J. S. W. WONG AND R. WONG, *On asymptotic solutions of the Renewal equation*, J. Math. Anal. Appl., 53 (1976), pp. 243–250.

[12] R. WONG, *Explicit error terms for asymptotic expansions of Mellin convolutions*, J. Math. Anal. Appl., 72 (1979), pp. 740–756.

[13] ———, *Asymptotic Approximations of Integrals*, Academic Press, Boston, 1989.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1554–1561, November 1994

007

MULTIPLE SOLUTIONS FOR A NONLINEAR DIRICHLET PROBLEM*

ALFONSO CASTRO† AND JORGE COSSIO‡

Abstract. The authors prove that a semilinear elliptic boundary value problem has five solutions when the range of the derivative of the nonlinearity includes at least the first two eigenvalues. Extensive use is made of Lyapunov–Schmidt reduction arguments, the mountain pass lemma, and characterizations of the local degree of critical points.

Key words. nonlinear elliptic equations, multiplicity of solutions, local degree, mountain pass lemma

AMS subject classifications. 35J65, 35J20

1. Introduction.

Let $f\colon \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $f(0) = 0$, and

$$f'(\infty) = \lim_{|u|\to\infty} \frac{f(u)}{u} \in \mathbb{R}. \tag{1.1}$$

Let Ω be a smooth bounded region in $\mathbb{R}^n$, and Δ the Laplacian operator. Let $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_k \leq \ldots$ be the eigenvalues of $-\Delta$ with Dirichlet boundary condition in Ω.

The solvability of the boundary value problem

$$\begin{cases} \Delta u + f(u) = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.2}$$

has proven to be closely related to the position of the numbers $f'(0)$, $f'(\infty)$ with respect to the spectrum of $-\Delta$. In fact, Castro and Lazer in [11] showed that if the interval $(f'(0), f'(\infty)) \cup (f'(\infty), f'(0))$ contains the eigenvalues $\lambda_k, \ldots, \lambda_j$ and $f'(t) < \lambda_{j+1}$ for all $t \in \mathbb{R}$ then (1.2) has at least three solutions. The proofs in [11] are based on global Lyapunov–Schmidt arguments applied to variational problems. Subsequently Chang (see [12]) approached the same problems using Morse theory, and Hofer (see [14]) obtained the existence of five solutions when $f'(\infty) < \lambda_1$. For other results in the study of this problem we refer the reader to [3], [4], [6], [8], [10], [17], [18], and [19], among others.

Here we prove the following.

THEOREM A. *If* $f'(0) < \lambda_1$, $f'(\infty) \in (\lambda_k, \lambda_{k+1})$ *with* $k \geq 2$, *and* $f'(t) \leq \gamma < \lambda_{k+1}$, *then* (1.2) *has at least five solutions. Moreover, one of the following cases occur:*

(a) k *is even and* (1.2) *has two solutions that change sign.*

(b) k *is even and* (1.2) *has six solutions, three of which are of the same sign.*

(c) k *is odd and* (1.2) *has two solutions that change sign.*

* Received by the editors April 27, 1992; accepted for publication (in revised form) August 6, 1993. Partially supported by National Science Foundation grant DMS-8905936, the Texas Advanced Research Program, and Colciencias Grant 168-93.

† Department of Mathematics, University of North Texas, Denton, Texas 76203-5116 (acastro@unt.edu).

‡ Departamento de Matemáticas, Universidad Nacional de Colombia, Apartado Aéreo 3840, Medellín Colombia (jcossio@sigma.eafit.edu.co).

(d) *k is odd and* (1.2) *has three solutions of the same sign.*

The assumption $k \geq 2$ is sharp; Theorem B of [11] gives sufficient conditions for (1.2) to have exactly three solutions when $k = 1$. We prove Theorem A by using Lyapunov–Schmidt arguments to reduce the solvability of (1.2) to a finite-dimensional problem, and then we use degree and index theories applied to critical points. We make intensive use of the fact that the Leray–Schauder degree is invariant under the Lyapunov–Schmidt reduction process. In order to calculate various indices and degrees we prove that in large regions the Leray–Schauder degree of maps arising in problems like (1.2) where f' crosses the first eigenvalue

$$\left(\lim_{u \to -\infty} \frac{f(u)}{u} < \lambda_1 < \lim_{u \to \infty} \frac{f(u)}{u} \right)$$

is equal to zero. We also use "mountain pass arguments" of the Ambrosetti–Rabinowitz type (see [5]).

In §2 we recall the framework that allows studying solutions to (1.2) in terms of variational functionals and the Lyapunov–Schmidt reduction method. In §3 we calculate the index of the trivial solution when the nonlinearity crosses the first eigenvalue, establish the existence of positive and negative solutions, and compute their indices. In §4 we prove Theorem A.

2. Preliminaries and notation. First we state a global version of the Lyapunov–Schmidt method. For the sake of completeness we recall that if Φ is a functional of class C^1 and u_0 is a critical point of Φ then u_0 is called of mountain pass type if for every open neighborhood U of u_0 $\Phi^{-1}(-\infty, \Phi(u_0)) \cap U \neq \emptyset$ and $\Phi^{-1}(-\infty, \Phi(u_0)) \cap U$ is not path connected.

LEMMA 2.1. *Let M be a real separable Hilbert space. Let X and Y be closed subspaces of M such that $M = X \oplus Y$. Let $j: M \to \mathbb{R}$ be a functional of class C^1. If there are $m > 0$ and $\alpha > 1$ such that*

$$\langle \nabla j(x + y) - \nabla j(x + y_1), y - y_1 \rangle \geq m \|y - y_1\|^\alpha \quad \text{for all} \quad x \in X,\ y,\ y_1 \in Y \tag{2.1}$$

then we have the following.

(i) *There exists a continuous function $\psi: X \to Y$ such that*

$$j(x + \psi(x)) = \min_{y \in Y} j(x + y).$$

Moreover, $\psi(x)$ is the unique member of Y such that

$$\langle \nabla j(x + \psi(x)), y \rangle = 0 \quad \text{for all} \quad y \in Y. \tag{2.2}$$

(ii) *The function $\tilde{j}: X \to \mathbb{R}$ defined by $\tilde{j}(x) = j(x + \psi(x))$ is of class C^1, and*

$$\langle \nabla \tilde{j}(x), x_1 \rangle = \langle \nabla j(x + \psi(x)), x_1 \rangle \quad \text{for all} \quad x,\ x_1 \in X. \tag{2.3}$$

(iii) *An element $x \in X$ is a critical point of $\tilde{j}$ if and only if $x + \psi(x)$ is a critical point of j.*

(iv) *Let $\dim X < \infty$ and P be the projection onto X across Y. Let $S \subset X$ and $\Sigma \subset M$ be open bounded regions such that*

$$\{x + \psi(x); x \in S\} = \Sigma \cap \{x + \psi(x); x \in X\}.$$

If $\nabla \tilde{j}(x) \neq 0$ for $x \in \partial S$ then

$$d(\nabla \tilde{j}, S, 0) = d(\nabla j, \Sigma, 0),$$

where d denotes the Leray–Schauder degree.

(v) *If $u_0 = x_0 + \psi(x_0)$ is a critical point of mountain pass type of j then x_0 is a critical point of mountain pass type of $\tilde{j}$.*

Proof. The reader is referred to [9] for the proof of parts (i)–(iii). The proof of part (iv) follows by arguing as in Lemma 2.6 of [16]. Now we proceed with the proof of part (v).

Suppose x_0 is not of mountain pass type of $\tilde{j}$. Let V be an open neighborhood of x_0 in X such that either $\tilde{j}^{-1}(-\infty, \tilde{j}(x_0)) \cap V$ is empty or path connected. If $\tilde{j}^{-1}(-\infty, \tilde{j}(x_0)) \cap V$ is empty, by part (i) we see that $\{x + y; x \in V, y \in Y\} \cap j^{-1}(-\infty, j(u_0))$ is also empty. Thus u_0 is not of mountain pass type for j. On the other hand if $\tilde{j}^{-1}(-\infty, \tilde{j}(x_0)) \cap V$ is path connected, letting $W = \{x + y; x \in V, \|y - \psi(x)\| < 1\}$ and using again part (i) it is easily seen that $W \cap j^{-1}(-\infty, j(u_0))$ is also path connected. This concludes the proof of Lemma 2.1. □

For each positive integer m let φ_m denote an eigenfunction corresponding to the eigenvalue λ_m. Let H be the Sobolev space $H_0^1(\Omega)$ which is the completion of the inner product space consisting of real C^1 functions having support contained in Ω with inner product

$$\langle u, v \rangle = \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx.$$

As it is well known, the set $\{\varphi_m\}$ can be assumed to be complete and orthonormal in H.

We say that $u \in H$ is a weak solution to (1.2) if for every $\varphi \in H$

$$\int_\Omega (\nabla u . \nabla \varphi - f(u)\, \varphi)\, dx = 0$$

By standard regularity for elliptic operators (see [11]) it follows that weak solutions are classical solutions when f is continuous and sublinear, i.e., when f is continuous and there is a positive constant a such that

$$|f(u)| \le a(1 + |u|). \tag{2.4}$$

Let $J\colon H \to \mathbb{R}$ denote the functional defined by

$$J(u) = \int_\Omega \left(\frac{1}{2} \|\nabla u\|^2 - F(u) \right) dx, \tag{2.5}$$

where $F(\xi) = \int_0^\xi f(s)\, ds$. Since $f'(\infty) \in (\lambda_k, \lambda_{k+1})$, f satisfies (2.4). Thus $J \in C^1(H, \mathbb{R})$ (see [19]) and

$$\langle \nabla J(u), \varphi \rangle = \int_\Omega (\nabla u . \nabla \varphi - f(u)\, \varphi)\, dx \quad \text{for} \quad \varphi \in H. \tag{2.6}$$

In particular u is a weak solution of (1.2) if and only if u is a critical point of J.

Let X denote the subspace of H spanned by $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$, Y its orthogonal complement, and J the functional defined by (2.5). We claim J satisfies hypothesis (2.1). Indeed, from (2.6) and the mean value theorem

$$\langle \nabla J(x + y) - \nabla J(x + y_1), y - y_1 \rangle = \|y - y_1\|^2 - \int_\Omega f'(\xi)(y - y_1)^2. \tag{2.7}$$

Denoting by $\|\ \|_0$ the usual $L^2(\Omega)$ norm and using that $f'(\xi) \le \gamma < \lambda_{k+1}$, we have

$$\begin{aligned} \langle \nabla J(x+y) - \nabla J(x+y_1), y - y_1 \rangle &\ge \|y - y_1\|^2 - \gamma \|y - y_1\|_0^2 \\ &\ge \left(1 - \frac{\gamma}{\lambda_{k+1}}\right) \|y - y_1\|^2, \end{aligned} \tag{2.8}$$

where we have used that $\|z\|^2 \ge \lambda_{k+1}\|z\|_0^2$ for all $z \in Y$. Thus (2.1) holds with $m = 1 - \gamma/(\lambda_{k+1})$ and $\alpha = 2$.

3. Index of the trivial solution when the nonlinearity crosses the first eigenvalue. For $\gamma > \lambda_1$ let $p(\gamma) := p$ be the homogeneous function defined by

$$p(x) = \begin{cases} \gamma x & \text{for } x \ge 0, \\ f'(0)x & \text{for } x < 0. \end{cases}$$

Let P be the primitive of p with $P(0) = 0$, and $\pi\colon H \to \mathbb{R}$ be the functional defined by

$$\pi(u) = \int_\Omega \left(\frac{1}{2}\|\nabla u\|^2 - P(u)\right) dx. \tag{3.1}$$

As observed in §2 (see (2.4)) π is a functional of class C^1, and its critical points are the weak solutions to

$$\begin{cases} \Delta u + p(u) = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.2}$$

Because $f'(0) < \lambda_1$ and the principal eigenvalue of the Laplacian in any subregion of Ω is bigger than or equal to λ_1, we see that if $u \ne 0$ is a weak solution to (3.2) then u is a positive eigenfunction. Since this contradicts that $\gamma > \lambda_1$, we conclude that $u = 0$ is the only critical point of π.

LEMMA 3.1. *If B is a ball in H containing zero then $d(\nabla\pi, B, 0) = 0$.*

Proof. By the definition of the Leray–Schauder degree if Z denotes the subspace spanned by $\varphi_1, \varphi_2, \dots, \varphi_l$ with l big enough

$$d(\nabla\pi, B, 0) = d(P\nabla\pi, B \cap Z, 0), \tag{3.3}$$

where P denotes the orthogonal projection onto Z. Since $\gamma > \lambda_1$ we see that $h(t) := p(t) - \lambda_1 t > 0$ for $t \ne 0$. Because φ_1 is in Z we have

$$\begin{aligned} \langle P\nabla\pi(x), \varphi_1 \rangle &= \langle \nabla\pi(x), \varphi_1 \rangle \\ &= \int_\Omega (\nabla x.\nabla\varphi_1 - \lambda_1 x\varphi_1 - h(x)\varphi_1)\, dz \\ &= \int_\Omega (-h(x)\varphi_1)\, dz < 0 \quad \text{if} \quad x \in Z \cap \partial B, \end{aligned} \tag{3.4}$$

where we have used that φ_1 is positive in Ω. From (3.4) we have now, for each $s \in [0,1]$ and $x \in Z \cap \partial B$,

$$\langle sP\nabla\pi(x) + (1-s)(-\varphi_1), \varphi_1 \rangle < 0. \tag{3.5}$$

Hence by invariance under homotopy of the Leray–Schauder degree we have

$$d(P\nabla\pi, B \cap Z, 0) = d(K, B \cap Z, 0) = 0, \tag{3.6}$$

where $K(x) = -\varphi_1$ for all $x \in Z$. From (3.3) and (3.6) the lemma is proven. □

Let f^+ be the function defined by

$$f^+(\xi) = \begin{cases} f(\xi) & \text{if } \xi \geq 0, \\ f'(0)\xi & \text{if } \xi < 0. \end{cases}$$

Let $F^+(\xi) = \int_0^\xi f^+(s)\,ds$, and $J^+ : H \to \mathbb{R}$ be the functional defined by

$$J^+(u) = \int_\Omega \left(\frac{1}{2}\|\nabla u\|^2 - F^+(u) \right) dx. \tag{3.7}$$

Imitating the proof of Corollary 2.23 of [19] it readily follows that J^+ satisfies the hypotheses of the mountain pass theorem. Hence J^+ has a critical point u^+, which by the maximum principle is a positive solution to (1.2). Therefore, by Theorems 1 and 2 of [15], if the set of critical points of J^+ is discrete then at least one of them is of mountain pass type and has local degree -1. Similar arguments produce either infinitely many negative solutions to (1.2) or a negative solution u^- which is a critical point of mountain pass type and has local degree -1.

Let $\gamma = f'(\infty)$ and π as in Lemma 3.1. Since $f'(\infty)$ is not an eigenvalue of $-\Delta$ with zero Dirichlet boundary conditions, for $\rho > 0$ big enough and $s \in [0,1]$ the function $s\nabla J^+ + (1-s)\nabla\pi$ has no zero on the sphere centered at 0 with radius ρ. Hence by Lemma 3.1 we have

$$d(\nabla J^+,\ B_\rho,\ 0) \ = \ 0 \tag{3.8}$$

for ρ big enough. For future reference we summarize the above discussion into the following lemma.

LEMMA 3.2. *Under the hypotheses of Theorem* A, (1.2) *possesses a positive (respectively, a negative) solution. If the set of positive (respectively, negative) solutions is discrete then at least one of them is a critical point of mountain pass type and its local degree is* -1.

Since 0 is an isolated local minimum of J^+ and J we have

$$d(\nabla J^+,\ B,\ 0) \ = \ 1 \ = \ d(\nabla J,\ B,\ 0), \tag{3.9}$$

where B is a ball centered at zero containing no other critical point (see [2]). Hence if Σ is a bounded region containing the positive solutions and no other critical point of J we have

$$\begin{aligned} d(\nabla J,\ \Sigma,\ 0) \ &= \ d(\nabla J^+,\ \Sigma,\ 0) \\ &= \ d(\nabla J^+,\ B_\rho - \overline{B},\ 0) \\ &= \ d(\nabla J,\ B_\rho,\ 0) - d(\nabla J,\ B,\ 0) \\ &= \ -1. \end{aligned} \tag{3.10}$$

Similarly we see that if Σ_1 is a bounded region containing the negative solutions to (1.2) and no other critical point of J then

$$d(\nabla J,\ \Sigma_1,\ 0) \ = \ -1. \tag{3.11}$$

4. Proof of Theorem A. First, we show that there exists $u_0 \in H$ such that $\nabla J(u_0) = 0$ and, if isolated, then

$$d\,(\nabla J, V, 0) = (-1)^k \tag{4.1}$$

in any region V containing no other critical point of J. In fact, by Lemma 2.1 and (2.8) there exists $\psi\colon X \to Y$ such that

$$J(x+\psi(x)) = \min_{y\in Y} J(x+y).$$

Moreover, $\psi(x)$ is the unique member of Y such that

$$\langle \nabla J(x+\psi(x)), y\rangle = 0 \quad \text{for all} \quad y \in Y, \tag{4.2}$$

the function $\tilde{J}\colon X \to \mathbb{R}$ defined by $\tilde{J}(x) = J(x+\psi(x))$ is of class C^1, and

$$\langle \nabla \tilde{J}(x), x_1\rangle = \langle \nabla J(x+\psi(x)), x_1\rangle \quad \text{for all} \quad x, x_1 \in X. \tag{4.3}$$

We now claim that for $x \in X$

$$J(x) \to -\infty \quad \text{as} \quad \|x\| \to \infty. \tag{4.4}$$

Because $f'(\infty) \in (\lambda_k, \lambda_{k+1})$ there exists $b \in \mathbb{R}$ and $\overline{\gamma} > \lambda_k$ such that $F(\xi) \geq (\overline{\gamma}\xi^2/2) + b$. Hence

$$J(x) = \frac{1}{2}\|x\|^2 - \int_\Omega F(x) \leq \frac{1}{2}\|x\|^2 - \frac{\overline{\gamma}}{2}\int_\Omega x^2 - b|\Omega|.$$

Since $\langle x, x\rangle \leq \lambda_k \langle x, x\rangle_0$ for $x \in X$, we obtain

$$J(x) \leq \frac{1}{2}\|x\|^2\left(1 - \frac{\overline{\gamma}}{\lambda_k}\right) - b|\Omega| \longrightarrow -\infty \quad \text{as} \quad \|x\| \to \infty.$$

Because $\tilde{J}(x) \leq J(x)$, (4.4) implies that

$$\tilde{J}(x) \to -\infty \quad \text{as} \quad \|x\| \to \infty. \tag{4.5}$$

Since $\dim X < \infty$ there exists $x_0 \in X$ such that

$$\tilde{J}(x_0) = \max_{x\in X} J(x+\psi(x)).$$

Taking $u_0 = x_0 + \psi(x_0)$ we have (see Lemma 2.1) $\nabla J(u_0) = 0$. Suppose now that x_0 is an isolated critical point of $\tilde{J}$, hence u_0 is an isolated critical point of J. Since $-\tilde{J}$ has a local minimum at x_0, taking $W = \{x \in X; x+\psi(x) \in V\}$ then $d(\nabla \tilde{J}, W, 0) = (-1)^k$. Therefore by part (iv) of Lemma 2.1 we have (4.1).

Suppose k is even. Let R be large enough so that if $\nabla \tilde{J}(x) = 0$ then $\|x\| < R$. Because $f'(t) \leq \gamma < \lambda_{k+1}$, there exist positive numbers c_1 and c_2 such that for all $x \in X$ $\|\psi(x)\| \leq c_1 + c_2\|x\|$. Thus if $u = x + y$ is a critical point of J then $\|x\| \leq R$ and $\|y\| \leq c_1 + c_2\|x\|$. Because $-\tilde{J}$ is coercive, $d(\nabla \tilde{J}, B_R, 0) = (-1)^k = 1$. Thus by part (iv) of Lemma 2.1 $d(\nabla J, C, 0) = 1$ where $C = \{x+y; \|x\| < R, \|y\| < c_1 + c_2 R\}$. Suppose that K, the set of critical points of J, is finite. Let S_1, S_2 and S_3 be disjoint open bounded regions in H such that $\overline{S_1} \cap K = \{0\}$, $\overline{S_2} \cap K$ is the set of positive solutions to (1.2), and $\overline{S_3} \cap K$ is the set of negative solutions to (1.2). By (3.10) and (3.11) we have

$$d(\nabla J, S_2, 0) = d(\nabla J, S_3, 0) = -1. \tag{4.6}$$

If $u_0 = x_0 + \psi(x_0) \notin S_2 \cup S_3$ we let S_4 denote an open bounded region disjoint from $\overline{S_1 \cup S_2 \cup S_3}$ such that $\overline{S_4} \cap K = \{u_0\}$. By the excision property of the Leray–Schauder

degree we have

$$\begin{aligned}1 = d(\nabla J, C, 0) &= d(\nabla J, S_1, 0) + d(\nabla J, S_2, 0) + d(\nabla J, S_3, 0) + d(\nabla J, S_4, 0)\\ &\quad + d(\nabla J, C - \overline{(S_1 \cup S_2 \cup S_3 \cup S_4)}, 0)\\ &= 1 - 1 - 1 + 1 + d(\nabla J, C - \overline{(S_1 \cup S_2 \cup S_3 \cup S_4)}, 0).\end{aligned}$$

Thus, by the existence property of the Leray–Schauder degree we see that there exists $u_1 \in C - \overline{(S_1 \cup S_2 \cup S_3 \cup S_4)}$ such that $\nabla J(u_1) = 0$, which proves that (1.2) has at least five solutions. In this case both u_0 and u_1 change sign.

Suppose now that $u_0 \in S_2 \cup S_3$; without loss of generality we can assume that $u_0 \in S_2$. Let S_4 be a neighborhood of u_0 such that $\overline{S_4} \cap K = \{u_0\}$. By Lemma 3.2 there exists a critical point of mountain pass type $u_1 \in S_2$ such that if S_5 is a neighborhood of u_1 containing no other critical point of J^+ then $d(\nabla J, S_5, 0) = -1$. Thus

$$\begin{aligned}-1 = d(\nabla J, S_2, 0) &= d(\nabla J, S_4, 0) + d(\nabla J, S_5, 0) + d(\nabla J, S_2 - \overline{S_4 \cup S_5}, 0)\\ &= 1 - 1 + d(\nabla J, S_2 - \overline{S_4 \cup S_5}, 0).\end{aligned}$$

Thus, by the existence property of the Leray–Schauder degree there exists $u_2 \in S_2 - \overline{S_4 \cup S_5}$ with $\nabla J(u_2) = 0$. Finally,

$$\begin{aligned}1 = d(\nabla J, C, 0) &= d(\nabla J, S_1, 0) + d(\nabla J, S_2, 0) + d(\nabla J, S_3, 0)\\ &\quad + d(\nabla J, C - \overline{(S_1 \cup S_2 \cup S_3)}, 0)\\ &= 1 - 1 - 1 + d(\nabla J, C - \overline{(S_1 \cup S_2 \cup S_3)}, 0).\end{aligned}$$

Thus there exists $u_3 \in C - \overline{(S_1 \cup S_2 \cup S_3)}$ with $\nabla J(u_3) = 0$. Thus the set $\{0, u_0, u_1, u_2, u_3\}$ together with a critical point of J in S_3 shows that (1.2) has six solutions. Since $u_3 \notin S_2 \cup S_3$ and $u_0, u_1, u_2 \in S_2$, u_3 is a sign changing solution and u_0, u_1, u_2 have the same sign. This completes the proof of Theorem A when k is even.

Suppose k is odd. Let S_i, $i = 1, 2, 3$ be as above. If $u_0 \notin S_2 \cup S_3$ the proof follows very closely that of the case k even; the details are left to the reader. Suppose $u_0 \in S_2 \cup S_3$, say, $u_0 \in S_2$. Because $u_0 > 0$ in Ω and $\partial u_0/\partial \eta < 0$ in $\partial\Omega$ (here $\partial/\partial\eta$ denotes the outward unit normal derivative), using that X is finite-dimensional and standard regularity theory of elliptic operators it follows that for some $\epsilon > 0$ $x + \psi(x) > 0$ in Ω if $\|x - x_0\| < \epsilon$. Thus $\tilde{J}$ and $\tilde{J}^+$ coincide in $\{x; \|x - x_0\| < \epsilon\}$. Thus $\tilde{J}^+$ has a local maximum at x_0. Since we are assuming (1.2) to have only finitely many solutions, x_0 is a strict local maximum of $\tilde{J}^+$. Let $\delta > 0$ be such that $\tilde{J}^+(x) < \tilde{J}^+(x_0)$ if $\|x - x_0\| < \delta$. Since $k > 2$, $\{x; 0 < \|x - x_0\| < \delta\}$ is connected. Thus x_0 is not a critical point of mountain pass type. By Lemma 3.2 J^+ has a critical point of mountain pass type $u_1 = x_1 + \psi(x_1)$ such that if V is a neighborhood of u_1 containing no other critical point of J^+ in its closure then $d(\nabla J^+, V, 0) = -1$. In particular, by part (v) of Lemma 2.1 $x_0 \neq x_1$. Let V_0 (respectively, V_1) be a neighborhood of u_0 (respectively, $u_1 = x_1 + \psi^+(x_1)$) containing no other critical point in its closure. Thus

$$\begin{aligned}-1 = d(\nabla J^+, S_2, 0) &= d(\nabla J^+, V_0, 0) + d(\nabla J^+, V_1, 0) + d(\nabla J^+, S_2 - \overline{(V_0 \cup V_1)}, 0)\\ &= -2 + d(\nabla J^+, S_2 - \overline{(V_0 \cup V_1)}, 0).\end{aligned}$$

Thus by the existence property of the Leray–Schauder degree there exists a third positive solution $u_2 \in S_2 - \overline{(V_0 \cup V_1)}$. Since by the existence property of the Leray–Schauder degree (1.2) has a solution $u_3 \in S_3$, we see that (1.2) has five solutions,

namely $0, u_0, u_1, u_2, u_3$. Since $u_0, u_1, u_2 \in S_2$ they have the same sign. This proves Theorem A. □

REFERENCES

[1] S. AGMON, *The L^p approach to the Dirichlet problem*, Ann. Scuola Norm. Sup. Pisa, 13 (1959), pp. 405–408.

[2] H. AMMAN, *A note on degree theory for gradient mappings*, Proc. Amer. Math. Soc., 85 (1982), pp. 591–595.

[3] H. AMMAN AND E. ZEHNDER, *Nontrivial solutions for a class of nonresonance problems and applications to nonlinear differential equations*, Ann. Scuola Norm. Sup. Pisa, 7 (1980), pp. 539–603.

[4] A. AMBROSETTI AND G. MANCINI, *Sharp nonuniqueness results for some nonlinear problems*, Nonlinear Anal. Theory Methods, 3 (1979), pp. 635–645.

[5] A. AMBROSETTI AND P. RABINOWITZ, *Dual variational methods in critical point theory*, J. Funct. Anal., 14 (1973), pp. 343–381.

[6] L. BOCCARDO AND T. GALLOUËT, *Homogenization for "Castro-Lazer Equation,"* Nonlinear Anal., 14 (1990), pp. 81–91.

[7] H. BRÉZIS, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland, 1973.

[8] N. P. CÁC, *On nontrivial solutions of an asymptotically linear Dirichlet problem*, J. Differential Equations, 75 (1988), pp. 103–107.

[9] A. CASTRO, *Métodos de reducción via minimax*, Primer simposio Colombiano de Análisis Funcional, Medellín, Colombia, 1981.

[10] ———, *Métodos variacionales y análisis funcional no lineal*, X Coloquio Colombiano de Matemáticas, Paipa, Colombia, 1980.

[11] A. CASTRO AND A. C. LAZER, *Critical point theory and the number of solutions of a nonlinear Dirichlet problem*, Ann. Mat. Pura Appl., (4) 120 (1979), pp. 113–137.

[12] K. C. CHANG, *Solutions of asymptotically linear operator equations via Morse theory*, Comm. Pure Appl. Math., 34 (1981), pp. 693–712.

[13] H. HOFER, *A geometric description of the neighborhood of a critical point given by the mountain-pass theorem*, J. London Math. Soc., 31 (1985), pp. 566–570.

[14] ———, *Variational and topological methods in partially ordered Hilbert spaces*, Math. Ann., 261 (1982), pp. 493–514.

[15] ———, *The topological degree at a critical point of mountain-pass type*, Proc. Sympos. Pure Math., 45 (1986), pp. 501–509.

[16] A. C. LAZER AND J. P. MCKENNA, *Multiplicity results for a class of semilinear elliptic and parabolic boundary value problems*, J. Math. Anal. Appl., 107 (1985), pp. 371–395.

[17] A. C. LAZER AND S. SOLIMINI, *Nontrivial solutions of operator equations and Morse indices of critical points of min-max type*, Nonlinear Anal., 12 (1988), pp. 761–775.

[18] S. LI AND J. Q. LIU, *Nontrivial critical points for asymptotically quadratic function*, Report of International Center for Theoretical Physics IC/390, Trieste, Italy, 1986.

[19] P. RABINOWITZ, *Minimax methods in critical point theory with applications to differential equations*, Regional Conference Series in Mathematics, number 65, American Mathematical Society, Providence, RI, 1986.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1562–1576, November 1994

008

EXPLICIT HEAT KERNEL ON GENERALIZED CONES*

HENDRIK W. K. ANGAD-GAUR†, BERNARD GAVEAU‡, AND MASAMI OKADA§

Abstract. The authors compute explicitly the heat kernel on the surface of cones as well as on their generalizations. A procedure similar to the Fourier transform is employed in order to combine two Green's functions: one for the Bessel equation on the positive half-line and another for the Laplacian on graph networks. An analogue of the Poisson summation formula is derived from the residue theorem applied to the Green's function. Numerical computations are also implemented to determine some geometric quantity via the asymptotic expansion of the spectral function as t goes to zero.

Key words. explicit heat kernel, Bessel function, graph networks

AMS subject classifications. 35R, 58G

1. Introduction.

It is usually difficult to compute explicitly the heat kernel on two-dimensional models except the euclidean space or the disk with suitable metric. In this paper we would like to show that the explicit heat kernel can be computed on (the surface of) cones and on their generalizations endowed with usual euclidean metric.

It seems that few people had been interested in computing the explicit heat kernel on various two-dimensional spaces. In 1967, however, the heat kernel on the Riemann surface of $\log z$ was computed for the first time, as far as we know, by Edwards [1] in his investigation of the statistical mechanics of polymers. Of cource, since the heat kernel is a natural object in the classical probability theory, we can also find general useful information in the book of Itô and McKean [3]. Later in 1973, numerical computation among others was done by Saito and Chen [5]. Also see Nechaev [4] as a recent reference. Besides these, it seems difficult to find systematic treatment of heat kernels on the surface of cones in the literature.

In §2 of this paper, we will outline how the heat kernel was computed on the Riemann surface of $\log z$ via the Fourier transform for the sake of completeness. This is partly because we would like to elucidate by comparison our function theoretic method based on the Green's function which is described in §3.

In the next section, we start with the definition of generalized cones. These are real objects in our three-dimensional world and the explicit computation of the heat kernel on the surface of generalized cones is motivated by consideration of engineering models of two-dimensional singular polyhedrons with intersections. After generalizing the domain of definition of the Green's function, we shall state Theorem 2, one of our main results. As an application, we get an analogue of the Poisson summation formula for the heat kernel. Also some geometric quantity can be computed by means of the asymptotic expansion of the heat kernel.

In §4 we show an asymptotic expansion of the spectral function for some basic models in order to compare with that for classical smooth models.

In the last section, we shall present numerical computation of the coefficients in the asymptotic expansion.

* Received by the editors May 7, 1992; accepted for publication (in revised form) July 29, 1993.
† Tougaloo College, Tougaloo, Mississippi 39174.
‡ Université Pierre et Marie Curie, Paris, France.
§ Tôhoku University, Sendai, Japan.

2. Heat kernel on the Riemann surface of $\log z$. Let $\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}$ be the polar coordinate expression of the usual Laplacian on the plane, where $\theta = 0$ and $\theta = 2\pi$ are identified.

Then the heat kernel $P_t^{\mathcal{C}}$ is defined as follows.

DEFINITION 1. *$P_t^{\mathcal{C}}(r,\theta) = P_t^{\mathcal{C}}((r,\theta),(r_0,\theta_0))$ is the unique solution of the following heat equation:*

$$(1) \qquad \begin{cases} \left(\dfrac{\partial}{\partial t} - \Delta\right) P_t^{\mathcal{C}} = 0, & t > 0, \\ P_t^{\mathcal{C}} \to \delta_{r=r_0}\,\delta_{\theta=\theta_0}\cdot\dfrac{1}{r_0}, & \text{as } t \downarrow 0, \end{cases}$$

where δ is the Dirac delta function.

It represents the temperature at the point (r,θ) at the time t provided that an unit calorie is injected at the point (r_0,θ_0) at the time $t = 0$. Note that the multiplication by $1/r_0$ is required in the initial condition of (1) in order to have $\int P_t^{\mathcal{C}}(r,\theta)\, r\, dr\, d\theta = 1$. Then it is not difficult to compute $P_t^{\mathcal{C}}(r,\theta)$, since $\mathcal{C}$ is the direct product of two real lines. In fact, as is well known, it is equal to $(1/4\pi t)e^{-d/4t}$, where $d = r_0^2 + r^2 - 2r_0 r\cos(\theta - \theta_0)$.

Now let us proceed to the computation of the heat kernel on $\mathcal{R}$ the Riemann surface of $\log z$.

DEFINITION 2. *The heat kernel on $\mathcal{R}$ denoted by $P_t(r,\theta)$ is defined in the same way as the solution of the heat equation* (1).

However note that in this case θ varies on the whole real line $(-\infty,\infty)$, i.e., $\theta = 0$ and $\theta = 2\pi$ are no longer identified.

2.1. Computation of the heat kernel on the Riemann surface. To compute $P_t(r,\theta) = P_t((r,\theta),(r_0,0))$, Edwards [1] used the Fourier inversion theorem with respect to θ. In the sequel we shall essentially follow his argument.

Step 1. By definition P_t satisfies (1) with $\theta_0 = 0$. First, we denote the Fourier transform of P_t by $\widehat{P}_t$, which is defined by

$$\widehat{P}_t(r,\xi) = \int_{-\infty}^{\infty} P_t(r,\theta)\, e^{-i\theta\xi}\, d\theta = 2\int_0^{\infty} P_t(r,\theta)\cos\theta\xi\, d\theta\ .$$

Here we used the fact that P_t is an even function of θ by symmetry. Then by applying the Fourier transform to the heat equation (1), we obtain

$$(2) \qquad \begin{cases} \left[\dfrac{\partial}{\partial t} - \left(\dfrac{\partial^2}{\partial r^2} + \dfrac{1}{r}\dfrac{\partial}{\partial r} - \dfrac{\xi^2}{r^2}\right)\right]\widehat{P}_t(r,\xi) = 0\ , & t > 0, \\ \widehat{P}_t(r,\xi) \to \dfrac{1}{r_0}\,\delta_{r=r_0}, \end{cases}$$

because of the property of the Fourier transform of derivatives.

Therefore it is reduced to solve the heat equation for one space variable r. For convenience of computation, we shall assume $\xi \geq 0$ from now on since $\widehat{P}_t(r,\xi)$ is an even function of ξ.

Step 2. Let us consider the Laplace transform of the first equation of (2) with respect to t.

DEFINITION 3. $G_0 = G_0(r, r_0, \mu, -\xi^2)$ *is the Laplace transform with respect to* t *of* $r_0 \widehat{P}_t(r,\xi)$ *with* $Re\ \mu > 0$, *i.e.,*

$$G_0 = \int_{+0}^{\infty} e^{-\mu t}\, r_0\, \widehat{P}_t(r,\xi)\, dt. \tag{3}$$

Then we get

$$\left[\mu - \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} - \frac{\xi^2}{r^2}\right)\right] G_0 = \delta_{r=r_0}, \tag{4}$$

since by integration of parts,

$$\int_{+0}^{\infty} e^{-\mu t} \frac{\partial}{\partial t}\widehat{P}_t\, dt = -\widehat{P}_{+0} + \frac{\mu}{r_0} G_0\,,$$

and $\widehat{P}_t(r,\xi) \to (1/r_0)\delta_{r=r_0}$ *as* $t \downarrow 0$. *Therefore it turns out that* G_0 *is the Green's function for the modified Bessel equation. Now let us see how* G_0 *is computed.*

LEMMA 1.

$$G_0(r, r_0, \mu, -\xi^2) = \frac{\pi r_0}{2\sin\xi\pi}\left[I_\xi(\sqrt{\mu}\, r_0) I_{-\xi}(\sqrt{\mu}\, r) - I_\xi(\sqrt{\mu}\, r_0) I_\xi(\sqrt{\mu} r)\right], \quad r_0 \le r\,, \tag{5}$$

where $I_{\pm\xi}(z) = e^{\mp i\xi\pi/2} J_{\pm\xi}(iz)$ *is the modified Bessel function.*

Proof. First, note that $I_{\pm\xi}(\sqrt{\mu}\, r)$ are solutions of (4) with δ replaced by 0 and that $I_\xi \in L^2_{loc}$ near $r = 0$ when $\mathrm{Re}\ \xi > -\frac{1}{2}$ in view of the asymptotic expansion of I_ξ. Next, we introduce another modified Bessel function $K_\xi(z)$ defined by $K_\xi(z) = \pi(I_{-\xi}(z) - I_\xi(z))/[2\sin\xi\pi]$. It is known that $K_\xi(\sqrt{\mu}\, r) \to 0$ as $r \to \infty$ since $\mathrm{Re}\ \sqrt{\mu} > 0$ and the Wronskian $W(I_\xi(z), K_\xi(z)) = -1/z$. See Watson [6] for these facts on Bessel functions.

Now we put $u(r) = I_\xi(\sqrt{\mu}\, r)$ and $v(r) = K_\xi(\sqrt{\mu}\, r)$. Then the Green's function $G_0 = G_0(r, r_0, \mu, -\xi^2)$ is expressed as $G_0 = -u(r_0)\, v(r)/W(u,v)$ for $r_0 \le r$, according to the standard fact of Green's function for Sturm–Liouville equations. The rest of the proof is immediate. □

Step 3. Therefore by taking the inverse Laplace transform of G_0,

$$\begin{aligned}\widehat{P}_t(r,\xi) &= \frac{1}{2i\pi r_0}\int_{c-i\infty}^{c+i\infty} e^{t\mu} G_0\, d\mu \\ &= \frac{1}{4i\sin\xi\pi}\int_{c-i\infty}^{c+i\infty} e^{t\mu}\left[I_\xi(\sqrt{\mu}\, r_0)\, I_{-\xi}(\sqrt{\mu}\, r) - I_\xi(\sqrt{\mu}\, r_0)\, I_\xi(\sqrt{\mu}\, r)\right] d\mu\,.\end{aligned}$$

Now for convenience of computation we modify the path of integration, i.e., we choose the path Γ around the negative real axis.

Then, since $I_\xi(\sqrt{\mu}\, r_0) I_{-\xi}(\sqrt{\mu}\, r)$ is holomorphic and bounded in $\{\mu \in \mathbb{C}\,;\, Re\mu \le c\}$, with $c > 0$, its integral turns out to be zero. Therefore

$$\widehat{P}_t(r,\xi) = \frac{i}{4\sin\xi\pi}\int_\Gamma e^{t\mu}\, I_\xi(\sqrt{\mu}\, r_0)\, I_\xi(\sqrt{\mu}\, r)\, d\mu\,.$$

See Fig. 1.

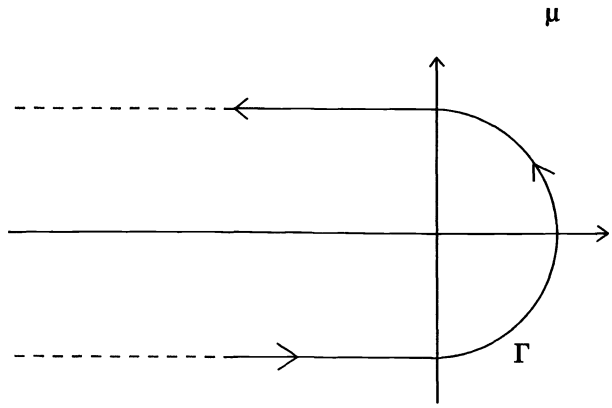


FIG. 1.

Since $I_\xi(z) = e^{-i\xi\pi/2} J_\xi(iz)$, we obtain

$$\widehat{P}_t(r,\xi) = \frac{e^{-i\xi\pi}}{2i\sin\xi\pi} \int_{\infty+i0}^{-\infty+i0} e^{-t\eta^2} J_\xi(r_0\eta)\, J_\xi(r\eta)\, \eta\, d\eta$$

by the change of variables $\eta = i\sqrt{\mu}$. Let us divide the integral into two parts. Then the above integral is reduced to

$$\begin{aligned} &-\int_0^\infty e^{-t\eta^2} J_\xi(r_0\eta)\, J_\xi(r\eta)\eta\, d\eta + \int_0^\infty e^{-t\eta^2} J_\xi(-r_0\eta)\, J_\xi(-r\eta)\eta\, d\eta \\ &\quad = (e^{2\pi\xi i} - 1) \int_0^\infty e^{-t\eta^2} J_\xi(r_0\eta)\, J_\xi(r\eta)\eta\, d\eta, \end{aligned}$$

since $J_\xi(-z) = e^{\xi\pi i} J_\xi(z)$. Then it follows that

$$\begin{aligned} \widehat{P}_t(r,\xi) &= \int_0^\infty e^{-t\eta^2} J_\xi(r_0\eta)\, J_\xi(r\eta)\eta\, d\eta \\ &= \frac{1}{2t} e^{-(r_0^2+r^2)/4t} I_\xi\left(\frac{r_0 r}{2t}\right), \end{aligned} \tag{6}$$

from Weber's second integral formula.

Step 4. Therefore by taking the inverse Fourier transform of $\widehat{P}_t$,

$$\begin{aligned} P_t(r,\theta) &= \frac{2}{\pi} \int_0^\infty \left(\int_0^\infty P_t(r,\phi)\cos(\phi\xi)\, d\phi \right) \cos(\theta\xi)\, d\xi \\ &= \frac{1}{\pi} \int_0^\infty \widehat{P}_t(r,\xi)\cos(\theta\xi)\, d\xi \\ &= \frac{1}{4\pi t} e^{-(r_0^2+r^2)/4t} \int_{-\infty}^\infty I_{|\xi|}\left(\frac{r_0 r}{2t}\right) e^{i\theta\xi}\, d\xi. \end{aligned}$$

Now we recall that

$$I_\xi(q) = \frac{1}{\pi}\int_0^\pi \cos(\xi u)\, e^{q\cos u}\, du - \frac{\sin\xi\pi}{\pi}\int_0^\infty e^{-\xi v - q\cosh v}\, dv \tag{7}$$

(Watson, [6, p. 181]), from which Saito and Chen [5] obtained the next expression.

PROPOSITION 1.

(8)

$$\int_{-\infty}^{\infty} I_{|\xi|}(q)\, e^{i\xi\theta}\, d\xi = \begin{cases} \dfrac{1}{\pi}\displaystyle\int_0^{\infty} \frac{dy}{1+y^2}\left(e^{-q\cosh((\pi+\theta)y)} - e^{-q\cosh((\pi-\theta)y)}\right) & \theta \le -\pi \\ e^{q\cos\theta} - \dfrac{1}{\pi}\displaystyle\int_0^{\infty} \frac{dy}{1+y^2}\left(e^{-q\cosh((\pi+\theta)y)} + e^{-q\cosh((\pi-\theta)y)}\right) & -\pi < \theta < \pi \\ \dfrac{1}{\pi}\displaystyle\int_0^{\infty} \frac{dy}{1+y^2}\left(e^{-q\cosh((\pi-\theta)y)} - e^{-q\cosh((\pi+\theta)y)}\right) & \pi \le \theta \end{cases}$$

$$\equiv J(q,\theta).$$

Proof. This can be verified by direct computation and we refer to [5] for the details. □

And thus we have finally the following.

THEOREM 1 (Edwards). *Let* $P_t(\,(r,\theta),(r_0,0)\,)$ *be the heat kernel on the Riemann surface of* $\log z$. *Then*

$$P_t(\,(r,\theta),(r_0,0)\,) = \frac{1}{4\pi t} e^{-(r_0^2+r^2)/4t}\, J\left(\frac{r_0 r}{2t},\, \theta\right) \tag{9}$$

where $J(\frac{r_0 r}{2t},\, \theta)$ *is defined in* Proposition 1.

COROLLARY 1.

$$P_t(\,(r,0),(r,0)\,) = \frac{1}{4\pi t} - \frac{1}{2\pi^2 t}\, e^{-r^2/2t} \int_0^{\infty} \frac{dy}{1+y^2}\left(e^{-(r^2/2t)\cosh(\pi y)}\right). \tag{10}$$

Proof. The proof is immediate from Proposition 1 and Theorem 1 with θ replaced by 0. □

3. On the surface of generalized cones. First we define the graph network.

DEFINITION 4. *A graph network is a branched one-dimensional manifold with finite vertices, namely, an ordinary circuit formed by joining a finite number of points with segments called edges.*

For a given graph network X let us define CX, the cone over X as follows.

DEFINITION 5. CX *is the quotient space* $X\times[0,\infty)/X\times 0$ *endowed with usual two-dimensional euclidean metric. Without the metric,* CX *is an object of the homotopy theory. The metric is important in our study as the following example shows.*

Example 1. C_α. This is is the cone over $\boldsymbol{R}/\alpha\boldsymbol{Z}$ according to our definition and is actually nothing but the (surface of) ordinary cone of apparatus α.

Note that the euclidean plane corresponds to $\alpha = 2\pi$. Moreover the Riemann surface of $\log z$, $\mathcal{R}$ is expressed as $C\boldsymbol{R}$. With these examples in mind, it is natural to call CX a generalized cone. See Fig. 2 for some more general examples.

In this section we shall show how to construct the heat kernel P_t on CX using the Green's function on X and the Green's function for the Bessel equation.

First we recall the Green's function on X. Let Δ_X be the Laplacian on X which is simply the second derivative on each edge with the domain of definition reflecting the conservation law of heat flux at each vertex. See [2] for details.

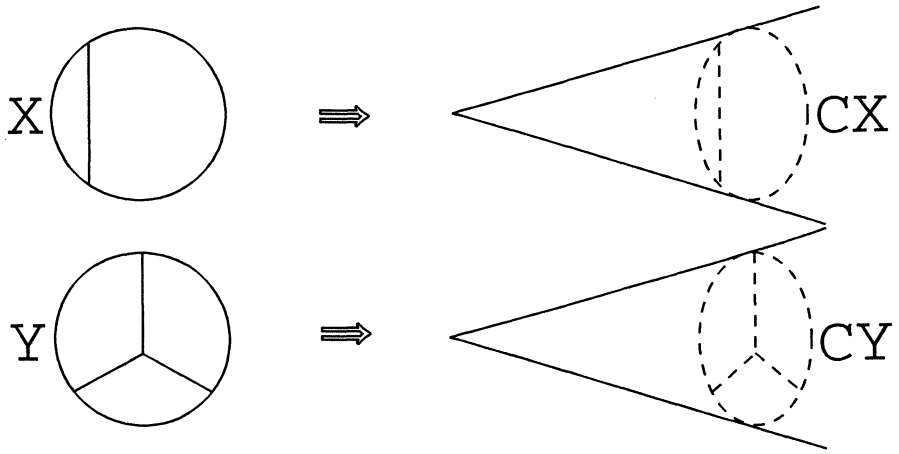


FIG. 2.

Now, here goes the definition of the Green's function on CX.

DEFINITION 6. *For a fixed point y of X, the Green's function $g = g(x, y, \lambda)$ on X is defined symbolically by $g = (\lambda I - \Delta_X)^{-1} \delta_y$, where $\lambda \notin \boldsymbol{R}_-$,the negative real axis.*

Next we would like to define the Green's function on $[0, \infty)$ for the Bessel equation by $G = \frac{1}{r_0} G_0$, where G_0 has been already defined in (4) for $\lambda = -\xi^2 \in \boldsymbol{R}_-$. Now, for a general complex value λ, $G(r, r_0, \mu, \lambda)$ is defined by the following.

DEFINITION 7. *G is defined for $\lambda \notin \boldsymbol{R}_+$ by*

$$G(r, r_0, \mu, \lambda) = \frac{1}{r_0} \left[\mu - \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{\lambda}{r^2} \right) \right]^{-1} \delta_{r=r_0}.$$

Remark. We have to be careful about the domain of definition with respect to the variable λ. We have chosen the analytic branch of g which is continuous across the positive real axis. However, in contrast to g, G should have continuous analytic extension across the negative real axis. The consequence is that $G(r, r_0, \mu, \lambda)$ is uniquely determined for $\lambda = -\xi^2 \in \boldsymbol{R}_-$ and that we have as in (6)

$$\frac{1}{2\pi i} \int_\Gamma e^{\mu t} G \, d\mu = \frac{1}{2t} e^{-(r_0^2 + r^2)/4t} I_{\sqrt{-\lambda}} \left(\frac{r_0 r}{2t} \right),$$

where for any λ, $\operatorname{Re} \sqrt{-\lambda} \geq 0$.

Now let us define P_t^X the heat kernel on CX.

DEFINITION 8. *$P_t^X = P_t^X((r, x), (r_0, x_0))$ is the solution of the following equation.*

$$\text{(11)} \qquad \begin{cases} \left[\dfrac{\partial}{\partial t} - \left(\dfrac{\partial^2}{\partial r^2} + \dfrac{1}{r} \dfrac{\partial}{\partial r} + \dfrac{1}{r^2} \Delta_X \right) \right] P_t^X = 0 , & t > 0, \\ P_t^X \to \dfrac{1}{r_0} \delta_{r=r_0} \, \delta_{x=x_0}, & \text{as } t \downarrow 0. \end{cases}$$

Then the following integral representation is a consequence of a natural generalization of the separation of variable technique and is an extension of Laplace–Fourier transform method of §2.

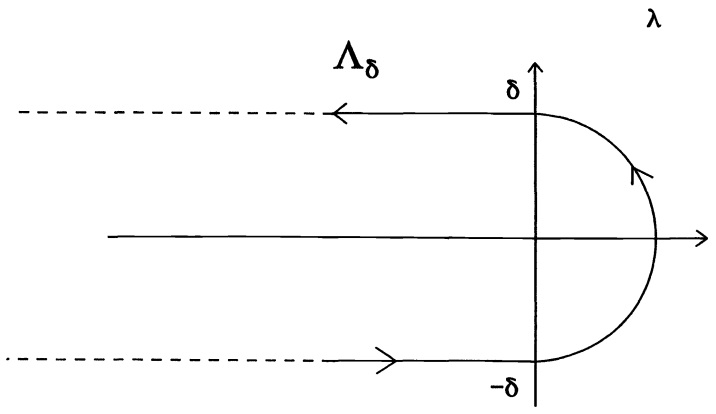


FIG. 3.

Let Λ_δ be the contour along the negative real axis which approaches it as $\delta \downarrow 0$. See Fig. 3.

THEOREM 2. *Let g and P_t^X be the Green's function on X and the heat kernel on the cone over X endowed with the euclidean metric, respectively. Then*

$$P_t^X = \frac{1}{2\pi i}\int_\Lambda g(x,x_0,\lambda)\,d\lambda\left(\frac{1}{2\pi i}\int_\Gamma e^{\mu t}\,G(r,r_0,\mu,\lambda)\,d\mu\right) \tag{12}$$

where $\Lambda = \lim_{\delta\downarrow 0}\Lambda_\delta$. Here the improper integral with respect to λ is interpreted as

$$\lim_{\epsilon\downarrow 0}\frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda}\,g\,d\lambda(\cdots)\ .$$

Proof.

Step 1. Let $u_\epsilon(t,r,x) = \frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda}\,g\,d\lambda(\cdots)$. Then on the one hand,

$$\frac{\partial u_\epsilon}{\partial t} = \frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda} g\,d\lambda\left(\frac{1}{2\pi i}\int_\Gamma \mu\,e^{\mu t}\,G\,d\mu\right).$$

On the other hand,

$$\begin{aligned}\left(\frac{\partial^2}{\partial r^2}+\frac{1}{r}\frac{\partial}{\partial r}+\frac{1}{r^2}\Delta_X\right)u_\epsilon &= \frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda}g\,d\lambda\left(\frac{1}{2\pi i}\int_\Gamma e^{\mu t}\left\{\left(\mu-\frac{\lambda}{r^2}\right)G-\frac{\delta_{r=r_0}}{r_0}\right\}d\mu\right)\\ &\quad+\frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda}(\lambda g-\delta_{x=x_0})\frac{d\lambda}{r^2}\left(\frac{1}{2\pi i}\int_\Gamma e^{\mu t}G\,d\mu\right)\\ &= \frac{\partial u_\epsilon}{\partial t}-\frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda}g\,d\lambda\left(\frac{1}{2\pi i}\int_\Gamma e^{\mu t}\,d\mu\frac{\delta_{r=r_0}}{r_0}\right)\\ &\quad-\frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda}\delta_{x=x_0}\frac{d\lambda}{r^2}\left(\frac{1}{2\pi i}\int_\Gamma e^{\mu t}G\,d\mu\right).\end{aligned}$$

Here note that the second term of the right-hand side $= 0$ since $\int_\Gamma e^{\mu t}\,d\mu = 0$ and the third term is also zero because we have shown that $\frac{1}{2\pi i}\int_\Gamma e^{\mu t}G\,d\mu = \frac{1}{2t}\,e^{-(r_0^2+r^2)/4t}$

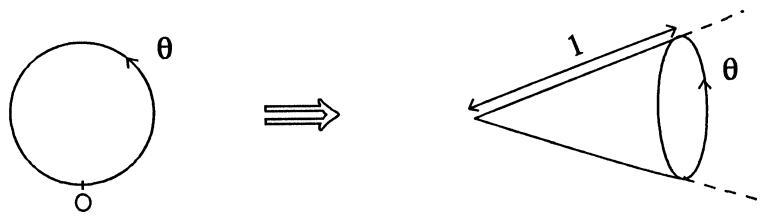


FIG. 4.

$I_{\sqrt{-\lambda}}\left(\frac{r_0 r}{2t}\right)$ from which follows $\int_\Lambda e^{\epsilon\lambda} I_{\sqrt{-\lambda}}\, d\lambda = 0$. See the preceeding remark. Consequently, u_ϵ satisfies the heat equation.

Step 2. Let us show the second statement of (11). First we see that $\frac{1}{2\pi i}\int_\Gamma e^{\mu t} G\, d\mu \to \frac{\delta_{r=r_0}}{r_0}$ as $t \downarrow 0$ by definition of the resolvent owing to the operational calculus : $\frac{1}{2\pi i}\int_\Gamma (\mu I - L)^{-1}\, d\mu = I$ (identity), where $L = \frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} + \frac{\lambda}{r^2}$. Therefore,

$$u_\epsilon \to \frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda} g\, d\lambda \frac{\delta_{r=r_0}}{r_0} \quad \text{as} \quad t \downarrow 0\,.$$

Next, note that $h_\epsilon^X(x, x_0) \equiv \frac{1}{2\pi i}\int_\Lambda e^{\epsilon\lambda} g\, d\lambda$ was shown to be the heat kernel on X at $t = \epsilon$ and $\lim_{\epsilon\downarrow 0} h_\epsilon^X = \delta_{x=x_0}$ [2]. This implies that $P_t^X = \lim_{\epsilon\downarrow 0} u_\epsilon$ is actually the desired heat kernel by the unicity of heat kernel. □

Remarks. (i) In the case where X is the real line, then $\frac{1}{2\pi i}\int_\Gamma e^{\mu t} G(r, r_0, \mu, -\xi^2)\, d\mu$ corresponds to $\widehat{P}_t$ of the preceding section and $g(x, x_0, -\xi^2)$ to $(e^{i\xi(x-x_0)})/2i\xi$.

(ii) We have employed the Green's function g which plays the role of a resolution of identity. Therefore it would be interesting to know if we could give other integral representations of the heat kernel based on various integral formulas such as Hankel, Kantrovich-Lebedev, Meijer and so on.

(iii) This type of decoupling procedure would hold in a more general case, for example, in the case of several space variables.

Example 2. q_t^α.

DEFINITION 9. *Let* $q_t^\alpha((r,\theta),(r_0,0))$ *be the heat kernel on* C_α *the surface of cone of apparatus* α. *See Fig.* 4.

We have then $g(\theta, 0, \lambda) = (e^{\sqrt{\lambda}\theta} + e^{\sqrt{\lambda}(\alpha-\theta)})/(2\sqrt{\lambda}(e^{\sqrt{\lambda}\alpha} - 1))$ by a simple computation.

Therefore, after the Taylor expansion of g in terms of $e^{\sqrt{\lambda}\alpha}$, we can use Proposition 1 to apply Theorem 2. The result is

$$q_t^\alpha = \sum_{m=-\infty}^{\infty} P_t\big((r, \theta + m\alpha), (r_0, 0)\big), \tag{13}$$

where P_t is the heat kernel in (9), which is also easy to check from the geometry.

As a consequence of Theorem 2 we have an equivalent formula (Poisson summation formula) as follows.

COROLLARY 2.

$$q_t^\alpha\big((r,\theta),(r_0,0)\big) = \frac{1}{\alpha t}\, e^{-(r_0^2+r^2)/4t}\left(\sum_{m=0}^{\infty} \cos\frac{2m\pi\theta}{\alpha}\, I_{2m\pi/\alpha}\left(\frac{r_0 r}{2t}\right) - \frac{1}{2} I_0\left(\frac{r_0 r}{2t}\right)\right).$$

Proof. It suffices to apply Theorem 2 using the residue formula to the integral derived from the right-hand side of (12)

$$\frac{1}{2\pi}\int_{-\infty-i0}^{\infty-i0}\frac{e^{i\xi\theta}+e^{i\xi(\alpha-\theta)}}{e^{i\xi\alpha}-1}\,I_{|\xi|}\left(\frac{r_0 r}{2t}\right)d\xi. \qquad \square$$

Moreover, the classical reflection principle gives another equivalent expression for special values of α. Let n be a positive integer. We define Q_t^α by

$$Q_t^\alpha(\,(r,\theta),\,(r_0,0)\,)=\begin{cases}\sum_{m=-n+1}^{n} P_t^{\mathcal{C}}(\,(r,\theta),\,(r_0,m\alpha)\,) & \text{if} \quad \alpha=\frac{2\pi}{2n},\\ \sum_{m=-n}^{n} P_t^{\mathcal{C}}(\,(r,\theta),\,(r_0,m\alpha)\,) & \text{if} \quad \alpha=\frac{2\pi}{2n+1}.\end{cases}$$

Then we have the following.

PROPOSITION 2. *Let q_t^α be defined in Definition* 9. *Then $Q_t^\alpha = q_t^\alpha$.*

Proof. First Q_t^α satisfies the heat equation. Next as $r \to r_0$ and $\theta \to 0$, $Q_t^\alpha \sim \frac{1}{4\pi t}$. Furthermore we easily see that Q_t^α is smoothly connected at $\theta = \pm\alpha/2$. Therefore we are done by the unicity of heat kernels. $\square$

Notation. Let us denote by $C(\alpha)$ the following quantity.

(14)
$$\int_0^\alpha d\theta\left(\int_0^\infty\left\{q_t^\alpha\big((r,\theta),(r,\theta)\big)-\frac{1}{4\pi t}\right\}r\,dr\right)=\alpha\int_0^\infty\left\{q_t^\alpha(\,(r,0),(r,0)\,)-\frac{1}{4\pi t}\right\}r\,dr.$$

Then we can compute $C(\alpha)$ as follows.

PROPOSITION 3. *Let $K(s)=\frac{1}{2\pi^2}\int_0^\infty \frac{dy}{1+y^2}\left(\frac{1}{1+\cosh sy}\right)$. Then*

$$\begin{aligned}C(\alpha) &= -\alpha\,K(\pi)+\alpha\textstyle\sum_{m=1,2,\cdots,[\pi/\alpha]}\left\{\frac{1}{2\pi(1-\cos m\alpha)}-K(\pi-m\alpha)-K(\pi+m\alpha)\right\}\\ &\quad+\alpha\textstyle\sum_{m=[\pi/\alpha]+1}^{\infty}\left\{K(\pi-m\alpha)-K(\pi+m\alpha)\right\},\end{aligned}$$

where the second sum is supposed to be zero if $\alpha > \pi$.

Proof. We compute each term of the series in (13) by means of Theorem 1 and Proposition 1. $\square$

Remarks.

(i) From the definition of $C(\alpha)$, an asymptotic expansion follows:

$$\int_0^\alpha\int_0^l q_t^\alpha(\,(r,\theta),\,(r,\theta)\,)\,r\,dr\,d\theta\sim\frac{\alpha l^2}{8\pi t}+C(\alpha)\ \text{modulo}\ O(e^{-cl^2/t})\ \text{as}\ t\downarrow 0\ ,$$

where l is an arbitrary positive constant and c is a positive constant depending on α.

(ii) A direct consequence of Proposition 2 is the following.

$$C(\alpha)=\begin{cases}\dfrac{\alpha}{8\pi}+\displaystyle\sum_{k=1}^{n-1}\frac{\alpha}{2\pi(1-\cos k\alpha)} & \text{if } \alpha=\frac{2\pi}{2n},\\ \displaystyle\sum_{k=1}^{n}\frac{\alpha}{2\pi(1-\cos k\alpha)} & \text{if } \alpha=\frac{2\pi}{2n+1}.\end{cases}$$

(iii) Another interesting quantity χ is defined by

$$\chi(\alpha)=2\alpha\int_0^\infty\sum_{m=-\infty}^{\infty}P_t\big(\,(r,m\alpha),\,(r,0)\,\big)\,(1-\cos m\alpha)\,r\,dr\ .$$

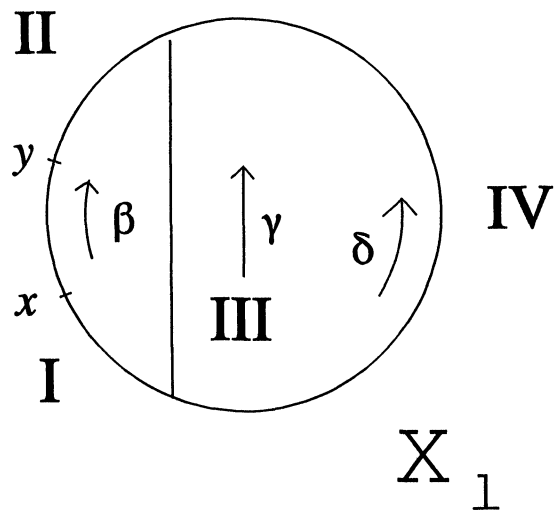


FIG. 5.

Then we can compute χ as in Proposition 3 and the result is:

$$\begin{aligned}\chi(\alpha) = 2\alpha \sum_{m=1}^{[\pi/\alpha]} (1-\cos m\alpha)\Big\{ &\frac{1}{2\pi(1-\cos m\alpha)} - K(\pi + m\alpha) \\ &- K(\pi - m\alpha)\Big\} \\ +2\alpha \sum_{m=[\pi/\alpha]+1}^{\infty} &(1-\cos m\alpha)\left[K(\pi - m\alpha) - K(\pi + m\alpha)\right] \\ = 1 - \frac{\alpha}{2\pi}&.\end{aligned}$$

This last equality can be verified easily when $\alpha = \frac{2\pi}{2n}$ or $\frac{2\pi}{2n+1}$, $n = 1, 2, 3, \ldots$. For other values of α, numerical methods show that equality holds within the margin of error. This $\chi(\alpha)$ actually is the index of the cone C_α and the above numerical value suggests that the Gauss–Bonnet formula is also valid for compact surfaces with conic singularities. The full account of this will be published elsewhere.

Example 3. Let X_1 be a graph network with two vertices connected by three edges of length β, γ and δ as is depicted in Fig. 5.

Then the Green's function is given by

$$G(x,y,\lambda) = \begin{cases} x_1 C^x + x_2 C^{-x} & \text{if } x \in I, \\ x_3 C^x + x_4 C^{-x} & \text{if } x \in II, \\ x_5 C^x + x_6 C^{-x} & \text{if } x \in III, \\ x_7 C^x + x_8 C^{-x} & \text{if } x \in IV, \end{cases} \tag{15}$$

where $C = e^{\sqrt{\lambda}}$, $\lambda = \mu^2$ and x_1, x_2, x_3, x_4, x_5, x_6, x_7, and x_8 are defined by the

following:

$$D = (C^{-2\gamma} - 1)\Big[-\frac{3}{2}C^{-\beta-\delta} + 2 + C^{\gamma-\delta} - \frac{1}{2}C^{\beta+\delta} - 2C^{\beta+\gamma} - \frac{1}{2}C^{-\beta+\delta} + C^{\gamma+\delta} + \frac{1}{2}C^{\beta-\delta}\Big] + (C^{\delta-\gamma} - 1)\big[-C^{\gamma-\delta} + C^{-\beta-\delta} + 3C^{\beta+\gamma} - 4 - C^{\beta-\delta} + C^{-\beta-\gamma} + C^{-\gamma-\delta}\big] + (1 - C^{-\delta-\gamma})\big[C^{\gamma+\delta} - C^{-\beta+\delta} + C^{\beta+\gamma} - 4 + C^{\beta+\delta} + 3C^{-\beta-\gamma} - C^{\delta-\gamma}\big],$$

$$x_1 = \frac{1}{\mu}\Big[-C^{-y} - \frac{1}{4}C^{-y+\beta-\delta} - \frac{1}{4}C^{y-\beta-\delta} - C^{-y-\gamma+\delta} - 2C^{-y+\beta-\gamma} + \frac{3}{2}C^{-y+\beta+\delta} + 2C^{y-\beta-\gamma} - \frac{3}{4}C^{y-\beta+\delta} + C^{-y-\gamma-\delta} + C^{-y-2\gamma} + \frac{1}{4}C^{-y+\beta-2\gamma-\delta} - \frac{3}{4}C^{y-\beta-2\gamma-\delta} - \frac{1}{4}C^{-y+\beta-2\gamma+\delta} - \frac{1}{4}C^{y-\beta-2\gamma+\delta} + \frac{3}{4}C^{-y+\beta+\delta}\Big]/D,$$

$$x_2 = \frac{1}{\mu}\Big[-\frac{1}{4}C^{-y+\beta-\delta} - \frac{1}{4}C^{y-\beta-\delta} + C^{y} + C^{y-\gamma+\delta} - C^{y-\gamma-\delta} - \frac{3}{4}C^{-y+\beta-2\gamma-\delta} - \frac{1}{4}C^{-y+\beta-2\gamma+\delta} + 2C^{-y+\beta-\gamma} - \frac{1}{4}C^{y-\beta-2\gamma+\delta} - 2C^{y-\beta-\gamma} - \frac{3}{4}C^{-y+\beta+\delta} + \frac{1}{4}C^{y-\beta+\delta} + \frac{9}{4}C^{y-\beta-2\gamma-\delta} - C^{y-2\gamma}\Big]/D,$$

$$x_3 = \frac{1}{\mu}\Big[-2C^{-y-\beta-\gamma} + \frac{1}{2}C^{y-2\beta-\gamma+\delta} + 2C^{y-\beta-\gamma} + \frac{1}{4}C^{-y-\beta+\delta} - \frac{3}{4}C^{y-\beta+\delta} - \frac{1}{4}C^{y-\beta-\delta} - \frac{1}{4}C^{-y-\beta-2\gamma+\delta} - \frac{1}{4}C^{y-\beta-2\gamma+\delta} + \frac{9}{4}e^{-y-\beta-2\gamma-\delta} - \frac{3}{4}C^{y-\beta-2\gamma-\delta} - \frac{1}{4}C^{-y-\beta-\delta} + C^{-y-\gamma+\delta} + C^{-y} - \frac{1}{2}C^{y-2\beta-\gamma+\delta} - C^{-y-\gamma-\delta} - C^{-y-2\gamma}\Big]/D,$$

$$x_4 = \frac{1}{\mu}\Big[2C^{-y+\beta-\gamma} - C^{y-\gamma+\delta} - 2C^{y+\beta-\gamma} - C^{y} - \frac{3}{4}C^{-y+\beta+\delta} + \frac{9}{4}C^{y+\beta+\delta} + C^{y-\gamma-\delta} - \frac{1}{4}C^{-y+\beta-\delta} - \frac{1}{4}C^{y+\beta-\delta} - \frac{1}{4}C^{-y+\beta-2\gamma+\delta} + C^{y-2\gamma} - \frac{1}{4}C^{y+\beta-2\gamma+\delta} - \frac{3}{4}C^{-y+\beta-2\gamma-\delta}\Big]/D,$$

$$x_5 = \frac{1}{\mu}\Big[-\frac{1}{2}C^{-y-\gamma+\delta} + \frac{3}{2}C^{y-\gamma+\delta} - \frac{1}{2}C^{-y-\gamma-\delta} - \frac{1}{2}C^{y-\gamma-\delta} + C^{-y-2\gamma}$$
$$-\frac{1}{2}C^{-y+\beta-2\gamma-\delta} + \frac{3}{2}C^{y-\beta-2\gamma-\delta} - \frac{1}{2}C^{-y+\beta-2\gamma+\delta} + C^{-y+\beta-\gamma}$$
$$-\frac{1}{2}C^{y-\beta-2\gamma+\delta} - C^{y-\beta-\gamma} - C^{y-2\gamma}\Big]/D,$$

$$x_6 = \frac{1}{\mu}\Big[-C^{-y} - \frac{1}{2}C^{-y+\beta-\delta} - \frac{1}{2}C^{y-\beta-\delta} + C^{y} - \frac{1}{2}C^{-y-\gamma+\delta} - C^{-y+\beta-\gamma}$$
$$+\frac{3}{2}C^{-y+\beta+\delta} + C^{y-\beta-\gamma} - \frac{1}{2}C^{y-\gamma+\delta} - \frac{1}{2}C^{y-\beta+\delta} + \frac{3}{2}C^{-y-\gamma-\delta}$$
$$-\frac{1}{2}C^{y-\gamma-\delta}\Big]/D,$$

x_7 and x_8 are obtained if we interchange γ and δ in the expression of x_5 and x_6 respectively. The above values of $x_1, x_2, \ldots, x_8$ are obtained by computer by solving eight linear equations derived as in [2].

DEFINITION 10. *$C(\beta,\gamma,\delta)$ is defined by*

$$C(\beta,\gamma,\delta) = \int_{X_1} d\theta \left(\int_0^\infty \left\{ P_t^1\left((r,\theta),(r,\theta)\right) - \frac{1}{4\pi t} \right\} r\, dr \right) \tag{16}$$

where $P_t^1(\,(r,\theta),(r_0,\theta_0)\,)$ is the heat kernel on CX_1.

In §§4 and 5 we will only consider the special case $\beta = 2\gamma = \delta = \pi$ for simplicity.

4. Asymptotic expansion of spectral function. Let M be a two-dimensional complex and $P_t(x,x)$ be the heat kernel on it. M may have singular intersection but on each simplex the metric is the canonical one. Then the spectral function Z_t is defined by $Z_t = \int_M P_t(x,x)\,dx$, where dx is the canonical surface measure. In this section we shall compute the asymptotic expansion of Z_t for some basic polyhedrons of length L and $2L$ depicted in Fig. 6.

For simplicity we suppose that III is a box (rectangular parallelepiped) containing one rectangle inside with rectangular intersection.

4.1. Asymptotic expansion of P_t. We observe first that P_t has an asymptotic expansion

$$P_t(x,x) = \frac{1}{4\pi t} + O(e^{-c/t}), \qquad c > 0 \text{ as } t \downarrow 0$$

except at a neighborhood of points P, S, and T. Next, it is easy to see that P_t has the same asymptotic expansion as q_t^α of the previous section in a neighborhood of P.

Note also that at the vertex S the singular surface is considered to be the generalized cone CX_1, where X_1 has already been defined at the end of §3. Therefore it is reduced to compute $C(\beta,\gamma,\delta)$ defined in (16).

Finally at the point T, the intersecting surface is considered to be the direct product $(0,\epsilon) \times X_2$ where X_2 is a T-shaped network depicted in Fig. 7.

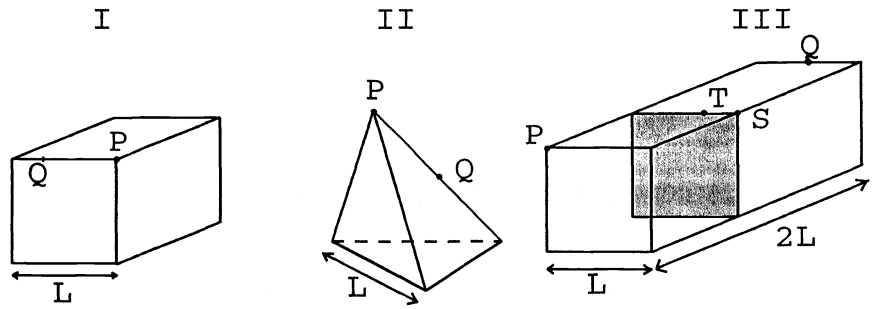


FIG. 6.

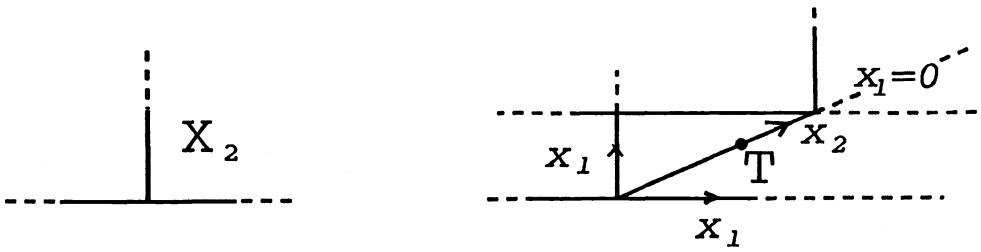


FIG. 7.

Therefore we know that for $x = (x_1, x_2)$, near $x_1 = 0$,

$$P_t(x,x) = \frac{1}{4\pi t}\left(1 - \frac{1}{3}e^{-x_1^2/t}\right) + O(e^{-c/t})$$

as $t \downarrow 0$ [2].

Consequently we get the following asymptotic expansions.

PROPOSITION 4. *Let $C(\cdot)$ and $C(\cdot,\cdot,\cdot)$ be as in* (14) *and* (16), *respectively.* Then *as $t \downarrow 0$ we get modulo $O(e^{-c/t})$, $c > 0$,*

(i) $Z_t^I \sim \frac{2L^2}{\pi t} + 8C(\frac{3}{2}\pi)$,

(ii) $Z_t^{II} \sim \frac{\sqrt{3}L^2}{4\pi t} + 4C(\pi)$,

(iii) $Z_t^{III} \sim \frac{11L^2}{4\pi t} - \frac{L}{\sqrt{4\pi t}} + 8C(\frac{3}{2}\pi) + 4C(2\pi, \pi, 2\pi)$ *where L is the length of edges of the models in Fig.* 6.

5. Numerical computation. In this section we use numerical methods to calculate $C(\alpha)$ as a function of α. The programs written were standard, using Simpson's rule and they were run on a VAX/VMS mainframe. The numerical results obtained in this way showed that the values of $C(\alpha)$ obtained by the formulas in Propositions 2 and 3 are in agreement for the special values of α for which Proposition 2 is valid.

We generate Table 1 and present the results graphically in Fig. 8.

TABLE 1

α	$C(\alpha)$ (reflection)	$C(\alpha)$ (numerical)
∞	—	$-\infty$
13.0	—	−0.132
10.8	—	−0.095
9.0	—	−0.061
8.0	—	−0.041
7.0	—	−0.018
6.28	—	0.0
5.0	—	0.038
4.0	—	0.078
3.14	—	0.129
2.09	—	0.222
1.57	0.313	0.313
1.31	—	0.383
1.05	0.486	0.484
0.79	0.656	0.655
0.63	0.825	0.824
0.52	0.993	0.992
0.45	1.161	1.160
0.39	1.328	1.328
0.35	1.495	1.495
0.31	1.663	1.663
0.20	2.664	2.664
⋮		⋮
0		∞

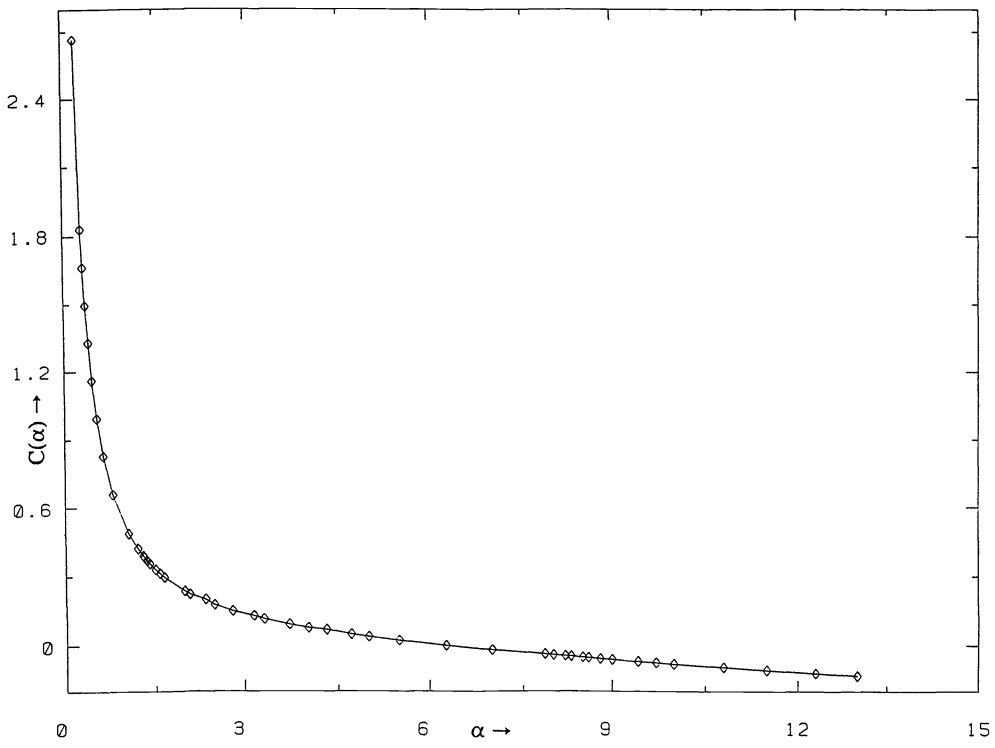


FIG. 8.

Acknowledgment. This note was prepared while the first and last authors were visiting Courant Institute of Mathematical Sciences. Special thanks are due to Professors H. P. McKean, Jr. and S. R. S. Varadhan, for stimulating discussions.

We are also grateful to the referee for comments which enabled us to make many improvements of the presentation.

REFERENCES

[1] S. F. EDWARDS, *Statistical mechanics with topological constraints:* I, Proc. Phys. Soc., 91 (1967), pp. 513–519.

[2] B. GAVEAU, M. OKADA, AND T. OKADA, *Explicit heat kernels on graphs and spectral analysis: Several complex variables,* Princeton Univ. Press, Math. Notes, 38 (1993), pp. 360–384.

[3] K. ITÔ AND H. P. MCKEAN, JR., *Diffusion processes and their sample paths,* Springer-Verlag, Berlin, 1965.

[4] S. K. NECHAEV, *Topological properties of a two-dimensional polymer chain in the lattice of obstacles*, J. Phys. A: Math. Gen., 21 (1988), pp. 3659–3671.

[5] N. SAITO AND Y. D. CHEN, *Statistics of a random coil chain in the presence of a point or line obstacle,* J. Chem. Phys., 59 (1973), pp. 3701–3709.

[6] G. N. WATSON, *A treatise on the theory of Bessel functions*, Cambridge Univ. Press, Cambridge, UK, 1944.

SIAM J. MATH. ANAL.
Vol. 25, No. 6 pp. 1577–1594, November 1994

009

ANALYTICAL AND NUMERICAL SOLUTIONS FOR A CLASS OF NONLOCAL NONLINEAR PARABOLIC DIFFERENTIAL EQUATIONS*

YANPING LIN†

Abstract. The aim of this paper is to study a class of nonlocal nonlinear parabolic boundary value problems. First the existence, uniqueness, and continuous dependence of the solution upon the data are demonstrated, and then finite difference methods, backward Euler and Crank–Nicolson schemes are studied. It is proved that both numerical schemes are stable and convergent to the real solution. The results of some numerical examples are presented, which demonstrate the efficiency and rapid convergence of the methods.

Key words. nonlocal, parabolic, existence, finite difference, stability

AMS subject classifications. 35L65, 45K05, 65M10

1. Introduction.

In this paper we consider the following parabolic equation of finding $u = u(x,t)$ such that

$$u_t + Au + f(u) = 0 \text{ in } Q_T, \qquad u = 0 \text{ on } \partial\Omega \times (0,T], \tag{1}$$

with the nonlocal time weighting initial condition

$$u(x,0) = \sum_{k=1}^{M} \beta_k(x) u(x,T_k) + \psi(x), \quad x \in \Omega, \tag{2}$$

where $Q_T = \Omega \times (0,T]$, $\Omega \subset R^d$ $(d \geq 1)$ is an open bounded domain with smooth boundary $\partial\Omega$, $T > 0$, and $0 < T_1 < T_2 < \cdots < T_M = T$, $\beta_k(x), \psi(x)$ and $f(u)$ are known smooth functions with respect to their variables, and A is a strongly elliptic operator

$$A = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(a_{i,j}(x)\frac{\partial}{\partial x_i}\right) + a(x), \quad x \in \Omega,$$

where $a(x)$, $a_{i,j}(x) = a_{j,i}(x)$ are known smooth functions, and satisfy for some positive constants a_0, $a_1 > 0$ that

$$a_0|\xi|^2 \leq \sum_{i,j=1}^{d} a_{i,j}\xi_i\xi_j \leq a_1|\xi|^2, \quad x \in \Omega, \quad \xi \in R^d.$$

The problem (1)–(2) can be viewed as a generalization of the standard time-periodic parabolic problem ($M = 1$, $\beta_1 = 1$, $\psi = 0$, and $T_1 = T$). It can also arise from the study of atomic reactors [2], [3] and some inverse heat conduction problems

* Received by the editors January 19, 1993; accepted for publication (in revised form) July 7, 1993.

† Department of Mathematics, University of Alberta, Edmonton, Alberta T6G 2G1 Canada (ylin@hilbert.math.ualberta.ca). The work of this author was supported in part by NSERC, Canada.

for determining the unknown physical parameters [9], [12]. Recently, the problem (1)–(2), linear or nonlinear, has been given a considerable attention by several authors ([2]–[12] and the references cited there), where some existence, uniqueness, and continuous dependence of the solution upon the data by both classical approach [10], [11] and abstract semigroup theory [8] are proved. It is noticed that the problem above also enjoys the maximum principle [2], [4], [9], [10] like the standard parabolic initial-boundary value problems [9], [14], [15]. For uniqueness the multiplier method [6] is also employed. A similar nonlocal initial-boundary value condition (2) for hyperbolic equation is also considered in [5]. For physical interpretation of (2) we refer to [2] and [4].

In summary the work mentioned above relies on the following assumption, that is, the weights $\beta_k(x), k = 1, 2, \ldots, M$, must satisfy the inequality

$$\sum_{k=1}^{M} |\beta_k(x)| \leq 1, \quad x \in \Omega. \tag{3}$$

See [2] and [10] for examples. The condition (3) is very restrictive, but critical to the methods employed in [2]–[12], either by maximum principle [9], [10], potential theoretical representation of the solution [9], [10], or abstract semigroup approach [8]. The reason for imposing condition (3) could be that these authors do not take the full advantages of the problem under consideration—a diffusion-type process, i.e., under some appropriate assumptions on the data, $||u(\cdot, t)||$ decays exponentially as $t \to \infty$, where $||\cdot||$ denotes the maximum norm or $L^2(\Omega)$ norm. The author of [7] considered the case of $\sum_{k=1}^{M} |\beta_k(x)| \leq c$, where c is a positive constant, and showed a solution exists via Schauder's fixed point theorem, but very technical restrictions on various Lipschitz constants and other conditions were assumed to compensate this relaxation. In the first part of this paper we shall study the problem (1)–(2) under a weaker (natural condition) assumption imposed on $\beta_k(x)$ (which is discussed later); the existence, uniqueness, and continuous dependence of the solution upon the data are proved.

If $f(u)$ is linear, then a purely algebraic condition on the weights $\beta_k(x)$ will be given, which is an optimal condition for the well-posedness of the problem (1)–(2). It is also shown by an example that the violation of the proposed algebraic condition below will result in nonuniqueness. The second part of this paper is devoted to the study of finite difference approximations to the solution of (1)–(2) by backward Euler and Crank–Nicolson methods. Both finite difference schemes are shown to be stable and convergent to the real solution with an expected rate of accuracy. In actual computations algorithms proposed will be used with a simple but effective (natural) iteration procedure.

The rate of convergence of the iterative procedure is also given. For convenience let us list our assumptions on the data.

(H1) The function $f(u)$ is smooth, $f'(u) + \lambda_0 \geq \mu > 0$ for $u \in R$, and $\psi(x) \in L^2(\Omega)$;

(H2) Assume that $\beta_k(x) \in L^\infty(\Omega)$ and $\beta_k^* = ||\beta_k||_{L^\infty(\Omega)}$ such that

$$\sum_{k=1}^{M} \beta_k^* e^{-\mu T_k} = \rho < 1,$$

where $\mu > 0$ is a positive constant and $\lambda_0 = \lambda_0(\Omega, a_0, a_1) > 0$ is the first eigenvalue

of the elliptic problem

$$Au + \lambda u = 0, \quad x \in \Omega, \tag{4}$$
$$u = 0, \quad x \in \partial\Omega.$$

Remark 1.1. It is clear that (H2) is a weaker assumption than (3).

The outline of this paper is as follows: In §2 the existence, uniqueness, and continuous dependence of the solution of the problem (1)–(2) are proved via the fixed point principle, and for the linear problem the eigenexpansion method will be employed. A general nonlocal (time integration) condition than (2) is also considered there. In §3 finite difference schemes are proposed, and then shown to be stable and convergent to the real solution. Finally numerical computations of some examples are reported in §4.

DEFINITION 1.1. *A function* $u \in C((0,T];L^2(\Omega)) \cap L^2((0,T);H_0^1(\Omega))$ *is said to be a solution of* (1)–(2) *if*

$$\int_{Q_T} \left(-u\phi_t + \sum_{i,j=1}^{d} a_{i,j} u_{x_i} v_{x_j} + auv + f(u)\phi \right) dxdt \tag{5}$$
$$= \int_\Omega \left(\sum_{k=1}^{M} \beta_k(x) u(x,T_k) + \psi(x) \right) \phi(x,0)dx$$

for all smooth function ϕ *such that* $\phi(x,t) \in C_0^\infty(\Omega)$, *and* $\phi(x,T) = 0$.

Here and throughout this paper the standard notation [13], [14] will be used.

2. Existence and uniqueness. As stated in §1 we shall prove several existence results and a nonuniqueness result in this section. First let us begin by the following result.

THEOREM 2.1. *Under assumptions* (H1)–(H2), *there exists a unique solution* $u \in C((0,T];L^2(\Omega)) \cap L^2((0,T);H_0^1(\Omega))$ *such that for some positive constant* $C > 0$, *independent of the data,*

$$||u(t)|| \le C(||\psi|| + |f(0)|), \quad t \in (0,T], \tag{6}$$

where $||\cdot||$ *is the* $L^2(\Omega)$ *norm.*

Proof. The proof is given by a fixed point principle.

Let mapping $S : L^2(\Omega) \to L^2(\Omega)$ be defined as follows: For $v \in L^2(\Omega)$,

$$Sv = \sum_{k=1}^{m} \beta_k(x) u(x,T_k,v) + \psi(x), \quad x \in \Omega,$$

where $u(x,t,v)$ is the solution of the following parabolic problem:

$$u_t + Au + f(u) = 0 \text{ in } Q_T, \tag{7}$$
$$u = 0 \text{ on } \partial\Omega \times (0,T],$$
$$u(x,0) = v(x) \in L^2(\Omega).$$

According to our assumptions (H1)–(H2) and the standard parabolic theory [13], [14], [1], [16], $u(x,t,v) \in L^\infty((0,T];L^2(\Omega)) \cap L^2((0,T);H_0^1(\Omega))$ exists such that

$$\int_{Q_T} \left(-u\phi_t + \sum_{i,j=1}^{d} a_{i,j} u_{x_i} v_{x_j} + auv + f(u)\phi \right) dxdt = \int_\Omega v(x)\phi(x,0)dx \tag{8}$$

for all smooth function ϕ such that $\phi(x,t) \in C_0^\infty(\Omega)$ and $\phi(x,T) = 0$. Using the regularity theory [1], [15] we find that $u(x,t,v) \in C((0,T];L^2(\Omega))$, and thus Sv is well defined.

Now let $u(x,t,v_m)$ be the solution of (7) with the initial data $v_m(x) \in L^2(\Omega)$, $m = 1,2$, respectively, and let $w = u(x,t,v_1) - u(x,t,v_2)$; it is easy to see that w satisfies

$$\begin{aligned} w_t + Aw + f^* w &= 0 \text{ in } Q_T, \\ w &= 0 \text{ on } \partial\Omega \times (0,T], \\ w(x,0) &= v_1(x) - v_2(x), \quad x \in \Omega, \end{aligned} \tag{9}$$

where

$$f^* = \int_0^1 \frac{df}{du}(\theta u(x,t,v_1) + (1-\theta)u(x,t,v_2))d\theta \geq \mu - \lambda_0.$$

Here assumption (H1) has been used.

Let $\{\lambda_l, \phi_l(x)\}$, $l = 0,1,\ldots$, be the eigenvalues and eigenfunctions of the problem (4) with $0 < \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_m \leq \cdots$ and be orthogonalized as $\int_\Omega \phi_l(x)\phi_m(x)dx = \delta_{lm}$, l, $m = 0,1,\ldots$. Let $w = \sum_{l=0}^\infty \phi_l(x)\eta_l(t)$ be the solution of (9) [14], [15] using the eigenpairs expansion.

If we multiply (9) by w and integrate over Ω, it is easy to see from the orthogonality, $f^* \geq \mu - \lambda_0$, and $\lambda_0 \leq \lambda_n$ for all $n \geq 1$ that

$$(w_t, w) + \mu(w,w) \leq 0, \tag{10}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$.

It is easy to see from (10) that

$$\frac{d}{dt}||w(t)|| + \mu||w(t)|| \leq 0.$$

Thus, we have

$$||w(t)|| \leq ||w(0)||e^{-\mu t} \quad \text{for} \quad t \in (0,T].$$

The definition of S and assumption (H2) now imply that

$$\begin{aligned} ||Sv_1 - Sv_2|| &= \left\| \sum_{k=1}^M \beta_k w(\cdot, T_k) \right\| \\ &\leq \sum_{k=1}^M \beta_k^* e^{-\mu T_k} ||w(0)|| \leq \rho ||v_1 - v_2||. \end{aligned}$$

Thus, we find that $S : L^2(\Omega) \to L^2(\Omega)$ is a contraction mapping since $0 < \rho < 1$.

That is, there is a unique $v \in L^2(\Omega)$ such that $Sv = v \in L^2(\Omega)$. Hence by (8) and (5) the solution of (7) with this unique fixed point $v(x)$ of S as the initial data will be the solution of (1)–(2).

For the estimate (6), we write (2) as

$$u_t + Au + f^{**}u = -f(0) \quad \text{with} \quad f^{**} = \int_0^1 \frac{df}{du}(\theta u)d\theta$$

and let $u = V + W$, where W satisfies

$$\begin{aligned} W_t + AW + f^{**}W &= -f(0) \text{ in } Q_T, \\ W &= 0 \text{ on } \partial\Omega \times (0,T], \\ W(x,0) &= 0 \quad x \in \Omega, \end{aligned}$$

and V satisfies

$$\begin{aligned} V_t + AV + f^{**}V &= 0 \text{ in } Q_T, \\ V &= 0 \text{ on } \partial\Omega \times (0,T], \\ V(x,0) &= \sum_{k=1}^{M} \beta_k(x) V(x,T_k) + \psi(x) + \sum_{k=1}^{M} \beta_k(x) W(x,T_k). \end{aligned} \tag{11}$$

It follows from the standard argument [14], [15] that $||W(t)|| \le C|f(0)|$, and from a similar argument to the above that $||V(t)|| \le ||V(0)||e^{-\mu t}$, thus from the initial condition of (11) and (H2),

$$||V(0)|| \le \frac{C}{1-\rho}\left(||\psi|| + \max_{0<t<T} ||W(t)||\right), \quad t \in (0,T].$$

Hence, (6) follows from the triangle inequality. □

Now we turn our attention to two special cases. That is to find $u(x,t)$ such that

$$\begin{aligned} u_t + Au &= 0 \text{ in } Q_T, \\ u &= 0 \text{ on } \partial\Omega \times (0,T], \\ u(x,0) &= \sum_{k=1}^{M} \beta_k(x) u(x,T_k) + g(x), \end{aligned} \tag{12}$$

where $g(x) \in L^2(\Omega)$ and $\beta_k(x) \in L^\infty(\Omega)$ for $k = 1,2,\dots,M$.

Let g be expanded using the eigenvalues and eigenfunctions of (4) by

$$g(x) = \sum_{m=0}^{\infty} g_m \phi_m(x) \quad \text{with} \quad g_l = \int_\Omega g(x)\phi_l(x)dx, \quad l = 0,1,\dots.$$

We seek the solution of the following form:

$$u(x,t) = \sum_{m=0}^{\infty} a_m \phi_m(x) e^{-\lambda_m t}, \tag{13}$$

where a_m, $m = 0,1,\dots$ are to be determined. Substituting (13) into the initial condition of (12), it follows from the linear independence of the eigenfunctions $\phi_m(x)$ that

$$a_m = a_m \left(\sum_{k=1}^{M} \beta_k e^{-\lambda_m T_m}\right) + g_m, \quad m = 0,1,\dots, \tag{14}$$

provided that $\beta_k(x) = \beta_k$ are constants for $k = 1,2,\dots,M$. If

$$1 - \sum_{k=1}^{M} \beta_k e^{-\lambda_m T_k} \neq 0 \quad \text{for} \quad m = 0,1,\dots, \tag{15}$$

then we have from (14) that

$$a_m = \frac{g_m}{1 - \sum_{k=1}^{M} \beta_k e^{-\lambda_m T_k}}, \quad m = 0, 1, \ldots, \tag{16}$$

and $\sum_{m=1}^{\infty} a_m^2 < \infty$ since $\sum_{m=1}^{\infty} g_m^2 < \infty$ and $\lambda_m \to \infty$ as $m \to \infty$. Hence we have obtained the following result.

THEOREM 2.2. *If β_k are constants and* (15) *is satisfied, then the solution of* (12) *exists and is unique, which is given by* (13) *and* (16), *and satisfies*

$$\int_\Omega u^2(x,t)dx \le C \int_\Omega g^2(x)dx, \quad t \in (0,T].$$

Proof. See the above analysis. □

Now let us consider the general case when $\beta_k(x)$ are not constants. Using the inital condition (12) we find

$$\sum_{m=0}^{\infty} a_m \phi_m(x) = \sum_{m=0}^{\infty} g_m \phi_m(x) + \sum_{m=1}^{\infty} a_m \phi_m(x) \sum_{k=1}^{M} \beta_k(x) e^{-\lambda_m T_k}.$$

Multiplying the above equation by $\phi_j(x)$ and integrating over Ω together with the orthogonality of $\phi_j(x)$, we obtain

$$a_j = \sum_{m=0}^{\infty} a_m E_{m,j} + g_j, \quad j = 0, 1, \ldots, \tag{17}$$

where

$$E_{m,j} = \sum_{k=1}^{M} e^{-\lambda_m T_k} \int_\Omega \beta_k(x) \phi_m(x) \phi_j(x) dx, \quad m, j = 0, 1, \ldots.$$

Let $a = (a_m)$ and $G = (g_m)$ be two infinite vectors and $E = (E_{m,j})$ be the infinite matrix, then we have formally from (17) that

$$(I - E)a = G.$$

Let [14]

$$l^2 = \left\{ y = (y_m) \mid \sum_{m=0}^{\infty} y_m^2 < \infty, \ y_m \in R \right\}.$$

Thus we have the following result.

THEOREM 2.3. *Assume that $g \in L^2(\Omega)$ and $\beta_k(x) \in L^\infty(\Omega)$ for $k = 1, 2, \ldots, M$. If the linear operator $E : l^2 \to l^2$ is such that the inverse operator $(I - E)^{-1}$ exists and is bounded, then the solution of* (12) *given by* (13) *exists and is unique with $a = (I - E)^{-1} G$.*

Proof. See the above. □

Finally, we show that the assumptions in Theorems 2.2 and 2.3 upon $\beta_k(x)$ are optimal, that is, the violation of this will result in nonuniqueness. For this purpose

let $M = 1$ and $g = 0$. If we select $\beta_1 = e^{\lambda_0 T}$ then $u = \phi_0(x)e^{-\lambda_0 t}$ will be a nonzero solution of

$$\begin{aligned} u_t + Au &= 0 \text{ in } Q_T, \\ u &= 0 \text{ on } \partial\Omega \times (0,T], \\ u(x,0) &= e^{\lambda_0 T}u(x,T), \quad x \in \Omega, \end{aligned} \tag{18}$$

with $\rho = 1$ in (H2).

But $u = 0$ is also a solution. Thus, the conditions given on $\beta_k(x)$ in Theorems 2.2 and 2.3 are optimal in this sense.

Remark 2.1. Since the proof of Theorem 2.1 only uses the existence results of parabolic problems, thus we see that if we consider nonlinear problem

$$\begin{aligned} u_t - div\, a(x,t,u,\nabla u) + f(x,t,u) &= 0 \text{ in } Q_T, \\ u &= 0 \text{ on } \partial\Omega \times (0,T], \end{aligned} \tag{19}$$

with the nonlocal initial condition (2), a and f smooth functions, and assume that $f_u(x,t,u) \geq 0$ and

$$(a(x,t,u,\nabla u) - a(x,t,v,\nabla v))\,\nabla(u-v) \geq \frac{\mu}{\lambda_0}|\nabla(u-v)|^2$$

for all x, t, u and v, where $\lambda_0 > 0$ is the first eigenvalue of (4) with $A = -\Delta$, then the assumption (H2) will imply the existence and uniqueness of the solution of the problem (19) and (2). The proof is similar to the above we therefore omit.

Remark 2.2. As in [10], [11], if (2) is replaced by

$$u(x,0) = \sum_{k=1}^{\infty} \beta_k(x)u(x,T_k) + \psi(x), \quad x \in \Omega, \tag{20}$$

where $\{T_k\} \subset (0,T]$, $\sum_{k=1}^{\infty} \beta_k^* < \infty$ is a convergent series and satisfies

$$\sum_{k=1}^{\infty} \beta_k^* e^{-\mu T_k} = \rho < 1,$$

where $\mu > 0$ is defined in §1, then the solution of (1) and (20) still exists and is unique following from the same proof of Theorem 2.1. We want to point out that $\inf_k\{T_k\} > 0$ is not required in our case, but in [10] and [11]. Finally let us consider the problem (1) with the following initial condition [2]:

$$u(x,0) = \sum_{k=1}^{M} \beta_k(x)F_k(u) + \psi(x), \quad x \in \Omega, \tag{21}$$

where

$$F_k(u) = \frac{1}{T_{2k} - T_{2k-1}} \int_{T_{2k-1}}^{T_{2k}} u(x,s)ds, \quad k = 1,2,\ldots,M,$$

and where $0 \leq T_1 < T_2 < \cdots < T_{2M} = T$. For problems (1) and (21) the assumption (H2) needs to be replaced by the following.

(H3) Assume that $\beta_k(x) \in L^\infty(\Omega)$ and $\beta_k^* = ||\beta_k||_{L^\infty(\Omega)}$ such that

$$\sum_{k=1}^{M} \frac{\beta_k^*}{T_{2k} - T_{2k-1}} \int_{T_{2k-1}}^{T_{2k}} e^{-\mu s} ds = \sum_{k=1}^{M} \frac{\beta_k^*}{(T_{2k} - T_{2k-1})\mu} \left(e^{-\mu T_{2k-1}} - e^{-\mu T_{2k}}\right) = \rho < 1.$$

where $\mu > 0$ is defined in assumption (H1).

Thus we have the following result.

THEOREM 2.4. *Under assumptions* (H1) *and* (H3), *problems* (1) *and* (21) *possess a unique solution* $u \in C((0,T]; L^2(\Omega)) \cap L^2((0,T); H_0^1(\Omega))$ *such that*

$$\int_{Q_T} \left(-u\phi_t + \sum_{i,j=1}^{d} a_{i,j} u_{x_i} v_{x_j} + auv + f(u)\phi \right) dxdt$$
$$= \int_\Omega \left(\sum_{k=1}^{M} \beta_k(x) F_k(u) + \psi(x) \right) \phi(x,0) dx$$

for all smooth function ϕ *such that* $\phi(x,t) \in C_0^\infty(\Omega)$ *and* $\phi(x,T) = 0$, *and satisfies the estimate* (6).

Proof. It follows from an argument similar to that given in Theorem 2.1. □

In fact, (21) is a more general condition than (2) since (2) can be viewed as a discrete version of (21), which will be seen in next section.

Remark 2.3. Like the nonuniqueness of (18), if we let $u = \phi_0(x)e^{-\lambda_0 t}$, where λ_0 and $\phi_0(x)$ are the first eigenvalue and eigenfunction of (4), then u is a nontrival solution of

$$\begin{aligned} u_t + Au &= 0 \text{ in } Q_T, \\ u &= 0 \text{ on } \partial\Omega \times (0,T], \\ u(x,0) &= \beta(x) \frac{1}{T} \int_0^T u(x,s) ds, \quad x \in \Omega \end{aligned}$$

with $\beta(x) = \lambda_0 T/(1 - e^{-\lambda_0 T})$ and $\rho = 1$. This example demonstrates that assumption (H3) is also optimal for problems (1) and (21).

3. Finite difference schemes. In this section we shall consider two finite difference schemes for problems (1)–(2), namely, the backward Euler method and the Crank–Nicolson method. For simplicity we study the following one-dimensional problem:

$$\begin{aligned} &u_t - u_{xx} + f(u) = g(x,t), \quad 0 < x < 1, \quad 0 < t \le T, \\ &u(0,t) = u(1,t) = 0, \quad 0 < t \le T, \\ &u(x,0) = \sum_{k=1}^{M} \beta_k(x) u(x, T_k), \quad 0 < x < 1. \end{aligned} \tag{22}$$

For this simple model the first eigenvalue of $w_{xx} = \lambda w$ with $w(0) = w(1) = 0$ is $\lambda_0 = \pi^2$. Without loss of generality we assume that $f(0) = 0$, $f'(u) \ge 0$ and $\beta_k(x)$ satisfies

$$\sum_{k=1}^{M} \beta_k^* e^{-\pi^2 T_k} = \rho < 1.$$

Let $\Delta x > 0$ small and $x_i = i\Delta x$, $i = 0, 1, \ldots, M_1$, where $M_1 = 1/\Delta x$ is a positive integer, be the partition of $[0, 1]$, and $0 = t_0 < t_1 < \cdots < t_N = T$ be the partition of $[0, T]$ such that $T_k = t_{N_k}$ are the nodes for some $N_k \in \{1, 2, \ldots, N\}$, $\tau_k = t_k - t_{k-1}$ and $\tau = \max_{1 \le n \le N} \tau_n$. Thus the backward Euler scheme reads as follows: Find $\{u_i^n\}$ such that

$$
\begin{aligned}
&\frac{u_i^n - u_i^{n-1}}{\tau_n} - \Delta_h u_i^n + f(u_i^n) = g_i^n, \quad 1 \le i \le M_1 - 1, \quad 1 \le n \le N, \\
&u_0^n = u_{M_1}^n = 0, \quad 0 \le n \le N, \\
&u_i^0 = \sum_{k=1}^{M} \beta_{k,i} u_i^{N_k}, \quad 1 \le i \le M_1 - 1,
\end{aligned} \tag{23}
$$

where $g_i^n = g(x_i, t_n)$, $\beta_{k,i} = \beta_k(x_i)$ for $i = 0, 1, \ldots, M_1$, $k = 1, 2, \ldots, M$ and $n = 0, 1, \ldots, N$, and

$$
\Delta_h u_i^n = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2}.
$$

Likewise the Crank–Nicolson scheme $\{u_i^n\}$ is defined by

$$
\begin{aligned}
&\frac{u_i^n - u_i^{n-1}}{\tau_n} - \Delta_h \frac{u_i^n + u_i^{n-1}}{2} + \frac{f(u_i^n) + f(u_i^{n-1})}{2} = g_i^{n+1/2}, \\
&1 \le i \le M_1 - 1, \quad 1 \le n \le N, \\
&u_0^n = u_{M_1}^n = 0, \quad 0 \le n \le N, \\
&u_i^0 = \sum_{k=1}^{M} \beta_{k,i} u_i^{N_k}, \quad 1 \le i \le M_1 - 1,
\end{aligned} \tag{24}
$$

where $g_i^{n+1/2} = (g_i^n + g_i^{n-1})/2$. We are now ready to state and prove our stability and convergence results.

THEOREM 3.1. *Let $\{u_i^n\}$ be the solution of the backward Euler scheme; then there exists $C > 0$, independent of Δt and Δx, and $\tau_0 > 0$ such that for all $0 < \tau \le \tau_0$*

$$
\max_{0 \le n \le N} ||U^n|| \le C \sum_{n=0}^{N} \tau_n ||F^n||,
$$

where $U^n = (u_1^n, \ldots, u_{M_1-1}^n)^T$, $F^n = (g_1^n, \ldots, g_{M_1-1}^n)^T$ and $||\cdot||$ denotes the natural norm on R^{M_1-1}.

Proof. Writing the finite difference equation (23) into the matrix form, we find

$$
(I + \tau_n(A + D_n))\, U^n = U^{n-1} + \tau_n F^n, \quad 1 \le n \le N,
$$

where $D_n = \operatorname{diag}(D_1^n, D_2^n, \ldots, D_{M_1}^n)$ with $D_i^n = \int_0^1 f'(\theta u_i^n) d\theta$ for all i and n, and

$$
A = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{pmatrix}.
$$

Letting $B_n = (I + \tau_n(A + D_n))^{-1}$, it follows that $U^n = B_n U^{n-1} + B_n \tau_n F^n$, and then one finds by induction that

$$U^n = \prod_{m=1}^{n} B_m U^0 + \sum_{m=1}^{n} \left(\prod_{l=m}^{n} B_l \right) \tau_m F^m. \tag{25}$$

But since A is a symmetric and D_n is diagonal, we have from the standard matrix theory [15] that

$$||B_n|| = ||(I + \tau_n(A + D_n))^{-1}|| \leq \max_{1 \leq i \leq M_1 - 1} \frac{1}{1 + \tau_n \nu_i^n},$$

where ν_i^n $(i = 1, 2, \ldots, M_1 - 1)$ are the eigenvalues of the matrix $A + D_n$, and that $\min_{1 \leq i \leq M_1,\, 0 \leq n \leq N} \{\nu_i^n\} \geq \pi^2$. Thus we have

$$||B_n|| \leq \frac{1}{1 + \pi^2 \tau_n}, \quad n = 0, 1, \ldots, N. \tag{26}$$

Therefore, we find from (25) and (26) that

$$||U^{N_k}|| \leq \prod_{m=1}^{N_k} \frac{1}{1 + \pi^2 \tau_m} ||U^0|| + \sum_{m=0}^{N_k} \tau_m ||F^m||. \tag{27}$$

From elementary calculus we know

$$\frac{1}{1+w} < \exp\left\{ \frac{-w}{1+w} \right\}, \quad w > 0,$$

so it is easy to see that

$$\begin{aligned} \prod_{m=1}^{N_k} \frac{1}{1 + \pi^2 \tau_m} &\leq \prod_{m=1}^{N_k} \exp\left\{ \frac{-\pi^2 \tau_m}{1 + \pi^2 \tau_m} \right\} \leq \exp\left\{ - \sum_{m=1}^{N_k} \frac{\pi^2 \tau_m}{1 + \pi^2 \tau_m} \right\} \\ &\leq \exp\left\{ \frac{-\pi^2 T_k}{1 + \pi^2 \tau} \right\}, \quad k = 1, 2, \ldots, M, \end{aligned}$$

where $\tau = \max_{1 \leq k \leq N} \tau_k$. But, we have for τ small that

$$\frac{\pi^2 T_k}{1 + \pi^2 \tau} = -\pi^2 T_k (1 - \pi^2 \tau + \cdots) \leq -\pi^2 T_k + \pi^4 T \tau \leq -\pi^2 T_k + \pi^4 T \tau$$

and then it follows

$$\prod_{m=1}^{N_k} \frac{1}{1 + \pi^2 \tau_m} \leq e^{-\pi^2 T_k} e^{\pi^4 T \tau}, \quad k = 1, 2, \ldots, M.$$

We thus obtain that

$$||U^{N_k}|| \leq e^{-\pi^2 T_k} e^{\pi^4 T \tau} ||U^0|| + \sum_{m=0}^{N_k} \tau_m ||F^m||.$$

Using the initial condition, we now find

$$||U^0|| \leq \sum_{k=1}^{M} \beta_k^* ||U^{N_k}|| \leq \sum_{k=1}^{M} \beta_k^* e^{-\pi^2 T_k} e^{\pi^4 T \tau} ||U^0|| + C \sum_{m=0}^{N} \tau_m ||F^m||.$$

Hence, there exists a small $\tau_0 >$ such that for all $0 < \tau \leq \tau_0$,

$$||U^0|| \leq \frac{C}{1 - \rho e^{\pi^4 T\tau}} \sum_{m=0}^{N} \tau_m ||F^m|| \tag{28}$$

due to $0 < \rho < 1$. In fact τ_0 can be selected to be $\tau_0 = -\log\rho/(\pi^4 T)$. Finally, substituting (28) into (27), we obtain

$$\begin{aligned} \max_{0\leq n\leq N} ||U^n|| &\leq C\left(||U^0|| + \sum_{m=0}^{N} \tau_m ||F^m||\right) \\ &\leq C \sum_{m=0}^{N} \tau_m ||F^m||. \end{aligned}$$

Hence, Theorem 3.1 is proved. □

THEOREM 3.2. *Let $\{u_i^n\}$ be the solution of Crank–Nicolson scheme, then there exists $C > 0$, independent of Δt and Δx, and $\tau_0 > 0$ such that for all $0 < \tau \leq \tau_0$,*

$$\max_{0\leq n\leq N} ||U^n|| \leq C \sum_{n=0}^{N} \tau_n ||F^n||.$$

Proof. Following the proof of Theorem 3.1, we write (24) into the matrix form

$$\begin{aligned} \left(I + \frac{1}{2}\tau_n(A + D_n)\right) U^n = \left(I - \frac{1}{2}\tau_n(A + D_{n-1})\right) U^{n-1} \\ + \tau_n \frac{F^n + F^{n+1}}{2}, \quad 1 \leq n \leq N, \end{aligned} \tag{29}$$

with D_n defined as before. Also (29) can be written in a compact form

$$U^n = K_n U^{n-1} + L_n \tau_n \frac{F^n + F^{n+1}}{2}, \quad n = 1, 2, \ldots, N,$$

where

$$\begin{aligned} K_n &= \left(I + \frac{1}{2}\tau_n(A + D_n)\right)^{-1} \left(I - \frac{1}{2}\tau_n(A + D_{n-1})\right), \\ L_n &= \left(I + \frac{1}{2}\tau_n(A + D_n)\right)^{-1}. \end{aligned}$$

From the standard matrix theory [15], we have that

$$||K_n|| \leq \max_{1\leq i\leq M_1 - 1} \left(\frac{1 - \frac{1}{2}\tau_n \nu_i^n}{1 + \frac{1}{2}\tau_n \nu_i^n}\right) \leq \frac{1 - \frac{1}{2}\pi^2 \tau_n}{1 + \frac{1}{2}\pi^2 \tau_n}, \quad n = 1, \ldots, N.$$

Let w_m be such that

$$\frac{1}{1 + w_m} = \frac{1 - \frac{1}{2}\pi^2 \tau_m}{1 + \frac{1}{2}\pi^2 \tau_m} \quad \text{with} \quad w_m = \frac{\pi^2 \tau_m}{1 - \frac{1}{2}\pi^2 \tau_m};$$

it follows from a simple calculation as above that

$$\prod_{m=1}^{n} \frac{1-\frac{1}{2}\pi^2\tau_m}{1+\frac{1}{2}\pi^2\tau_m} \le \prod_{m=1}^{n} \frac{1}{1+w_m} < \prod_{m=1}^{n} \exp\left\{\frac{-w_m}{1+w_m}\right\}$$
$$\le \exp\left\{-\sum_{m=1}^{n} \frac{\pi^2\tau_m}{1+\frac{1}{2}\pi^2\tau_m}\right\} \le e^{-\pi^2 t_n} e^{\frac{1}{2}\pi^4 T\tau}, \quad 1 \le n \le N.$$

Thus, we find that

$$\prod_{m=1}^{n} ||K_m|| \le e^{-\pi^2 t_n} e^{\frac{1}{2}\pi^4 T\tau}, \quad 1 \le n \le N.$$

Hence, the remainder of the proof follows from an argument similar to that given in Theorem 3.1. We omit the details. □

We are now ready to state our convergence results.

THEOREM 3.3. (I) *Let u be the solution of* (22) *such that $u \in C^{4,2}(\bar{Q}_T)$ and u_i^n be the backward Euler solution* (23). *Then there exists a positive constant $C > 0$, independent of the step sizes Δx and Δt, and $\tau > 0$ small such that for all $0 < \tau \le \tau_0$*

$$\max_{0\le n\le N} ||u(t_n) - U^n|| \le C(\tau + \Delta x^2);$$

(II) *Let u be the solution of* (22) *such that $u \in C^{4,3}(\bar{Q}_T)$ and u_i^n be the Crank-Nicolson solution* (24). *Then there exists a positive constant $C > 0$, independent of the step sizes Δx and Δt, and $\tau_0 > 0$ small such that for all $0 < \tau \le \tau_0$*

$$\max_{0\le n\le N} ||u(t_n) - U^n|| \le C(\tau^2 + \Delta x^2),$$

where

$$||u(t_n) - U^n|| = \sqrt{\sum_{i=1}^{M_1-1} |u(x_i, t_n) - u_i^n|^2 \, \Delta x}, \qquad n = 0, 1, \ldots, N.$$

Proof. Let $e_i^n = u(x_i, t_n) - u_i^n$ be the error, then e_i^n satisfies (23) with $F_i^n = O(\tau + \Delta x^2)$ for the backward Euler solution or (24) with $F_i^n = O(\tau^2 + \Delta x^2)$ for the Crank–Nicolson solution; then the results follow from the stability estimates of Theorems 3.1 and 3.2. □

Remark 3.1. We only give the stability proofs for one-dimensional problem above, but our method of proof can easily be extended to any d-dimensional problems without any technical complications. The assumption on the nonlinear function $f(u)$ can also be relaxed to assumption (H1) in Theorem 2.1, that is, $f'(u) + \lambda_0 = \mu > 0$ where λ_0 is the first eigenvalue of the problem (4).

Remark 3.2. For finite difference approximations to the solution of (1) and (21), we see that if $F_k(u)$ $(k = 1, 2, \ldots, M)$ is discretized by some numerical quadrature formula

$$F_k(u) = \frac{1}{T_{2k} - T_{2k-1}} \sum_{j=0}^{M_k} w_{k,j} u(x, \sigma_j) + \quad O(Error), \qquad k = 1, 2, \ldots, M,$$

where $T_{2k-1} \le \sigma_j \le T_{2k}$ for $j = 0, 1, \ldots, M_k$ and $k = 1, 2, \ldots, M$, we see now that the discrete version of (21) is as the same as the condition (2) except for different weights

functions $\beta(x)'s$. Therefore the stability of the finite difference schemes of problems (1) and (21), backward Euler or Crank–Nicolson, can be proved in the same manner above, and we omit the detailed discussions. We have proved that both backward Euler and Crank–Nicolson schemes are stable and convergent to the real solution as $\Delta x,\ \Delta t \to 0$. Obviously if a direct method is used to solve $\{u_i^n\}$ one must deal with a $(M_1-1)(N+1)\times(M_1-1)(N+1)$ nonlinear system. Fortunately a simple iteration procedure can be used because of the parabolic nature of the problem, that is, let $(u_i^0)^{(0)}=0$ be the initial guess, then the $(u_i^0)^{(l+1)}$ is defined as the following:

$$(u_i^0)^{(l+1)}=\sum_{k=1}^{M}\beta_{k,i}(u_i^{N_k})^{(l)},\quad l=0,1,\ldots,\quad i=1,2,\ldots,M_1-1,$$

where $(u_i^n)^{(l)}$ is the finite difference solution of backward Euler or Crank–Nicolson method with the initial data $(u_i^0)^{(l)}$. We have the following convergence estimates for this iterative procedure.

THEOREM 3.4. *Let $r=\rho e^{\pi^4 T\tau_0}<1$ where $\tau_0>0$ is selected as in Theorems* 3.2 *or* 3.3 *with respect the finite difference schemes, then it holds*

$$\|(U^0)^{(l+1)}-(U^0)^{(l)}\|\le r\|(U^0)^{(l)}-(U^0)^{(l-1)}\|,\quad l=1,2,\ldots. \tag{30}$$

Proof. It follows from an argument similar to that given above. □

A direct consequence of Theorem 3.4 is the following estimate:

$$\max_{0\le n\le N}\|(U^n)^{(l+1)}-(U^n)^{(l)}\|\le r^l\|(U^0)^{(1)}-(U^0)^{(0)}\|,\quad l=1,2,\ldots. \tag{31}$$

In fact (31) follows from (30) and the discrete stability estimate of backward Euler or Crank–Nicolson schemes for parabolic equations [16].

In theory above it seems that the time step size τ must be selected so small such that $\rho e^{\pi^4 T\tau_0}<1$, which is the convergence rate of iteration procedure, but in actual computations the restriction of τ can be relaxed in a great deal and is not important. Therefore the restriction on τ may be due to the method used in deriving the estimates in the proofs of Theorems 3.2 and 3.3, It is the author's conjecture that the stability of finite difference schemes proposed above are independent of the step sizes just like the stability of parabolic finite difference methods [16]. Numerical examples in next section show that only a few iterations are needed in order to obtain an acceptable solution.

4. Numerical examples. In this section we report some results of our numerical calculations using finite difference scheme proposed in the previous sections.

Example 1. Assume that $Q_T=(0,1)^2\times(0,T]$. We consider the following problem: Find $u=u(x,t)$ such that

$$\begin{aligned} u_t-\Delta u&=f(x,t)\ \text{ in }\ Q_T,\\ u&=0\ \text{ on }\ \partial\Omega\times(0,T],\end{aligned}$$

with the following nonlocal condition

$$u(x,0)=\beta_1 u(x,T_1)+\beta_2 u(x,T_2)+\psi(x),\quad x\in\Omega,$$

where $T_2=T=1$ and $0<T_1<T_2$. Let $u=\sin(\pi x)\sin(\pi y)e^{-t}$ be the solution of (22) with data $\psi(x)=\sin(\pi x)\sin(\pi y)(1-\beta_1 e^{T_1}-\beta_2 e^{T_2})$ and $f(u)=0$ and $g(x,t)=$

TABLE 1
The Crank–Nicolson Scheme with $T_1 = 0.5$.

Δx	Δt	L^∞ error	L^2 error	# Iter.	β_1	β_2
0.1	0.1	3.93×10^{-2}	1.34×10^{-2}	3	1	1
0.05	0.05	1.95×10^{-2}	7.14×10^{-3}	4	1	1
1/30	1/30	1.31×10^{-2}	4.87×10^{-3}	4	1	1
1/40	1/40	9.98×10^{-3}	3.69×10^{-3}	4	1	1
0.1	0.1	4.01×10^{-2}	1.35×10^{-2}	3	1	-1
0.05	0.05	2.92×10^{-2}	7.87×10^{-3}	4	1	-1
1/30	1/30	2.20×10^{-2}	5.5×10^{-3}	4	1	-1
0.05	0.05	5.59×10^{-2}	9.96×10^{-3}	4	5	-1
0.05	0.05	5.45×10^{-2}	9.83×10^{-3}	4	5	-5
0.05	0.05	1.11×10^{-1}	1.39×10^{-2}	4	1	20
0.05	0.05	1.19×10^{-1}	1.48×10^{-2}	4	20	20

TABLE 2
The Crank–Nicolson Scheme with $\Delta x = \Delta t = 0.05$.

T_1	# Iter.	L^∞ error	L^2 error
0.8	2	2.93×10^{-2}	7.88×10^{-3}
0.5	3	2.92×10^{-2}	7.88×10^{-3}
0.4	3	2.92×10^{-2}	7.87×10^{-3}
0.3	3	2.96×10^{-2}	7.87×10^{-3}
0.2	4	2.91×10^{-2}	7.87×10^{-3}
0.1	7	2.99×10^{-2}	7.92×10^{-3}
0.05	11	3.24×10^{-2}	8.09×10^{-3}

$(-1 + 2\pi^2)\sin(\pi x)\sin(\pi y)e^{-t}$ for any two constants β_1 and β_2. The Crank–Nicolson method is used in this example. Table 1 shows the computational results of errors in maximum norm, L^2 norm and the numbers of iterations via various Δx, Δt, β_1 and β_2. In order to keep the accuracy the stopping criteria TOL of the iteration tolerance is chosen by $TOL = 0.5(\Delta x^2 + \Delta t^2)$ in all our computations, which is half of the truncation error. That is, let the initial guess $u^0 = 0$ and if

$$||(U^0)^{(l+1)} - (U^0)^{(l)}|| \leq TOL$$

for some l, then $(u_i^0)^{l+1}$ and $(u_i^N)^l$ will be accepted as the numerical initial value and final value, respectively, and the computations will be terminated. We see from Table 1 that the accuracy of numerical calculation not only depends on the step sizes, but also on the discontinuity.

Example 2. In this example we shall see the impact of T_1, which affects the numbers of iterations. For this purpose we use the same example as in Example 1 with fixed $\Delta x = \Delta t = 0.05$, $\beta_1 = 1$, and $\beta_2 = -1$. When T_1 is "far" away from 0, such as $T_1 = 0.5$ and $T_1 = 0.8$, only two iterations are needed. This is because that no matter what initial value we start with, solution $u(x,t)$ approaches the steady state quickly as t increases, thus $u(x,t)$ can be calculated accurately for larger t which in turn gives a good next step initial update data. When T_1 gets close to 0, the numbers of iterations increase in general. Table 2 shows this phenomenon.

Example 3. In this example we consider a simple nonlinear model problem of finding $u = u(x,t)$ such that

$$\begin{aligned} &u_t - u_{xx} + f(u) = g(x,t), \quad 0 < x < 1, \quad 0 < t \leq T, \\ &u(0,t) = u(1,t) = 0, \quad 0 < t \leq T, \\ &u(x,0) = \beta(x)u(x,T) + \psi(x), \quad 0 < x < 1, \end{aligned}$$

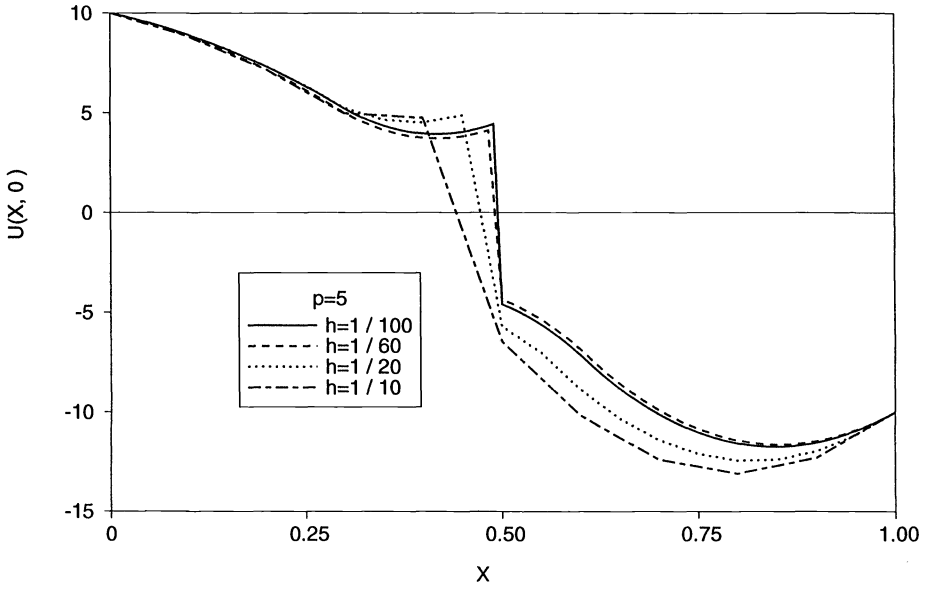


FIG. 1. *The initial values for $p = 5$ with $h = \Delta t = \Delta x$: rough data case.*

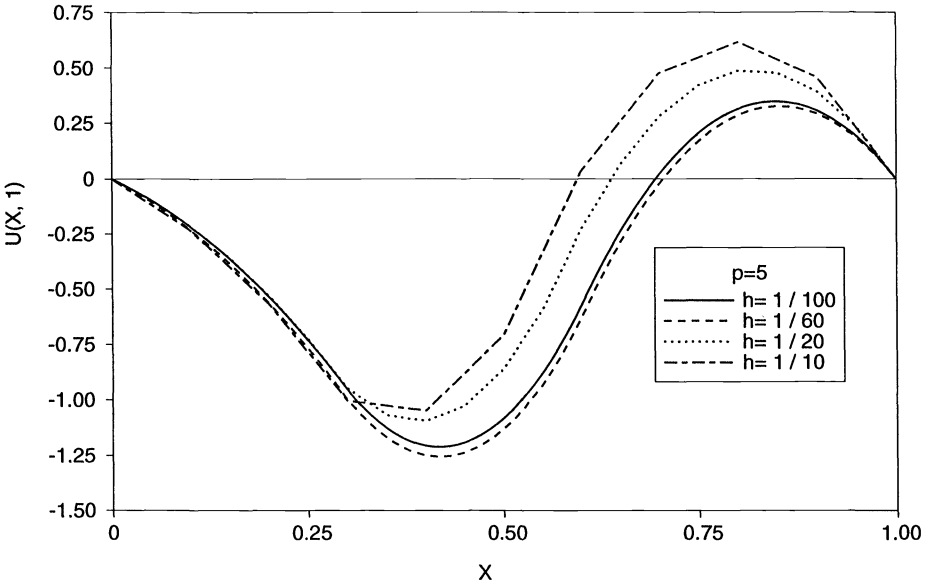


FIG. 2. *The final values for $p = 5$ with $h = \Delta t = \Delta x$: rough data case.*

where

$$\beta(x) = \begin{cases} 5, & 0 < x \le 0.5, \\ -5, & 0.5 < x < 1, \end{cases} \qquad \psi(x) = \begin{cases} 10, & 0 < x \le 0.5, \\ -10, & 0.5 < x < 1, \end{cases}$$

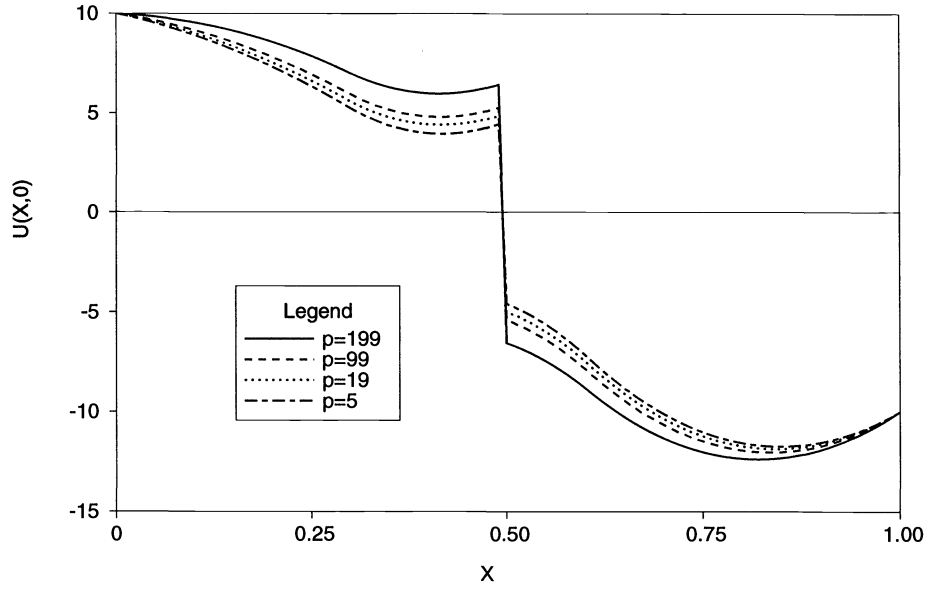


FIG. 3. *The initial values for* $h = \Delta t = \Delta x = 0.01$ *via* p: *rough data case.*

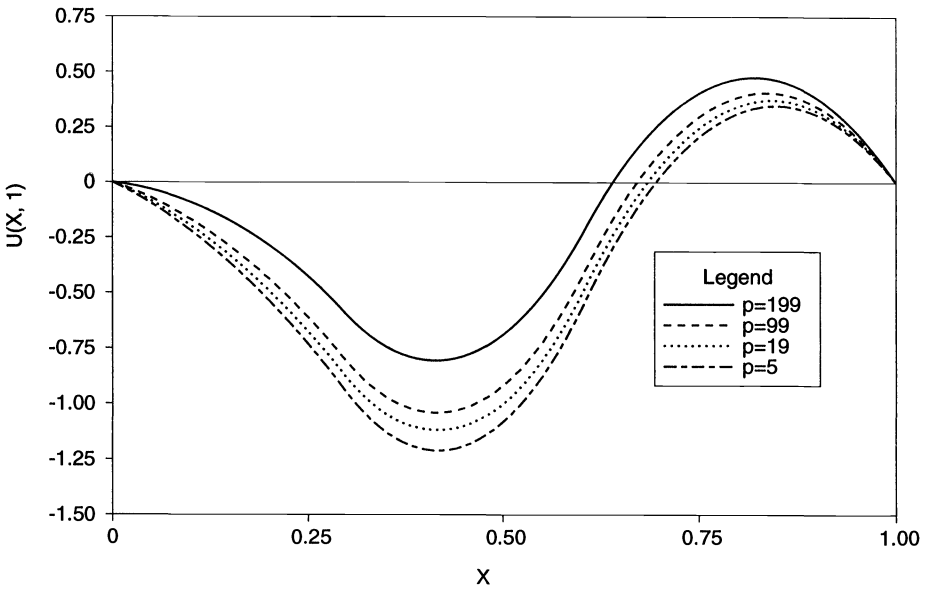


FIG. 4. *The final values for* $h = \Delta t = \Delta x = 0.01$ *via* p: *rough data case.*

and

$$f(u) = u^p, \quad p = 1, 3, 5, \ldots, \quad \text{and} \quad g(x,t) = \begin{cases} 10, & 0 < x \le 0.3, \\ -40, & 0.3 < x \le 0.6, \\ 30, & 0.6 < x < 1. \end{cases}$$

The backward Euler scheme to be used in this example, which is different from

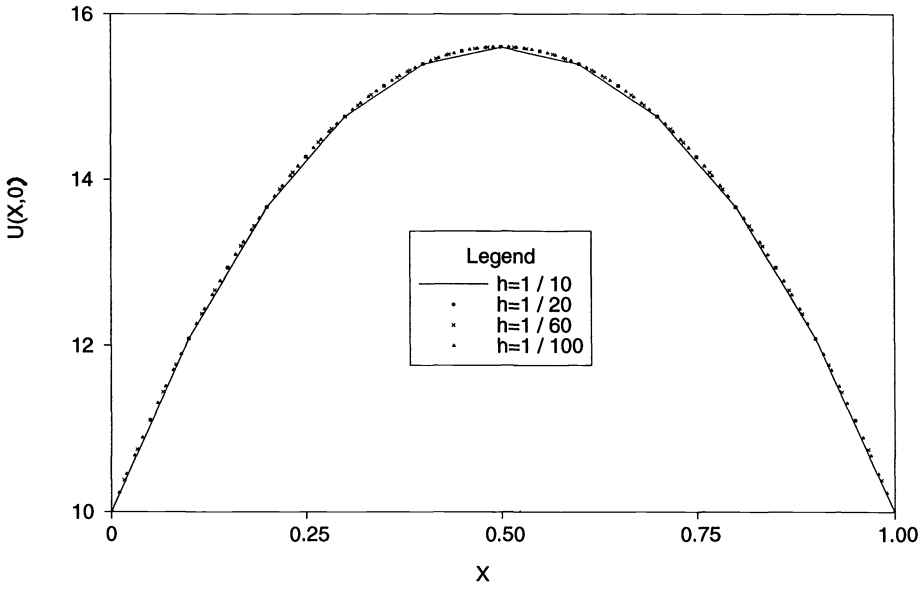


FIG. 5. *The initial values for* $h = \Delta t = \Delta x = 0.01$ *via p: smooth data case.*

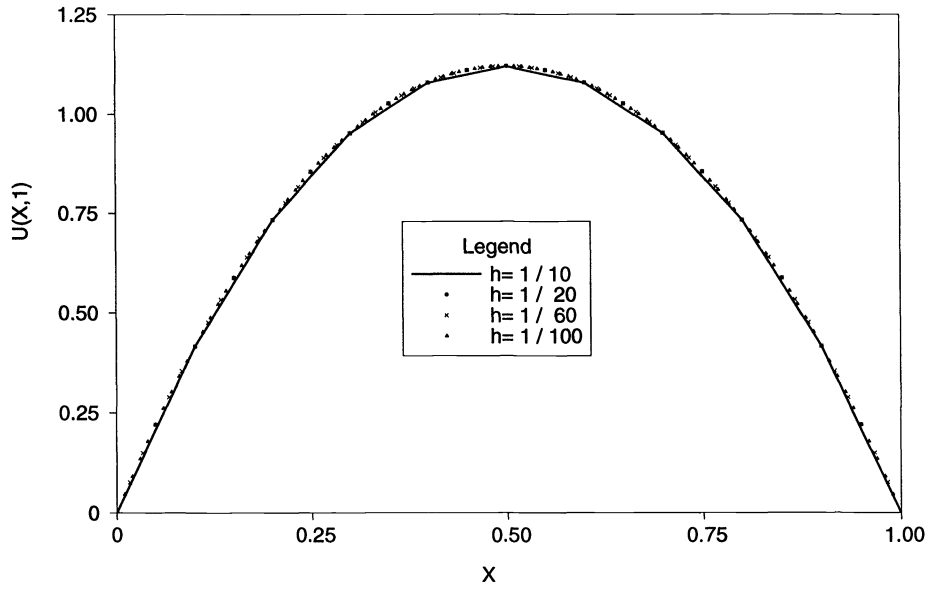


FIG. 6. *The final values for* $h = \Delta t = \Delta x = 0.01$ *via p: smooth data case.*

(23) in §3, is defined as

$$\frac{u_i^n - u_i^{n-1}}{\tau_n} - \Delta_h u_i^n + (u_i^{n-1})^{p-1} u_i^n = g_i^n, \quad 1 \le i \le M_1 - 1, \quad 1 \le n \le N,$$
$$u_0^n = u_{M_1}^n = 0, \quad 0 \le n \le N,$$
$$u_i^0 = \beta_i u_i^N, \quad 1 \le i \le M_1 - 1.$$

It can be proved in the same manner as in the previous section that this scheme is also stable. In this example $h = \Delta x = \Delta t$ is selected. For $p = 5$, Fig. 1 shows that the initial values obtained by using the various step sizes $h = 1/100$, $1/60$, $1/20$, and $1/10$. The numbers of iterations of all cases are about two or three. It also can be seen that small step sizes are needed in order to obtained an accepted solutions, which is due to the discountinuity of the data. Fig. 2 shows the final values obtained for the above example. For $h = 0.01$, the Fig. 3 demonstrates the impact of the power p in the nonlinear function $f(u)$ on the initial values, and the Fig. 4 on the final values.

Example 4. As in Example 3 we now take $\beta(x) = 5$, $g(x,t) = 10$, $\psi(x) = 10$ and $f(u) = u^5$, that is, smooth data is used. Figs. 5 and 6 show the computational results for the initial and final values. It is noticed that the numerical solutions are almost the same for all step sizes, or the numerical solution is not sensitive to the step sizes since the real solution is very smooth in this case. Examples 3 and 4 imply that for the nonsmooth data problem special care needs to be given to the step sizes near the discontinuity of the data in order to obtain a reasonable numerical solution without solving a relative large matrix system.

Acknowledgments. The author thanks the referee for comments and suggestions on the original manuscript which led to its improvement.

REFERENCES

[1] L. Boccardo and T. Gallouet, *Non-linear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 78 (1989), pp. 149–169.

[2] L. Bysezewski, *Strong maximum principle for parabolic nonlinear problems with nonlocal inequalities together with arbitrary functionals*, J. Math. Anal. Appl., 156 (1991), pp. 457–470.

[3] ———, *Theorem about the existence and uniqueness of solution of a semilinear evolution nonlocal Cauchy problem*, J. Math. Anal. Appl., 162 (1991), pp. 494–505.

[4] ———, *Strong maximum and minimum principles for parabolic problems with nonlocal inequalites*, Z. Angew. Math. Mech., 70 (1990), pp. 202–206.

[5] ———, *Theorem about existence and uniqueness of continuous solution of nonlocal problem for nonlinear hyperbolic equation*, Appl. Anal., 40 (1991), pp. 173–180.

[6] ———, *Uniqueness of solutions of parabolic semilinear nonlinear boundary problems*, J. Math. Anal. Appl., 165 (1992), pp. 427–478.

[7] ———, *Existence of a solution of a Fourier nonlocal quasilinear parabolic problem*, J. Appl. Math. Stochastic Anal., 5 (1992), pp. 43–68.

[8] L. Bysezewski and V. Lakshmikantham, *Theorem about the existence and uniqueness of a solution of a nonlocal abstract Cauchy problem in a Banach space*, Appl. Anal., 40 (1990), pp. 11–19.

[9] J. Cannon, *The One-Dimensional Heat Equation*, in Encyclopedia of Mathematics and Its Applications, Vol 23, Addison-Wesley Publishing Company, Menlo Park, CA, 1984.

[10] J. Chabrowski, *On non-local problems for parabolic equations*, Nagoya Math. J., 93 (1984), pp. 109–131.

[11] ———, *On the non-local problem with a functional for parabolic equation*, Funkcial. Ekvac., 27 (1984), pp. 101–123.

[12] J. Chadam and H-M. Yin, *Determination of an unknown function in a parabolic equation with an overspecified condition*, Math. Meth. Appl. Sci., 13(1990), pp. 421–430.

[13] A. Dall'aglio and L. Orsina, *Existence results for some nonlinear parabolic equations with nonregular data*, Differential Integral Equations, 5 (1992), pp. 1335–1354.

[14] O. Ladyzenskaja, V. Solonnikov and N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Transl., American Mathematical Society, Providence, RI, 1968.

[15] J. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems*, Springer-Verlag, New York, 1972.

[16] G. Marchuk, *Methods of Numerical Mathematics*, Springer-Verlag, New York, 1975.

[17] H. Yin, *Weak and classical solutions of some nonlinear Volterra intergrodifferential equations*, IMA preprint series 920, 1992.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1595–1621, November 1994

010

MONODROMY REPRESENTATIONS OF SYSTEMS OF DIFFERENTIAL EQUATIONS FREE FROM ACCESSORY PARAMETERS *

YOSHISHIGE HARAOKA†

Abstract. In a paper by Haraoka [SIAM J. Math. Anal., 25(1994), pp. 1203–1226], by following Okubo's theory [K. Okubo, *Seminar reports of Tokyo Metropolitan University,* 1987], the canonical forms of all generic classes of Fuchsian systems of differential equations on $\mathbf{P}^1(\mathbf{C})$ free from accessory parameters are obtained. Among them the explicit forms of six classes are new. In the present paper monodromy representations of the systems in the six classes are calculated. The technique employed is similar to one used to obtain the canonical forms. Hermitian forms invariant under those monodromy groups are also calculated. It turns out that the space of the invariant Hermitian forms for each system is real one-dimensional.

Key words. accessory parameter, monodromy representation, invariant Hermitian form

AMS subject classifications. 33C20, 33C65, 33E30

In his theory on Fuchsian systems of differential equations on the complex projective line, K. Okubo has shown that, if a Fuchsian system of differential equations is free from accessory parameters, we can compute its monodromy representation ([O]). Noting that in general we have no way to compute monodromy representations, we expect that systems free from accessory parameters are good systems of differential equations and will define new special functions after the Gauss hypergeometric function.

In our previous work [H2] we have determined all Fuchsian systems on $\mathbf{P}^1(\mathbf{C})$ that are irreducible and free from accessory parameters; we have followed Okubo's theory and used Yokoyama's [Y] classification theorem to obtain our result. This paper is devoted to computing monodromy representations of those systems. Here we also follow Okubo's theory and employ a technique similar to one in [H2].

Let S be a finite subset of $\mathbf{P}^1(\mathbf{C})$. Take a point x_0 in $\mathbf{P}^1(\mathbf{C}) \setminus S$. Let

$$\frac{dY}{dx} = A(x)\,Y \tag{F}$$

be a system of differential equations on $\mathbf{P}^1(\mathbf{C})$ of rank n with the set of singular points S. Fix a fundamental matrix solution $Y(x)$ defined in a neighborhood of x_0. The *monodromy representation* of the system (F) with respect to $Y(x)$ is a group homomorphism

$$R : \pi_1(\mathbf{P}^1(\mathbf{C}) \setminus S, x_0) \to \mathrm{GL}(n, \mathbf{C}),$$

which is defined by

$$Y(x)^{\gamma} = Y(x) \cdot R([\gamma])$$

for any loop γ in $\mathbf{P}^1(\mathbf{C}) \setminus S$ with the base point x_0, where $Y(x)^{\gamma}$ denotes the analytic continuation of $Y(x)$ along the loop γ.

According to Yokoyama's classification there are eight classes of systems of differential equations which are irreducible and free from accessory parameters: Systems

* Received by the editors December 31, 1992; accepted for publication August 4, 1993.

† Department of Mathematics, Faculty of General Education, Kumamoto University, Kumamoto, 860, Japan.

(I), (I*), (II), (II*), (III), (III*), (IV), and (IV*). System (I) is known to be transformed into the generalized hypergeometric equation, whose monodromy representation is known ([O], [OTY], [L], [BH]). System (I*) is known to be transformed into the Pochhammer equation, whose monodromy representation is also known ([M], [TB], [H1]). Then we deal with the other systems (II), (II*), (III), (III*), (IV), and (IV*). We note that the monodromy representations of systems (II) and (II*) of rank 4 have been obtained in [ST] and [S], respectively.

Systems (II), (III), (IV), (II*), (III*) and (IV*) are studied in §§1.1, 1.2, 1.3, 2.1, 2.2, and 2.3, respectively. For each system, we specify a fundamental matrix solution, give generators of monodromy representations (Theorems 1, 3, 5, 7, 9, 11) and, in the case that the parameters of the system are real numbers, determine invariant Hermitian forms for the monodromy group (Theorems 2, 4, 6, 8, 10, 12). As a result, it turns out that, for every system, the dimension of the space of invariant Hermitian forms over the field of the real numbers is one.

Theorems 1 and 2 are proved in §1.1. We can prove Theorems 3, 5, 7, 9, 11 by combining the methods of proving Theorem 1 and Theorem in [H2], so in this paper we omit their proofs. Similarly, the proofs of Theorems 4, 6, 8, 10, 12 are omitted, since they can be proved in ways analogous to the proof of Theorem 2.

Sections 1.1–1.3 are concerned with representations of the fundamental group of $\mathbf{P}^1(\mathbf{C}) \setminus \{\text{three points}\}$. Sections 2.1–2.3 are concerned with representations of the fundamental group of $\mathbf{P}^1(\mathbf{C}) \setminus \{\text{four points}\}$.

Notation.

$\mathbf{Z}_{<0}$: the set of negative integers.
I_k : the identity matrix of size k, for $k \in \mathbf{N}$.
O : zero matrix of an appropriate size.
$M(k, l; \mathbf{C})$: the set of $k \times l$ matrices with entries in $\mathbf{C}$, for $k, l \in \mathbf{N}$.
(α_{ij}) : matrix whose (i, j)-entry is α_{ij} for every i, j.
e_i : column vector which has the only nonzero entry 1 in the ith position.
$B(t, r) := \{x \in \mathbf{C}; |x - t| < r\}$, for $t \in \mathbf{C}$ and $r > 0$.
$e(\alpha) := \exp(2\pi\sqrt{-1}\alpha)$, for $\alpha \in \mathbf{C}$.

1. System (II), system (III), and system (IV). Let t_1, t_2 be mutually distinct points in $\mathbf{C}$. We set

$$B_1 = B(t_1, |t_2 - t_1|),$$
$$B_2 = B(t_2, |t_1 - t_2|),$$
$$W = B_1 \cap B_2.$$

Since $t_1 \neq t_2$, W is a nonempty, simply connected domain in $\mathbf{C}$. We take any point $x_0 \in W$ as a base point of $\mathbf{P}^1(\mathbf{C}) \setminus \{t_1, t_2, \infty\}$; for example, we can take

$$x_0 = \frac{t_1 + t_2}{2}.$$

For $i = 1, 2$, let γ_i be a loop that starts from x_0, encircles t_i once in the positive direction, and ends at x_0, not encircling $t_{i'}$ $(i' \neq i)$. Then the fundamental group $\pi_1(\mathbf{P}^1(\mathbf{C}) \setminus \{t_1, t_2, \infty\}, x_0)$ is generated by the homotopy classes $[\gamma_1]$ and $[\gamma_2]$.

This notation is fixed throughout this section.

1.1. Monodromy representation of system (II). Let n be an even integer equal to or greater than 4. We set $n = 2m$ with $m \in \{2, 3, 4, \ldots\}$. Let

$\lambda = (\lambda_1, \ldots, \lambda_m), \mu = (\mu_1, \ldots, \mu_m)$ be elements in $\mathbf{C}^m$, and let $\rho = (\rho_1, \rho_2, \rho_3)$ be an element in $\mathbf{C}^3$ satisfying

$$\lambda_i \neq \lambda_j, \quad \mu_i \neq \mu_j, \quad \rho_i \neq \rho_j$$

for $i \neq j$, and

$$\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i = m\rho_1 + (m-1)\rho_2 + \rho_3.$$

System $(\mathrm{II})_{\lambda,\mu,\rho}$ (or simply (II)) of rank n is the following system of differential equations for the unknown n-column vector y :

$$(xI_n - T_{\mathrm{II}}) \frac{dy}{dx} = A_{\mathrm{II}} y \tag{1.1.1}$$

with

$$T_{\mathrm{II}} = \begin{pmatrix} t_1 I_m & \\ & t_2 I_m \end{pmatrix},$$

$$A_{\mathrm{II}} = \left(\begin{array}{ccc|ccc} \lambda_1 & & & & & \\ & \ddots & & & (\alpha_{ij}) & \\ & & \lambda_m & & & \\ \hline & & & \mu_1 & & \\ & (\beta_{ij}) & & & \ddots & \\ & & & & & \mu_m \end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1) \prod_{\substack{1 \leq k \leq m \\ k \neq i}} \left(\frac{\lambda_k + \mu_j - \rho_1 - \rho_2}{\lambda_i - \lambda_k} \right),$$

$$\beta_{ij} = (\mu_i - \rho_1) \prod_{\substack{1 \leq l \leq m \\ l \neq i}} \left(\frac{\lambda_j + \mu_l - \rho_1 - \rho_2}{\mu_i - \mu_l} \right)$$

for $i, j = 1, \ldots, m$. The Jordan canonical form of the matrix A_{II} is

$$\begin{pmatrix} \rho_1 I_m & & \\ & \rho_2 I_{m-1} & \\ & & \rho_3 \end{pmatrix}.$$

The system (1.1.1) is Fuchsian over $\mathbf{P}^1(\mathbf{C})$ with regular singular points at $x = t_1, t_2, \infty$, and is free from accessory parameters [H2].

We assume

$$\rho_k \notin \mathbf{Z}_{<0}, \qquad \rho_k - \rho_l \notin \mathbf{Z}, \quad \text{for } k, l = 1, 2, 3, \qquad k \neq l, \tag{1.1.2}$$

$$\lambda_i, \mu_i, \lambda_i - \lambda_j, \quad \mu_i - \mu_j \notin \mathbf{Z}, \quad \text{for } i, j = 1, \ldots, m, \qquad i \neq j. \tag{1.1.3}$$

Then by the Frobenius method we obtain the following result.

PROPOSITION 1. *We assume* (1.1.2).

(i) *At* $x = t_1$ *the system* (1.1.1) *has the following* n *linearly independent solutions*: m *singular solutions*

$$y_i(x) = (x - t_1)^{\lambda_i} (e_i + O(x - t_1)), \qquad i = 1, \ldots, m, \tag{1.1.4}$$

which are convergent in B_1, *and* m *holomorphic solutions.*

(ii) *At* $x = t_2$ *the system* (1.1.1) *has the following* n *linearly independent solutions*: m *singular solutions*

$$z_i(x) = (x - t_2)^{\mu_i} (e_{m+i} + O(x - t_2)), \qquad i = 1, \ldots, m, \tag{1.1.5}$$

which are convergent in B_2, *and* m *holomorphic solutions.*

(iii) *At* $x = \infty$ *the system* (1.1.1) *has the following* n *linearly independent solutions*: m *solutions of the form*

$$x^{\rho_1} \sum_{k=0}^{\infty} b_{i,k} x^{-k}, \qquad i = 1, \ldots, m,$$

$(m-1)$ *solutions of the form*

$$x^{\rho_2} \sum_{k=0}^{\infty} b_{m+i,k} x^{-k}, \qquad i = 1, \ldots, m-1,$$

and one solution of the form

$$x^{\rho_3} \sum_{k=0}^{\infty} b_{n,k} x^{-k}.$$

Thus the solutions $y_1, \ldots, y_m$ and $z_1, \ldots, z_m$ given in (1.1.4) and (1.1.5), respectively, are convergent in $W = B_1 \cap B_2$. Set

$$Y_0(x) = (y_1(x) \cdots y_m(x)\, z_1(x) \cdots z_m(x)) \tag{1.1.6}$$

for $x \in W$. Then by the Gauss–Okubo formula ([O, Thm. 2.1, Chap. II]) we obtain the following result.

PROPOSITION 2. *We have*

$$\det Y_0(x) = \prod_{i=1}^{m} \left\{ (x - t_1)^{\lambda_i} (x - t_2)^{\mu_i} \right\} \times \frac{\prod_{i=1}^{m} \Gamma(\lambda_i + 1) \prod_{i=1}^{m} \Gamma(\mu_i + 1)}{\Gamma(\rho_1 + 1)^m \Gamma(\rho_2 + 1)^{m-1} \Gamma(\rho_3 + 1)}.$$

Thus $Y_0(x)$ is a fundamental matrix solution in W of the system (1.1.1).

THEOREM 1. *We assume* (1.1.2) *and* (1.1.3)

$$\begin{aligned} &\lambda_i - \rho_1 \notin \mathbf{Z}, \mu_i - \rho_1 \notin \mathbf{Z} \quad \textit{for } i = 1, \ldots, m, \\ &\lambda_i + \mu_j - (\rho_1 + \rho_2) \notin \mathbf{Z} \quad \textit{for } i, j = 1, \ldots, m. \end{aligned} \tag{1.1.7}$$

There is a diagonal matrix $D \in \mathrm{GL}(n, \mathbf{C})$ *such that the monodromy representation*

$$R_{\mathrm{II}} : \pi_1(\mathbf{P}^1(\mathbf{C}) \setminus \{t_1, t_2, \infty\}, x_0) \to \mathrm{GL}(n, \mathbf{C})$$

of the system (1.1.1) *with respect to* $Y(x) = Y_0(x)D$ *is given by*

$$R_{\mathrm{II}}([\gamma_1]) = \begin{pmatrix} E_m(\lambda) & (\xi_{ij}) \\ O & I_m \end{pmatrix}, \qquad R_{\mathrm{II}}([\gamma_2]) = \begin{pmatrix} I_m & O \\ (\eta_{ij}) & E_m(\mu) \end{pmatrix}, \tag{1.1.8}$$

where

$$E_m(\lambda) = \begin{pmatrix} e(\lambda_1) & & \\ & \ddots & \\ & & e(\lambda_m) \end{pmatrix}, \qquad E_m(\mu) = \begin{pmatrix} e(\mu_1) & & \\ & \ddots & \\ & & e(\mu_m) \end{pmatrix}$$

and

$$\xi_{ij} = (e(\lambda_i - \rho_1) - 1) \prod_{\substack{1 \le k \le m \\ k \ne i}} \left(\frac{e(\mu_j) - e(\rho_1 + \rho_2 - \lambda_k)}{e(\rho_1 + \rho_2 - \lambda_i) - e(\rho_1 + \rho_2 - \lambda_k)} \right),$$

$$\eta_{ij} = (e(\mu_i) - e(\rho_1)) \prod_{\substack{1 \le l \le m \\ l \ne i}} \left(\frac{e(\rho_1 + \rho_2 - \lambda_j) - e(\mu_l)}{e(\mu_i) - e(\mu_l)} \right) \tag{1.1.9}$$

for $i, j = 1, \ldots, m$.

Proof. We define fundamental matrix solutions $Y_1(x)$ and $Y_2(x)$ at $x = t_1$ and $x = t_2$, respectively, as follows. For every $j = 1, \ldots, m$, continue $z_j(x)$ analytically to $x = t_1$ to obtain

$$z_j(x) = u_{1j} y_1(x) + \cdots + u_{mj} y_m(x) + y_j^*(x), \tag{1.1.10}$$

where $u_{jk} \in \mathbf{C}, k = 1, \ldots, m$, and $y_j^*(x)$ is a holomorphic solution of (1.1.1) at $x = t_1$. Similarly we have

$$y_j(x) = v_{1j} z_1(x) + \cdots + v_{mj} z_m(x) + z_j^*(x), \tag{1.1.11}$$

where $v_{jk} \in \mathbf{C}, k = 1, \ldots, m$, and $z_j^*(x)$ is a holomorphic solution of (1.1.1) at $x = t_2$. A fundamental matrix solution $Y_1(x)$ at $x = t_1$ is defined by

$$Y_1(x) = (y_1(x) \cdots y_m(x)\, y_1^*(x) \cdots y_m^*(x)), \tag{1.1.12}$$

and a fundamental matrix solution $Y_2(x)$ at $x = t_2$ is defined by

$$Y_2(x) = (z_1^*(x) \cdots z_m^*(x)\, z_1(x) \cdots z_m(x)). \tag{1.1.13}$$

Then it follows from (1.1.6), (1.1.10), and (1.1.11) that

$$Y_0(x) = Y_1(x) \begin{pmatrix} I_m & U \\ & I_m \end{pmatrix}, \qquad Y_0(x) = Y_2(x) \begin{pmatrix} I_m & \\ V & I_m \end{pmatrix}, \tag{1.1.14}$$

where

$$U = (u_{jk})_{1 \le j,k \le m}, \qquad V = (v_{jk})_{1 \le j,k \le m}.$$

Since $y_j^*(x)$ is holomorphic at $x = t_1$ for $j = 1, \dots, m$, from (1.1.4) we obtain

$$Y_1(x)^{\gamma_1} = Y_1(x) \begin{pmatrix} E_m(\lambda) & \\ & I_m \end{pmatrix}. \tag{1.1.15}$$

Similarly by (1.1.5) we have

$$Y_2(x)^{\gamma_2} = Y_2(x) \begin{pmatrix} I_m & \\ & E_m(\mu) \end{pmatrix}. \tag{1.1.16}$$

Then from (1.1.14) we obtain

$$\begin{aligned} Y_0(x)^{\gamma_1} &= Y_0(x) \begin{pmatrix} E_m(\lambda) & (E_m(\lambda) - I_m) U \\ O & I_m \end{pmatrix} =: Y_0(x) M_1, \\ Y_0(x)^{\gamma_2} &= Y_0(x) \begin{pmatrix} I_m & O \\ (E_m(\mu) - I_m) V & E_m(\mu) \end{pmatrix} =: Y_0(x) M_2. \end{aligned} \tag{1.1.17}$$

Now we have

$$Y_0(x)^{\gamma_1 \cdot \gamma_2} = Y_0(x) M_2 M_1.$$

Since $\gamma_1 \cdot \gamma_2$ is a loop that encircles ∞ once in the negative direction, from Proposition 1(iii) it follows that the matrix $M := M_2 M_1$ is diagonalizable to the diagonal matrix

$$\begin{pmatrix} e(\rho_1) I_m & & \\ & e(\rho_2) I_{m-1} & \\ & & e(\rho_3) \end{pmatrix}.$$

Thus we have

$$\operatorname{rank}(M - e(\rho_1) I_n) = m, \tag{1.1.18}$$

$$\operatorname{rank}(M - e(\rho_1) I_n)(M - e(\rho_2) I_n) = 1. \tag{1.1.19}$$

Set

$$E_1 = e(\rho_1) I_m, \qquad E_2 = e(\rho_2) I_m.$$

Then

$$M - e(\rho_1) I_n = \begin{pmatrix} E_m(\lambda) - E_1 & (E_m(\lambda) - I_m) U \\ (E_m(\mu) - I_m) V E_m(\lambda) & (E_m(\mu) - I_m) V (E_m(\lambda) - I_m) U + E_m(\mu) - E_1 \end{pmatrix}.$$

Noting the assumption (1.1.7), we have $e(\lambda_i) \neq e(\rho_1)$ for every $i = 1, \dots, m$, and hence it follows from (1.1.18) that the last m columns of $M - e(\rho_1) I_n$ are linear combinations of the first m columns of that. Thus we obtain

$$\begin{aligned} &(E_m(\mu) - I_m) V (E_m(\lambda) - I_m) U + E_m(\mu) - E_1 \\ &\quad = (E_m(\mu) - I_m) V E_m(\lambda) (E_m(\lambda) - E_1)^{-1} (E_m(\lambda) - I_m) U. \end{aligned} \tag{1.1.20}$$

Again noting (1.1.7), (1.1.2), and (1.1.3), we have $e(\lambda_i) \neq 1, e(\mu_i) \neq 1$ and $e(\mu_i) \neq e(\rho_1)$ for $i = 1, \ldots, m$, and hence $\det U \neq 0$ and $\det V \neq 0$ by (1.1.20). Then from (1.1.20) we obtain

(1.1.21)

$$U = (E_m(\lambda) - I_m)^{-1} (E_m(\lambda) - E_1) E_1^{-1} V^{-1} (E_m(\mu) - I_m)^{-1} (E_m(\mu) - E_1).$$

On the other hand, by using (1.1.20) we have

$$\begin{aligned} &(M - e(\rho_1) I_n)(M - e(\rho_2) I_n) \\ &= \begin{pmatrix} I_m & \\ & (E_m(\mu) - I_m) V E_m(\lambda) (E_m(\lambda) - E_1)^{-1} \end{pmatrix} \\ &\times \left[\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes F \right] \begin{pmatrix} I_m & \\ & (E_m(\lambda) - E_1)^{-1} (E_m(\lambda) - I_m) U \end{pmatrix}, \end{aligned} \tag{1.1.22}$$

where we have set

(1.1.23)

$$F = (E_m(\lambda) - E_1)(E_m(\lambda) - E_2) + (E_m(\lambda) - I_m) U (E_m(\lambda) - I_m) V E_m(\lambda).$$

Then (1.1.19) holds if and only if

$$\operatorname{rank} F = 1. \tag{1.1.24}$$

Putting (1.1.21) into (1.1.23), we have

$$\begin{aligned} F = &(E_m(\lambda) - I_m)^{-1} (E_m(\lambda) - E_1) E_1^{-1} \\ &\times \left[E_1 (E_m(\lambda) - E_2) E_m(\lambda)^{-1} + V^{-1} (E_m(\mu) - E_1) V \right]. \end{aligned}$$

Thus, setting

$$F_1 = E_1 (E_m(\lambda) - E_2) E_m(\lambda)^{-1} + V^{-1} (E_m(\mu) - E_1) V, \tag{1.1.25}$$

from (1.1.24) we obtain

$$\operatorname{rank} F_1 = 1. \tag{1.1.26}$$

Here we quote two lemmas given in our previous work [H2]. Let $p_1, \ldots, p_m, q_1, \ldots, q_m$ be mutually distinct complex numbers.

LEMMA 1 ([H2, Prop. 3]). *Let*

$$P = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_m \end{pmatrix}, \qquad Q = \begin{pmatrix} q_1 & & \\ & \ddots & \\ & & q_m \end{pmatrix}$$

be diagonal matrices. If $V \in \mathrm{GL}(m, \mathbf{C})$ *satisfies*

$$\operatorname{rank} (V^{-1} Q V - P) = 1,$$

there are diagonal matrices $D_1, D_2 \in \mathrm{GL}(m, \mathbf{C})$ *such that*

$$V = D_1 C^{-1} D_2,$$

where

$$C = \left(\frac{1}{p_i - q_j} \right)_{1 \le i,j \le m}.$$

LEMMA 2 ([H2, Prop. 1]). *The* (i,j)*-entry* γ_{ij} *of the inverse matrix* C^{-1} *of*

$$C = \left(\frac{1}{p_i - q_j} \right)_{1 \le i,j \le m}$$

is

$$\gamma_{ij} = \frac{\prod_{k=1}^{m} (p_j - q_k)}{\prod_{\substack{1 \le l \le m \\ l \ne j}} (p_j - p_l)} \cdot \frac{\prod_{\substack{1 \le l \le m \\ l \ne j}} (q_i - p_l)}{\prod_{\substack{1 \le k \le m \\ k \ne i}} (q_i - q_k)}$$

for $i, j = 1, \dots, m$.

We apply these lemmas to (1.1.26) to obtain

$$V = D_1 (r_{ij})_{1 \le i,j \le m} D_2, \tag{1.1.27}$$

where $D_1, D_2 \in \mathrm{GL}(m, \mathbf{C})$ are diagonal matrices, and

(1.1.28)
$$r_{ij} = \frac{\prod_{k=1}^{m} (e(\mu_k) - e(\rho_1 + \rho_2 - \lambda_j)) \prod_{\substack{1 \le l \le m \\ l \ne j}} (e(\rho_1 + \rho_2 - \lambda_\ell) - e(\mu_i))}{\prod_{\substack{1 \le l \le m \\ l \ne j}} (e(\rho_1 + \rho_2 - \lambda_\ell) - e(\rho_1 + \rho_2 - \lambda_j)) \prod_{\substack{1 \le k \le m \\ k \ne i}} (e(\mu_k) - e(\mu_i))}$$

for $i, j = 1, \dots, m$. U is now determined by (1.1.21).

Set

$$\begin{aligned} f_i &= - \prod_{\substack{1 \le k \le m \\ k \ne i}} (e(\rho_1 + \rho_2 - \lambda_k) - e(\rho_1 + \rho_2 - \lambda_i)), \\ g_i &= \prod_{k=1}^{m} (e(\rho_1 + \rho_2 - \lambda_k) - e(\mu_i)) \cdot \frac{e(\mu_i) - 1}{e(\mu_i) - e(\rho_1)} \end{aligned} \tag{1.1.29}$$

for $i = 1, \dots, m$, and set

$$D = \begin{pmatrix} D_2^{-1} & \\ & D_1 \end{pmatrix} \begin{pmatrix} f_1 & & & & & \\ & \ddots & & & & \\ & & f_m & & & \\ & & & g_1 & & \\ & & & & \ddots & \\ & & & & & g_m \end{pmatrix}. \tag{1.1.30}$$

If we use a fundamental matrix solution $Y_0(x)D$ instead of $Y_0(x)$, the monodromy matrices M_1 and M_2 are replaced by $D^{-1}M_1D$ and $D^{-1}M_2D$, respectively. Thus the theorem follows from (1.1.27)–(1.1.30).

We denote by $\mathcal{M}_{\mathrm{II},\lambda,\mu,\rho}$ the image of the monodromy representation R_{II} given in Theorem 1; thus $\mathcal{M}_{\mathrm{II},\lambda,\mu,\rho}$ is the subgroup of $\mathrm{GL}(n,\mathbf{C})$ generated by $R_{\mathrm{II}}([\gamma_1])$ and $R_{\mathrm{II}}([\gamma_2])$ given by (1.1.8).

Let h be a Hermitian form over $\mathbf{C}^n$. h is said to be *invariant* under a subgroup G of $\mathrm{GL}(n,\mathbf{C})$ if

$$ {}^t\bar{M}HM = H $$

holds for every $M \in G$, where H is the Hermitian matrix associated with h.

From now on we assume that $\lambda_1,\dots,\lambda_m,\mu_1,\dots,\mu_m,\rho_1,\rho_2,\rho_3$ are *real* numbers, and give the Hermitian forms invariant under $\mathcal{M}_{\mathrm{II},\lambda,\mu,\rho}$.

THEOREM 2. *We assume* (1.1.2), (1.1.3), *and* (1.1.7). *Let H be the Hermitian matrix associated with a Hermitian form h invariant under $\mathcal{M}_{\mathrm{II},\lambda,\mu,\rho}$. Then there is a real number α such that*

$$ H = \alpha \left(\begin{array}{c|c} \begin{matrix} h_1 & & \\ & \ddots & \\ & & h_m \end{matrix} & (h_{ij}) \\ \hline {}^t(\bar{h}_{ij}) & \begin{matrix} k_1 & & \\ & \ddots & \\ & & k_m \end{matrix} \end{array} \right), $$

where

$$ h_i = \frac{\sin \pi\lambda_i}{\sin \pi(\lambda_i - \rho_1)} \cdot \prod_{l=1}^{m} \sin \pi(\rho_1 + \rho_2 - \lambda_i - \mu_l) \cdot \prod_{\substack{1 \le k \le m \\ k \ne i}} \sin \pi(\lambda_i - \lambda_k), $$

$$ k_j = \frac{\sin \pi\mu_j}{\sin \pi(\mu_j - \rho_1)} \cdot \prod_{k=1}^{m} \sin \pi(\rho_1 + \rho_2 - \lambda_k - \mu_j) \cdot \prod_{\substack{1 \le l \le m \\ l \ne j}} \sin \pi(\mu_j - \mu_l), \tag{1.1.31} $$

$$ h_{ij} = \frac{\xi_{ij}}{e(\lambda_i) - 1} \cdot h_i $$

for $i,j = 1,\dots,m$, and ξ_{ij} is given by (1.1.9).

Proof. It suffices to determine the entries of a Hermitian matrix H that satisfies

$$ {}^t\bar{M}_1 H M_1 = H \quad \text{and} \quad {}^t\bar{M}_2 H M_2 = H, \tag{1.1.32} $$

where $M_i = R_{\mathrm{II}}([\gamma_i])$ for $i = 1,2$. Set

$$ H = \begin{pmatrix} H_1 & H_2 \\ {}^t\bar{H}_2 & H_3 \end{pmatrix} $$

with $H_1, H_2, H_3 \in \mathrm{M}(m,m;\mathbf{C})$. Using (1.1.8), from (1.1.32) we obtain

$$ E_m(\lambda)^{-1} H_1 E_m(\lambda) = H_1, \tag{1.1.33} $$

$$ E_m(\lambda)^{-1} \{H_1(\xi_{ij}) + H_2\} = H_2, \tag{1.1.34} $$

$$ E_m(\mu)^{-1} H_3 E_m(\mu) = H_3, \tag{1.1.35} $$

$$ \{H_2 + {}^t(\bar{\eta}_{ij}) H_3\} E_m(\mu) = H_2. \tag{1.1.36} $$

By the assumptions (1.1.2) and (1.1.3), from (1.1.33) and (1.1.35) it follows that H_1 and H_3 are diagonal matrices. We set

$$H_1 = \begin{pmatrix} h_1 & & \\ & \ddots & \\ & & h_m \end{pmatrix}, \qquad H_3 = \begin{pmatrix} k_1 & & \\ & \ddots & \\ & & k_m \end{pmatrix},$$

and

$$H_2 = (h_{ij})_{1 \le i,j \le m}.$$

Noting (1.1.2) and (1.1.3), from (1.1.34) we obtain

$$h_{ij} = \frac{\xi_{ij}}{e(\lambda_i) - 1} h_i \tag{1.1.37}$$

for $i, j = 1, \ldots, m$. Then, using (1.1.36) and (1.1.37), we have

$$k_j = \frac{1}{e(\mu_j)\bar{\eta}_{ji}} \cdot \frac{1 - e(\mu_j)}{e(\lambda_i) - 1} \xi_{ij} \cdot h_i \tag{1.1.38}$$

for $i, j = 1, \ldots, m$. Since the left-hand side of (1.1.38) is independent of i, we have

$$\frac{1}{e(\lambda_i) - 1} \cdot \frac{\xi_{ij}}{\bar{\eta}_{ji}} \cdot h_i = \frac{1}{e(\lambda_1) - 1} \cdot \frac{\xi_{1j}}{\bar{\eta}_{j1}} \cdot h_1 \tag{1.1.39}$$

for $i = 2, \ldots, m$. By using (1.1.39), (1.1.38), and (1.1.37), we can express h_i, k_j, and h_{ij} in terms of h_1. Noting the formula

$$e(\omega/2) - e(-\omega/2) = 2\sqrt{-1} \sin \pi\omega$$

for any $\omega \in \mathbf{C}$, and noting that the diagonal entries of H are real numbers, we obtain (1.1.31). It is easy to see that the Hermitian matrix H thus determined satisfies (1.1.32). This completes the proof.

1.2. Monodromy representation of system (III). Let n be an odd integer equal to or greater than 5. We set $n = 2m + 1$ with $m \in \{2, 3, 4, \ldots\}$. Let $\lambda = (\lambda_1, \ldots, \lambda_{m+1}), \mu = (\mu_1, \ldots, \mu_m)$, and $\rho = (\rho_1, \rho_2, \rho_3)$ be elements in $\mathbf{C}^{m+1}, \mathbf{C}^m$, and $\mathbf{C}^3$, respectively, satisfying

$$\lambda_i \neq \lambda_j, \quad \mu_i \neq \mu_j, \quad \rho_i \neq \rho_j$$

for $i \neq j$, and

$$\sum_{i=1}^{m+1} \lambda_i + \sum_{i=1}^{m} \mu_i = m\rho_1 + m\rho_2 + \rho_3.$$

System $(\mathrm{III})_{\lambda,\mu,\rho}$ (or simply (III)) of rank n is the system of differential equations

$$(xI_n - T_{\mathrm{III}}) \frac{dy}{dx} = A_{\mathrm{III}} y \tag{1.2.1}$$

with

$$T_{\mathrm{III}} = \begin{pmatrix} t_1 I_{m+1} & \\ & t_2 I_m \end{pmatrix},$$

$$A_{\mathrm{III}} = \left(\begin{array}{ccc|ccc} \lambda_1 & & & & & \\ & \ddots & & & (\alpha_{ij}) & \\ & & \lambda_{m+1} & & & \\ \hline & & & \mu_1 & & \\ & (\beta_{ij}) & & & \ddots & \\ & & & & & \mu_m \end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1)(\lambda_i - \rho_2) \prod_{\substack{1 \le k \le m+1 \\ k \ne i}} \left(\frac{\lambda_k + \mu_j - \rho_1 - \rho_2}{\lambda_i - \lambda_k}\right), \qquad i = 1, \ldots, m+1, \quad j = 1, \ldots, m,$$

$$\beta_{ij} = \prod_{\substack{1 \le l \le m \\ l \ne i}} \left(\frac{\lambda_j + \mu_l - \rho_1 - \rho_2}{\mu_i - \mu_l}\right), \qquad i = 1, \ldots, m, \quad j = 1, \ldots, m+1.$$

The Jordan canonical form of the matrix A_{III} is

$$\begin{pmatrix} \rho_1 I_m & & \\ & \rho_2 I_m & \\ & & \rho_3 \end{pmatrix}.$$

We assume

$$\begin{array}{llll} \rho_k \notin \mathbf{Z}_{<0}, & \rho_k - \rho_l \notin \mathbf{Z}, & \text{for } k, l = 1, 2, 3, & k \ne l, \\ \lambda_i \notin \mathbf{Z}, & \lambda_i - \lambda_j \notin \mathbf{Z}, & \text{for } i, j = 1, \ldots, m+1, & i \ne j, \\ \mu_i \notin \mathbf{Z}, & \mu_i - \mu_j \notin \mathbf{Z}, & \text{for } i, j = 1, \ldots, m, & i \ne j. \end{array} \tag{1.2.2}$$

Then the system (1.2.1) has $m+1$ singular solutions

$$y_i(x) = (x - t_1)^{\lambda_i} (e_i + O(x - t_1)), \quad i = 1, \ldots, m+1$$

in the domain B_1, and m singular solutions

$$z_i(x) = (x - t_2)^{\mu_i} (e_{m+1+i} + O(x - t_2)), \qquad i = 1, \ldots, m$$

in the domain B_2. Hence we can define a matrix solution

$$Y_0(x) = (y_1(x) \cdots y_{m+1}(x)\, z_1(x) \cdots z_m(x))$$

for $x \in W = B_1 \cap B_2$. We see that $Y_0(x)$ is a fundamental matrix solution (cf. Proposition 2). Then we obtain the following result.

THEOREM 3. *We assume* (1.2.2) *and*

$$\begin{array}{lll} \lambda_i - \rho_k \notin \mathbf{Z} & \text{for } i = 1, \ldots, m+1, & k = 1, 2 \\ \mu_j - \rho_k \notin \mathbf{Z} & \text{for } j = 1, \ldots, m, & k = 1, 2, \\ \lambda_i + \mu_j - (\rho_1 + \rho_2) \notin \mathbf{Z} & \text{for } i = 1, \ldots, m+1, & j = 1, \ldots, m. \end{array} \tag{1.2.3}$$

There is a diagonal matrix $D \in \mathrm{GL}(n, \mathbf{C})$ *such that the monodromy representation*

$$R_{\mathrm{III}} : \pi_1(\mathbf{P}^1(\mathbf{C}) \setminus \{t_1, t_2, \infty\}, x_0) \to \mathrm{GL}(n, \mathbf{C})$$

of the system (1.2.1) *with respect to* $Y(x) = Y_0(x)D$ *is given by*

$$R_{\mathrm{III}}([\gamma_1]) = \begin{pmatrix} E_{m+1}(\lambda) & (\xi_{ij}) \\ O & I_m \end{pmatrix}, \qquad R_{\mathrm{III}}([\gamma_2]) = \begin{pmatrix} I_{m+1} & O \\ (\eta_{ij}) & E_m(\mu) \end{pmatrix},$$

where

$$E_{m+1}(\lambda) = \begin{pmatrix} e(\lambda_1) & & \\ & \ddots & \\ & & e(\lambda_{m+1}) \end{pmatrix}, \qquad E_m(\mu) = \begin{pmatrix} e(\mu_1) & & \\ & \ddots & \\ & & e(\mu_m) \end{pmatrix},$$

$$\text{(1.2.4)} \quad \xi_{ij} = \frac{(e(\lambda_i) - e(\rho_1))(e(\lambda_i) - e(\rho_2))}{e(\rho_1)^m e(\rho_2)^m} \prod_{\substack{1 \le k < m+1 \\ k \ne i}} \left(\frac{e(\lambda_k + \mu_j) - e(\rho_1 + \rho_2)}{e(\lambda_i) - e(\lambda_k)} \right)$$

for $i = 1, \ldots, m+1, j = 1, \ldots, m$, *and*

$$\eta_{ij} = \prod_{\substack{1 \le l \le m \\ l \ne i}} \left(\frac{e(\lambda_j + \mu_l) - e(\rho_1 + \rho_2)}{e(\mu_i) - e(\mu_l)} \right)$$

for $i = 1, \ldots, m, j = 1, \ldots, m+1$.

We denote by $\mathcal{M}_{\mathrm{III},\lambda,\mu,\rho}$ the image of the monodromy representation R_{III} given in Theorem 3; namely,

$$\mathcal{M}_{\mathrm{III},\lambda,\mu,\rho} = \langle R_{\mathrm{III}}([\gamma_1]), R_{\mathrm{III}}([\gamma_2]) \rangle .$$

Now we assume that $\lambda_1, \ldots, \lambda_{m+1}, \mu_1, \ldots, \mu_m, \rho_1, \rho_2, \rho_3$ are *real* numbers and give the Hermitian forms invariant under $\mathcal{M}_{\mathrm{III},\lambda,\mu,\rho}$.

THEOREM 4. *We assume* (1.2.2) *and* (1.2.3). *Let* H *be the Hermitian matrix associated with a Hermitian form* h *invariant under* $\mathcal{M}_{\mathrm{III},\lambda,\mu,\rho}$. *Then there is a real number* α *such that*

$$H = \alpha \left(\begin{array}{ccc|ccc} h_1 & & & & & \\ & \ddots & & & (h_{ij})_{\substack{1 \le i \le m+1 \\ 1 \le i \le m}} & \\ & & h_{m+1} & & & \\ \hline & & & k_1 & & \\ & {}^t(\bar{h}_{ij})_{\substack{1 \le i \le m+1 \\ 1 \le i \le m}} & & & \ddots & \\ & & & & & k_m \end{array} \right),$$

where

$$h_i = \frac{\sin \pi \lambda_i}{\sin \pi (\lambda_i - \rho_1) \cdot \sin \pi (\lambda_i - \rho_2)} \cdot \prod_{\substack{1 \le k \le m+1 \\ k \ne i}} \sin \pi (\lambda_i - \lambda_k)$$

$$\times \prod_{l=1}^{m} \sin \pi (\lambda_i + \mu_l - \rho_1 - \rho_2),$$

$$k_j = 4 \sin \pi \mu_j \cdot \prod_{\substack{1 \le l \le m \\ l \ne j}} \sin \pi (\mu_j - \mu_l) \cdot \prod_{k=1}^{m+1} \sin \pi (\lambda_k + \mu_j - \rho_1 - \rho_2),$$

$$h_{ij} = \frac{\xi_{ij}}{e(\lambda_i) - 1} \cdot h_i$$

for $i = 1, \dots, m+1, j = 1, \dots, m$, *and* ξ_{ij} *is given by* (2.4).

1.3. Monodromy representation of system (IV). Let $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, $\mu = (\mu_1, \mu_2)$ and $\rho = (\rho_1, \rho_2, \rho_3)$ be elements in $\mathbf{C}^4$, $\mathbf{C}^2$, and $\mathbf{C}^3$, respectively, satisfying

$$\lambda_i \neq \lambda_j, \quad \mu_i \neq \mu_j, \quad \rho_i \neq \rho_j$$

for $i \neq j$, and

$$\sum_{i=1}^{4} \lambda_i + \sum_{i=1}^{2} \mu_i = 2\rho_1 + 2\rho_2 + 2\rho_3.$$

System $(\mathrm{IV})_{\lambda,\mu,\rho}$ (or simply (IV)) is the system of differential equations

$$(xI_6 - T_{\mathrm{IV}}) \frac{dy}{dx} = A_{\mathrm{IV}} y \tag{1.3.1}$$

of rank 6 with

$$T_{\mathrm{IV}} = \begin{pmatrix} t_1 I_4 & \\ & t_2 I_2 \end{pmatrix},$$

$$A_{\mathrm{IV}} = \begin{pmatrix} \lambda_1 & & & & \alpha_{11} & \alpha_{12} \\ & \lambda_2 & & & \alpha_{21} & \alpha_{22} \\ & & \lambda_3 & & \alpha_{31} & \alpha_{32} \\ & & & \lambda_4 & \alpha_{41} & \alpha_{42} \\ \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} & \mu_1 & \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} & & \mu_2 \end{pmatrix},$$

where

$$\alpha_{ij} = \frac{\prod_{l=1,2,3} (\lambda_i - \rho_l)}{\prod_{\substack{1 \leq k \leq 4 \\ k \neq i}} (\lambda_i - \lambda_k)} \cdot a_{ij}, \qquad i = 1, \dots, 4, \quad j = 1, 2,$$

$$\beta_{ij} = \frac{1}{\mu_i - \mu_{i'}} \cdot b_{ij}, \qquad i = 1, 2, \quad j = 1, \dots, 4, \quad \{i, i'\} = \{1, 2\},$$

$$a_{11} = \prod_{k=2}^{4} (\lambda_1 + \lambda_k + \mu_2 - \rho_1 - \rho_2 - \rho_3),$$

$$a_{12} = \prod_{k=2}^{4} (\lambda_1 + \lambda_k + \mu_1 - \rho_1 - \rho_2 - \rho_3),$$

$$a_{ij} = \lambda_1 + \lambda_i + \mu_{j'} - \rho_1 - \rho_2 - \rho_3, \qquad i = 2, 3, 4, \quad j = 1, 2, \quad \{j, j'\} = \{1, 2\},$$

$$b_{11} = b_{21} = 1,$$

$$b_{ij} = \prod_{\substack{k=2,3,4 \\ k \neq j}} (\lambda_1 + \lambda_k + \mu_i - \rho_1 - \rho_2 - \rho_3), \qquad i = 1, 2, \quad j = 2, 3, 4.$$

The Jordan canonical form of the matrix A_{IV} is

$$\begin{pmatrix} \rho_1 I_2 & & \\ & \rho_2 I_2 & \\ & & \rho_3 I_2 \end{pmatrix}.$$

We assume

$$\begin{array}{llll} \rho_k \notin \mathbf{Z}_{<0}, & \rho_k - \rho_l \notin \mathbf{Z}, & \text{for } k,l = 1,2,3, & k \neq l \\ \lambda_i \notin \mathbf{Z}, & \lambda_i - \lambda_j \notin \mathbf{Z}, & \text{for } i,j = 1,\dots,4, & i \neq j, \\ \mu_1, \mu_2 \notin \mathbf{Z}, & \mu_1 - \mu_2 \notin \mathbf{Z}. & & \end{array} \tag{1.3.2}$$

Then the system (1.3.1) has 4 singular solutions

$$y_i(x) = (x - t_1)^{\lambda_i} (e_i + O(x - t_1)), \qquad i = 1,\dots,4$$

in the domain B_1, and 2 singular solutions

$$z_i(x) = (x - t_2)^{\mu_i} (e_{4+i} + O(x - t_2)), \qquad i = 1,2$$

in the domain B_2. Hence we can define a matrix solution

$$Y_0(x) = (y_1(x)\, y_2(x)\, y_3(x)\, y_4(x)\, z_1(x)\, z_2(x))$$

for $x \in W = B_1 \cap B_2$. We see that $Y_0(x)$ is a fundamental matrix solution (cf. Proposition 2). Then we obtain the following result.

THEOREM 5. *We assume* (1.3.2) *and*

$$\begin{array}{l} \lambda_i - \rho_k \notin \mathbf{Z} \quad \text{for } i = 1,\dots,4, \qquad k = 1,2,3, \\ \lambda_i + \lambda_j + \mu_k - (\rho_1 + \rho_2 + \rho_3) \notin \mathbf{Z} \quad \text{for } i,j = 1,\dots,4, \quad i \neq j, \quad k = 1,2. \end{array} \tag{1.3.3}$$

There is a diagonal matrix $D \in \mathrm{GL}(6,\mathbf{C})$ *such that the monodromy representation*

$$R_{\mathrm{IV}} : \pi_1(\mathbf{P}^1(\mathbf{C}) \setminus \{t_1, t_2, \infty\}, x_0) \to \mathrm{GL}(6,\mathbf{C})$$

of the system (2.2.1) *with respect to* $Y(x) = Y_0(x)D$ *is given by*

$$R_{\mathrm{IV}}([\gamma_1]) = \begin{pmatrix} E_4(\lambda) & (\xi_{ij}) \\ O & I_2 \end{pmatrix}, \qquad R_{\mathrm{IV}}([\gamma_2]) = \begin{pmatrix} I_4 & O \\ (\eta_{ij}) & E_2(\mu) \end{pmatrix},$$

where

$$E_4(\lambda) = \begin{pmatrix} e(\lambda_1) & & & \\ & e(\lambda_2) & & \\ & & e(\lambda_3) & \\ & & & e(\lambda_4) \end{pmatrix}, \qquad E_2(\mu) = \begin{pmatrix} e(\mu_1) & \\ & e(\mu_2) \end{pmatrix},$$

$$\xi_{ij} = \frac{\prod_{l=1}^{3} (e(\lambda_i) - e(\rho_l))}{e(\rho_1 + \rho_2 + \rho_3)^2\, e(\lambda_1) \prod_{\substack{1 \le k \le 4 \\ k \neq i}} (e(\lambda_i) - e(\lambda_k))} \cdot x_{ij}, \tag{1.3.4}$$

$$i = 1,\dots 4, \qquad j = 1,2,$$

$$\begin{aligned} x_{11} &= [12:2]\,[13:2]\,[14:2], \\ x_{12} &= e(\mu_2 - \mu_1)\,[12:1]\,[13:1]\,[14:1], \\ x_{21} &= [12:2], \qquad x_{22} = [12:1], \\ x_{31} &= [13:2], \qquad x_{32} = [13:1], \\ x_{41} &= [14:2], \qquad x_{42} = [14:1], \end{aligned}$$

$$\eta_{ij} = \frac{1}{e(\mu_i) - e(\mu_{i'})} \cdot y_{ij}, \qquad i = 1,2, \quad j = 1,\dots,4, \quad \{i,i'\} = \{1,2\},$$

$$\begin{aligned} y_{11} &= e(\mu_1 - \mu_2), & y_{21} &= 1, \\ y_{12} &= [13:1][14:1], & y_{22} &= [13:2][14:2], \\ y_{13} &= [12:1][14:1], & y_{23} &= [12:2][14:2], \\ y_{14} &= [12:1][13:1], & y_{24} &= [12:2][13:2]. \end{aligned}$$

Here we have set

$$[ij:k] = e(\lambda_i + \lambda_j + \mu_k) - e(\rho_1 + \rho_2 + \rho_3)$$

for $i,j = 1,\dots,4, k = 1,2$.

We denote by $\mathcal{M}_{\mathrm{IV},\lambda,\mu,\rho}$ the image of the monodromy representation R_{IV} given in Theorem 5; namely,

$$\mathcal{M}_{\mathrm{IV},\lambda,\mu,\rho} = \langle R_{\mathrm{IV}}([\gamma_1]), R_{\mathrm{IV}}([\gamma_2]) \rangle .$$

Now we assume that $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \mu_1, \mu_2, \rho_1, \rho_2, \rho_3$ are *real* numbers and give the Hermitian forms invariant under $\mathcal{M}_{\mathrm{IV},\lambda,\mu,\rho}$.

THEOREM 6. *We assume* (1.3.2) *and* (1.3.3). *Let* H *be the Hermitian matrix associated with a Hermitian form* h *invariant under* $\mathcal{M}_{\mathrm{IV},\lambda,\mu,\rho}$. *Then there is a real number* α *such that*

$$H = \alpha \begin{pmatrix} h_1 & & & & h_{11} & h_{12} \\ & h_2 & & & h_{21} & h_{22} \\ & & h_3 & & h_{31} & h_{32} \\ & & & h_4 & h_{41} & h_{42} \\ \bar{h}_{11} & \bar{h}_{21} & \bar{h}_{31} & \bar{h}_{41} & k_1 & \\ \bar{h}_{12} & \bar{h}_{22} & \bar{h}_{32} & \bar{h}_{42} & & k_2 \end{pmatrix},$$

where

$$h_1 = \frac{\sin \pi\lambda_1 \cdot \prod_{k=2}^{4} \sin \pi(\lambda_1 - \lambda_k)}{\prod_{l=1}^{3} \sin \pi(\lambda_1 - \rho_l)},$$

$$h_i = 2^4 \cdot \frac{\sin \pi\lambda_1 \cdot \prod_{\substack{1 \le k \le 4 \\ k \ne i}} \sin \pi(\lambda_i - \lambda_k)}{\prod_{l=1}^{3} \sin \pi(\lambda_i - \rho_l)} \times \prod_{\substack{2 \le k \le 4 \\ k \ne i}} \prod_{l=1,2} \sin \pi(\lambda_1 + \lambda_k + \mu_l - \rho_1 - \rho_2 - \rho_3), \qquad i = 2,3,4,$$

$$k_j = 2^4 \sin \pi\mu_j \cdot \sin \pi(\mu_j - \mu_{j'}) \cdot \prod_{k=2}^{4} \sin \pi(\lambda_1 + \lambda_k + \mu_{j'} - \rho_1 - \rho_2 - \rho_3), \qquad j = 1,2, \quad \{j,j'\} = \{1,2\},$$

$$h_{ij} = \frac{\xi_{ij}}{e(\lambda_i) - 1} \cdot h_i, \qquad i = 1,\dots,4, \quad j = 1,2,$$

and ξ_{ij} *is given by* (1.3.4).

2. System (II*), system (III*), and system (IV*). Let $t_1, t_2, t_3 \in \mathbf{C}$ be three points satisfying the following:

(i) t_1, t_2, t_3 are mutually distinct,
(ii) t_1, t_2, t_3 do not lie on a line,
(iii) when we trace, in the positive direction, the circle C on which t_1, t_2, t_3 lie, t_1, t_2, t_3 come in this order. We set

$$\begin{aligned} B_1 &= B\left(t_1, \min\left\{|t_2 - t_1|, |t_3 - t_1|\right\}\right), \\ B_2 &= B\left(t_2, \min\left\{|t_1 - t_2|, |t_3 - t_2|\right\}\right), \\ B_3 &= B\left(t_3, \min\left\{|t_1 - t_3|, |t_2 - t_3|\right\}\right), \\ W &= B_1 \cap B_2 \cap B_3. \end{aligned}$$

Then, by the assumptions on the t_j's, we see that W is a nonempty, simply connected domain. Take any x_0 in W and fix it. For $i = 1, 2, 3$, let γ_i be a loop that starts from and ends at x_0, encircles t_i once in the positive direction, and crosses C only twice on the arc $t_{i-1}t_it_{i+1}$, where we set $t_{-1} = t_3$ and $t_4 = t_1$. Then the fundamental group $\pi_1(\mathbf{P}^1(\mathbf{C}) \setminus \{t_1, t_2, t_3, \infty\}, x_0)$ is generated by the homotopy classes $[\gamma_1]$, $[\gamma_2]$, and $[\gamma_3]$.

This notation is fixed throughout this section.

2.1. Monodromy representation of system (II*). Let n be an even integer equal to or greater than 4. We set $n = 2m$ with $m \in \{2, 3, 4, \ldots\}$. Let $\lambda = (\lambda_1, \ldots, \lambda_m), \mu = (\mu_1, \ldots, \mu_{m-1}), \nu$ and $\rho = (\rho_1, \rho_2)$ be elements in $\mathbf{C}^m$, $\mathbf{C}^{m-1}$, $\mathbf{C}$, and $\mathbf{C}^2$, respectively, satisfying

$$\lambda_i \neq \lambda_j, \quad \mu_i \neq \mu_j, \quad \rho_i \neq \rho_j$$

for $i \neq j$, and

$$\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m-1} \mu_i + \nu = m\rho_1 + m\rho_2.$$

System $(\mathrm{II}^*)_{\lambda,\mu,\nu,\rho}$ (or simply (II*)) of rank n is the system of differential equations

$$(xI_n - T_{\mathrm{II}^*})\frac{dy}{dx} = A_{\mathrm{II}^*}y \tag{2.1.1}$$

with

$$T_{\mathrm{II}^*} = \begin{pmatrix} t_1 I_m & & \\ & t_2 I_{m-1} & \\ & & t_3 \end{pmatrix},$$

$$A_{\mathrm{II}^*} = \left(\begin{array}{ccc|ccc} \lambda_1 & & & & (\alpha_{ij}) & \\ & \ddots & & & & \\ & & \lambda_{m+1} & & & \\ \hline & (\beta_{ij}) & & \mu_1 & & \xi_1 \\ & & & & \ddots & \vdots \\ & & & & & \mu_{m-1} \quad \xi_{m-1} \\ & & & \eta_1 & \cdots & \eta_{m-1} \quad \nu \end{array}\right),$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1)(\lambda_i - \rho_2) \cdot \prod_{\substack{1 \le k \le m \\ k \ne i}} \left(\frac{\lambda_k + \mu_j - \rho_1 - \rho_2}{\lambda_i - \lambda_k} \right),$$

$$i = 1, \dots, m, \qquad j = 1, \dots, m-1,$$

$$\alpha_{im} = (\lambda_i - \rho_1)(\lambda_i - \rho_2) \cdot \prod_{\substack{1 \le k \le m \\ k \ne i}} \frac{1}{\lambda_i - \lambda_k}, \qquad i = 1, \dots, m,$$

$$\beta_{ij} = \prod_{\substack{1 \le l \le m-1 \\ l \ne i}} \left(\frac{\lambda_j + \mu_l - \rho_1 - \rho_2}{\mu_i - \mu_l} \right), \qquad i = 1, \dots, m-1, \quad j = 1, \dots, m,$$

$$\beta_{mj} = -\prod_{l=1}^{m-1} (\lambda_j + \mu_l - \rho_1 - \rho_2), \qquad j = 1, \dots, m,$$

$$\xi_i = \prod_{\substack{1 \le l \le m-1 \\ l \ne i}} \frac{1}{\mu_i - \mu_l}, \qquad i = 1, \dots, m-1,$$

$$\eta_j = -\prod_{k=1}^{m} (\lambda_k + \mu_j - \rho_1 - \rho_2), \qquad j = 1, \dots, m-1.$$

The Jordan canonical form of A_{II^*} is

$$\begin{pmatrix} \rho_1 I_m & \\ & \rho_2 I_m \end{pmatrix}.$$

The system (2.1.1) is Fuchsian over $\mathbf{P}^1(\mathbf{C})$ with regular singular points at $x = t_1, t_2, t_3, \infty$, and is free from accessory parameters.

We assume

$$\begin{array}{llll} \rho_1, \rho_2 \notin \mathbf{Z}_{<0}, & \rho_1 - \rho_2 \notin \mathbf{Z}, & & \\ \lambda_i \notin \mathbf{Z}, & \lambda_i - \lambda_j \notin \mathbf{Z}, & \text{for } i, j = 1, \dots, m, & i \ne j, \\ \mu_i \notin \mathbf{Z}, & \mu_i - \mu_j \notin \mathbf{Z}, & \text{for } i, j = 1, \dots, m-1, & i \ne j. \\ \nu \notin \mathbf{Z}. & & & \end{array} \tag{2.1.2}$$

Then the system (2.1.1) has m singular solutions

$$y_i(x) = (x - t_1)^{\lambda_i} (e_i + O(x - t_1)), \qquad i = 1, \dots m$$

in the domain B_1, $m-1$ singular solutions

$$z_i(x) = (x - t_2)^{\mu_i} (e_{m+i} + O(x - t_2)), \qquad i = 1, \dots, m-1$$

in the domain B_2 and one singular solution

$$w(x) = (x - t_3)^{\nu} (e_n + O(x - t_3))$$

in the domain B_3. Hence we can define a matrix solution

$$Y_0(x) = (y_1(x) \cdots y_m(x) z_1(x) \cdots z_{m-1}(x) w(x))$$

for $x \in W = B_1 \cap B_2 \cap B_3$. We see that $Y_0(x)$ is a fundamental matrix solution (cf. Proposition 2).

THEOREM 7. *We assume* (2.1.2) *and*

$$(2.1.3)\qquad \begin{array}{llll} \lambda_i - \rho_k \notin \mathbf{Z} & \text{for } i = 1, \dots, m, & k = 1, 2, \\ \lambda_i + \mu_j - (\rho_1 + \rho_2) \notin \mathbf{Z} & \text{for } i = 1, \dots, m, & j = 1, \dots, m-1 \end{array}$$

There is a diagonal matrix $D \in \mathrm{GL}(n, \mathbf{C})$ *such that the monodromy representation*

$$R_{\mathrm{II}^*} : \pi_1 \left(\mathbf{P}^1 (\mathbf{C}) \setminus \{t_1, t_2, t_3, \infty\}, x_0\right) \to \mathrm{GL}\,(n, \mathbf{C})$$

of the system (2.1.1) *with respect to* $Y(x) = Y_0(x)D$ *is given by*

$$R_{\mathrm{II}^*}([\gamma_1]) = \begin{pmatrix} E_m(\lambda) & (\xi_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le m}} \\ O & I_m \end{pmatrix},$$

$$R_{\mathrm{II}^*}([\gamma_2]) = \begin{pmatrix} I_m & O & O \\ (\eta_{ij})_{\substack{1 \le i \le m-1 \\ 1 \le j \le m}} & E_{m-1}(\mu) & (\eta_{in})_{1 \le i \le m-1} \\ O & O & 1 \end{pmatrix},$$

$$R_{\mathrm{II}^*}([\gamma_3]) = \begin{pmatrix} I_{n-1} & O \\ (\zeta_j)_{1 \le j \le n-1} & e(\nu) \end{pmatrix},$$

where

$$E_m(\lambda) = \begin{pmatrix} e(\lambda_1) & & \\ & \ddots & \\ & & e(\lambda_m) \end{pmatrix}, \qquad E_{m-1}(\mu) = \begin{pmatrix} e(\mu_1) & & \\ & \ddots & \\ & & e(\mu_{m-1}) \end{pmatrix},$$

(2.1.4)

$$\xi_{ij} = (e(\lambda_i) - e(\rho_1))(e(\rho_2 - \lambda_i) - 1) \prod_{\substack{1 \le k \le m \\ k \ne i}} \left(\frac{e(\mu_j) - e(\rho_1 + \rho_2 - \lambda_k)}{e(\rho_1 + \rho_2 - \lambda_i) - e(\rho_1 + \rho_2 - \lambda_k)} \right), \qquad i = 1, \dots, m, \quad j = 1, \dots, m-1,$$

$$\xi_{im} = (e(\lambda_i) - e(\rho_1))(e(\rho_2 - \lambda_i) - 1) \prod_{\substack{1 \le k \le m \\ k \ne i}} \frac{1}{e(\rho_1 + \rho_2 - \lambda_i) - e(\rho_1 + \rho_2 - \lambda_k)}, \qquad i = 1, \dots, m,$$

(2.1.5)

$$\eta_{ij} = \prod_{\substack{1 \le l \le m-1 \\ l \ne i}} \left(\frac{e(\rho_1 + \rho_2 - \lambda_j) - e(\mu_l)}{e(\mu_i) - e(\mu_l)} \right), \qquad i = 1, \dots, m-1, \quad j = 1, \dots, m,$$

$$\eta_{in} = \prod_{\substack{1 \le l \le m-1 \\ l \ne i}} \frac{1}{e(\mu_i) - e(\mu_l)}, \qquad i = 1, \dots, m-1,$$

and

$$\zeta_j = e(\lambda_j + \nu - \rho_1 - \rho_2) \prod_{l=1}^{m-1} (e(\rho_1 + \rho_2 - \lambda_j) - e(\mu_l)), \qquad j = 1, \dots, m,$$

$$\zeta_{m+j} = -\frac{\prod_{k=1}^{m} (e(\mu_j) - e(\rho_1 + \rho_2 - \lambda_k))}{e(\mu_j)}, \qquad j = 1, \dots, m-1.$$

We denote by $\mathcal{M}_{\mathrm{II}^*,\lambda,\mu,\nu,\rho}$ the image of the monodromy representation R_{II^*} given in Theorem 7; namely,

$$\mathcal{M}_{\mathrm{II}^*,\lambda,\mu,\nu,\beta} = \langle R_{\mathrm{II}^*}([\gamma_1]), R_{\mathrm{II}^*}([\gamma_2]), R_{\mathrm{II}^*}([\gamma_3]) \rangle .$$

Now we assume that $\lambda_1, \dots, \lambda_m, \mu_1, \dots, \mu_{m-1}, \nu, \rho_1, \rho_2$ are *real* numbers and give the Hermitian forms invariant under $\mathcal{M}_{\mathrm{II}^*,\lambda,\mu,\nu,\rho}$.

THEOREM 8. *We assume* (2.1.2) *and* (2.1.3). *Let H be the Hermitian matrix associated with a Hermitian form h invariant under $\mathcal{M}_{\mathrm{II}^*,\lambda,\mu,\nu,\rho}$. Then there is a real number α such that*

$$H = \alpha \left(\begin{array}{ccc|ccc} h_1 & & & & & \\ & \ddots & & & (h_{ij})_{\substack{1\le i\le m\\ 1\le j\le m}} & \\ & & h_m & & & \\ \hline & & & k_1 & & f_1 \\ & {}^t(\bar{h}_{ij})_{\substack{1\le i\le m\\ 1\le j\le m}} & & & \ddots & \vdots \\ & & & & k_{m-1} & f_{m-1} \\ & & & \bar{f}_1 & \cdots \quad \bar{f}_{m-1} & g \end{array} \right),$$

where

$$h_i = \frac{\sin \pi\lambda_i}{\sin \pi(\lambda_i - \rho_1) \cdot \sin \pi(\lambda_i - \rho_2)} \cdot \prod_{\substack{1\le k\le m\\ k\ne i}} \sin \pi(\lambda_i - \lambda_k) \times \prod_{l=1}^{m-1} \sin \pi(\lambda_i + \mu_l - \rho_1 - \rho_2), \qquad i = 1, \dots, m,$$

$$k_j = 4 \sin \pi\mu_j \cdot \prod_{\substack{1\le l\le m-1\\ l\ne j}} \sin \pi(\mu_j - \mu_l) \cdot \prod_{k=1}^{m} \sin \pi(\lambda_k + \mu_j - \rho_1 - \rho_2), \qquad j = 1, \dots, m-1,$$

$$g = -\frac{\sin \pi\nu}{4^{m-2}},$$

$$h_{ij} = \frac{\xi_{ij}}{e(\lambda_i) - 1} \cdot h_i, \qquad i, j = 1, \dots, m,$$

$$f_i = \frac{\eta_{in}}{e(\mu_i) - 1} \cdot k_i, \qquad i = 1, \dots, m-1,$$

and ξ_{ij}, η_{in} are given by (2.1.4), (2.1.5), *respectively.*

2.2. Monodromy representation of system (III*). Let n be an odd integer equal to or greater than 5. We set $n = 2m+1$ with $m \in \{2,3,4,\dots\}$. Let $\lambda = (\lambda_1,\dots,\lambda_m)$, $\mu = (\mu_1,\dots,\mu_m)$, ν and $\rho = (\rho_1,\rho_2)$ be elements in $\mathbf{C}^m$, $\mathbf{C}^m$, $\mathbf{C}$, and $\mathbf{C}^2$, respectively, satisfying

$$\lambda_i \neq \lambda_j, \quad \mu_i \neq \mu_j, \quad \rho_i \neq \rho_j$$

for $i \neq j$, and

$$\sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \mu_i + \nu = (m+1)\rho_1 + m\rho_2.$$

System $(\mathrm{III}^*)_{\lambda,\mu,\nu,\rho}$ (or simply (III*)) of rank n is the system of differential equations

$$(xI_n - T_{\mathrm{III}^*})\frac{dy}{dx} = A_{\mathrm{III}^*} y \tag{2.2.1}$$

with

$$T_{\mathrm{III}^*} = \begin{pmatrix} t_1 I_m & & \\ & t_2 & \\ & & t_3 I_m \end{pmatrix},$$

$$A_{\mathrm{III}^*} = \begin{pmatrix} \lambda_1 & & & \gamma_1 & & & \\ & \ddots & & \vdots & & (\alpha_{ij}) & \\ & & \lambda_m & \gamma_m & & & \\ \sigma_1 & \cdots & \sigma_m & \nu & \tau_1 & \cdots & \tau_m \\ & & & \theta_1 & \mu_1 & & \\ & (\beta_{ij}) & & \vdots & & \ddots & \\ & & & \theta_m & & & \mu_{m-1} \end{pmatrix},$$

where

$$\alpha_{ij} = (\lambda_i - \rho_1) \prod_{\substack{1\le k\le m \\ k\neq i}} \left(\frac{\rho_1+\rho_2-\lambda_k-\mu_j}{\lambda_i-\lambda_k}\right), \qquad i,j = 1,\dots,m,$$

$$\beta_{ij} = (\mu_i - \rho_1) \prod_{\substack{1\le l\le m \\ l\neq i}} \left(\frac{\rho_1+\rho_2-\lambda_j-\mu_l}{\mu_i-\mu_l}\right), \qquad i,j = 1,\dots,m,$$

$$\gamma_i = \frac{\lambda_i - \rho_1}{\prod_{\substack{1\le k\le m \\ k\neq i}} (\lambda_k - \lambda_i)}, \qquad i = 1,\dots,m,$$

$$\theta_i = \frac{\mu_i - \rho_1}{\prod_{\substack{1\le l\le m \\ l\neq i}} (\mu_i - \mu_l)}, \qquad i = 1,\dots,m,$$

$$\sigma_j = \prod_{l=1}^{m} (\rho_1+\rho_2-\lambda_j-\mu_l), \qquad j = 1,\dots,m,$$

$$\tau_j = -\prod_{k=1}^{m} (\lambda_k+\mu_j-\rho_1-\rho_2), \qquad j = 1,\dots,m.$$

The Jordan canonical form of A_{III^*} is

$$\begin{pmatrix} \rho_1 I_{m+1} & \\ & \rho_2 I_m \end{pmatrix}.$$

The system (2.2.1) is Fuchsian over $\mathbf{P}^1(\mathbf{C})$ with regular singular points at $x = t_1, t_2, t_3, \infty$, and is free from accessory parameters.

We assume

$$
\begin{array}{llll}
\rho_1, \rho_2 \notin \mathbf{Z}_{<0}, & \rho_1 - \rho_2 \notin \mathbf{Z}, & & \\
\lambda_i \notin \mathbf{Z}, & \lambda_i - \lambda_j \notin \mathbf{Z}, & \text{for } i,j = 1,\dots,m, & i \neq j, \\
\mu_i \notin \mathbf{Z}, & \mu_i - \mu_j \notin \mathbf{Z}, & \text{for } i,j = 1,\dots,m, & i \neq j, \\
\nu \notin \mathbf{Z}. & & &
\end{array} \tag{2.2.2}
$$

Then the system (2.2.1) has m singular solutions

$$y_i(x) = (x - t_1)^{\lambda_i} (e_i + O(x - t_1)), \qquad i = 1, \dots, m$$

in the domain B_1, one singular solution

$$w(x) = (x - t_2)^{\nu} (e_{m+1} + O(x - t_2))$$

in the domain B_2 and m singular solutions

$$z_i(x) = (x - t_3)^{\mu_i} (e_{m+1+i} + O(x - t_3)), \qquad i = 1, \dots, m$$

in the domain B_3. Hence we can define a matrix solution

$$Y_0(x) = (y_1(x) \cdots y_m(x)\, w(x)\, z_1(x) \cdots z_m(x))$$

for $x \in W = B_1 \cap B_2 \cap B_3$. We see that $Y_0(x)$ is a fundamental matrix solution (cf. Proposition 2).

THEOREM 9. *We assume* (2.2.2) *and*

$$
\begin{array}{ll}
\lambda_i - \rho_1 \notin \mathbf{Z} & \text{for } i = 1,\dots,m, \\
\mu_i - \rho_1 \notin \mathbf{Z} & \text{for } i = 1,\dots,m, \\
\lambda_i + \mu_j - (\rho_1 + \rho_2) \notin \mathbf{Z} & \text{for } i,j = 1,\dots,m.
\end{array} \tag{2.2.3}
$$

There is a diagonal matrix $D \in \mathrm{GL}(n, \mathbf{C})$ *such that the monodromy representation*

$$R_{\mathrm{III}^*} : \pi_1(\mathbf{P}^1(\mathbf{C}) \setminus \{t_1, t_2, t_3, \infty\}, x_0) \to \mathrm{GL}(n, \mathbf{C})$$

of the system (2.2.1) *with respect to* $Y(x) = Y_0(x)D$ *is given by*

$$R_{\mathrm{III}^*}([\gamma_1]) = \begin{pmatrix} E_m(\lambda) & (\xi_{ij})_{\substack{1 \le i \le m \\ 0 \le j \le m}} \\ O & I_{m+1} \end{pmatrix},$$

$$R_{\mathrm{III}^*}([\gamma_2]) = \begin{pmatrix} I_m & O & O \\ (\zeta_j)_{1 \le j \le m} & e(\nu) & (\zeta_{m+j})_{1 \le j \le m} \\ O & O & I_m \end{pmatrix},$$

$$R_{\mathrm{III}^*}([\gamma_3]) = \begin{pmatrix} I_{m+1} & O \\ (\eta_{ij})_{\substack{1 \le i \le m \\ 1 \le j \ge m+1}} & E_m(\mu) \end{pmatrix},$$

where

$$E_m(\lambda) = \begin{pmatrix} e(\lambda_1) & & \\ & \ddots & \\ & & e(\lambda_m) \end{pmatrix}, \qquad E_m(\mu) = \begin{pmatrix} e(\mu_1) & & \\ & \ddots & \\ & & e(\mu_m) \end{pmatrix}, \tag{2.2.4}$$

$$\xi_{i0} = -e(\lambda_i + \nu - \rho_1 - \rho_2)(e(\lambda_i) - e(\rho_1)) \times \prod_{\substack{1 \le k \le m \\ k \ne i}} \frac{1}{e(\rho_2 - \lambda_i) - e(\rho_2 - \lambda_k)}, \qquad i = 1, \dots, m,$$

$$\xi_{ij} = (e(\lambda_i) - e(\rho_1)) \prod_{\substack{1 \le k \le m \\ k \ne i}} \left(\frac{e(\mu_j - \rho_1) - e(\rho_2 - \lambda_k)}{e(\rho_2 - \lambda_i) - e(\rho_2 - \lambda_k)} \right), \qquad i, j = 1, \dots, m,$$

$$\zeta_j = -\prod_{l=1}^{m} (e(\rho_2 - \lambda_j) - e(\mu_l - \rho_1)), \qquad j = 1, \dots, m,$$

$$\zeta_{m+j} = e(\rho_1) \prod_{k=1}^{m} (e(\mu_j - \rho_1) - e(\rho_2 - \lambda_k)), \qquad j = 1, \dots, m,$$

and

$$\eta_{ij} = (e(\mu_i - \rho_1) - 1) \prod_{\substack{1 \le l \le m \\ l \ne i}} \left(\frac{e(\rho_2 - \lambda_j) - e(\mu_l - \rho_1)}{e(\mu_i - \rho_1) - e(\mu_l - \rho_1)} \right), \qquad i, j = 1, \dots, m,$$

$$\eta_{i\, m+1} = (e(\rho_1 - \mu_i) - 1) \prod_{\substack{1 \le l \le m \\ l \ne i}} \frac{1}{e(\mu_i - \rho_1) - e(\mu_l - \rho_1)}, \qquad i = 1, \dots, m. \tag{2.2.5}$$

We denote by $\mathcal{M}_{\mathrm{III}^*,\lambda,\mu,\nu,\rho}$ the image of the monodromy representation R_{III^*} given in Theorem 9; namely,

$$\mathcal{M}_{\mathrm{III}^*,\lambda,\mu,\nu,\rho} = \langle R_{\mathrm{III}^*}([\gamma_1]), R_{\mathrm{III}^*}([\gamma_2]), R_{\mathrm{III}^*}([\gamma_3]) \rangle.$$

Now we assume that $\lambda_1, \dots, \lambda_m, \mu_1, \dots, \mu_m, \nu, \rho_1, \rho_2$ are *real* numbers, and give the Hermitian forms invariant under $\mathcal{M}_{\mathrm{III}^*,\lambda,\mu,\nu,\rho}$.

THEOREM 10. *We assume* (2.2.2) *and* (2.2.3). *Let H be the Hermitian matrix associated with a Hermitian form h invariant under $\mathcal{M}_{\mathrm{III}^*,\lambda,\mu,\nu,\rho}$. Then there is a real number α such that*

$$H = \alpha \begin{pmatrix} h_1 & & & h_{10} & & & \\ & \ddots & & \vdots & & (h_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le m}} & \\ & & h_m & h_{m0} & & & \\ \bar{h}_{10} & \cdots & \bar{h}_{m0} & g & h_{m+11} & \cdots & h_{m+1m} \\ & & & \bar{h}_{m+11} & k_1 & & \\ & {}^t(\bar{h}_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le m}} & & \vdots & & \ddots & \\ & & & \bar{h}_{m+1m} & & & k_m \end{pmatrix},$$

where

$$h_i = \frac{\sin \pi \lambda_i}{\sin \pi (\lambda_i - \rho_1)} \cdot \prod_{\substack{1 \le k \le m \\ k \ne i}} \sin \pi (\lambda_i - \lambda_k)$$
$$\times \prod_{l=1}^{m} \sin \pi (\lambda_i + \mu_l - \rho_1 - \rho_2), \qquad i = 1, \dots, m,$$
$$g = -\frac{\sin \pi \nu}{4^{m-1}},$$
$$k_j = \frac{\sin \pi \mu_j}{\sin \pi (\mu_j - \rho_1)} \cdot \prod_{\substack{1 \le l \le m \\ l \ne j}} \sin \pi (\mu_j - \mu_l)$$
$$\times \prod_{k=1}^{m} \sin \pi (\lambda_k + \mu_j - \rho_1 - \rho_2), \qquad j = 1, \dots, m,$$
$$h_{ij} = \frac{\xi_{ij}}{e(\lambda_i) - 1} \cdot h_i, \qquad i = 1, \dots, m, \quad j = 0, \dots, m,$$
$$h_{m+1j} = \frac{e(\mu_j)\, \bar{\eta}_{jm+1}}{1 - e(\mu_j)} \cdot k_j, \quad j = 1, \dots, m,$$

and ξ_{ij}, η_{ji} *are given by* (2.2.4) *and* (2.2.5), *respectively.*

2.3. Monodromy representation of system (IV*). Let $\lambda = (\lambda_1, \lambda_2), \mu = (\mu_1, \mu_2), \nu = (\nu_1, \nu_2)$ and $\rho = (\rho_1, \rho_2)$ be elements in $\mathbf{C}^2$ satisfying

$$\lambda_1 \ne \lambda_2, \quad \mu_1 \ne \mu_2, \quad \nu_1 \ne \nu_2, \quad \rho_1 \ne \rho_2,$$

and

$$\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + \nu_1 + \nu_2 = 4\rho_1 + 2\rho_2.$$

System $(\mathrm{IV}^*)_{\lambda,\mu,\nu,\rho}$ (or simply (IV*)) is the system of differential equations of rank 6

$$(xI_6 - T_{\mathrm{IV}^*}) \frac{dy}{dx} = A_{\mathrm{IV}^*} y \tag{2.3.1}$$

with

$$T_{\mathrm{IV}^*} = \begin{pmatrix} t_1 I_2 & & \\ & t_2 I_2 & \\ & & t_3 I_2 \end{pmatrix},$$
$$A_{\mathrm{IV}^*} = \begin{pmatrix} \lambda_1 & & \alpha_{13} & \alpha_{14} & \alpha_{15} & \alpha_{16} \\ & \lambda_2 & \alpha_{23} & \alpha_{24} & \alpha_{25} & \alpha_{26} \\ \beta_{11} & \beta_{12} & \mu_1 & & \beta_{15} & \beta_{16} \\ \beta_{21} & \beta_{22} & & \mu_2 & \beta_{25} & \beta_{26} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \nu_1 & \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & & \nu_2 \end{pmatrix},$$

where

$$\alpha_{ij} = \frac{\lambda_i - \rho_1}{\lambda_i - \lambda_{i'}} \cdot \alpha_{ij}, \quad \text{for } i = 1, 2, \quad \text{with } \{i, i'\} = \{1, 2\}, \quad j = 3, 4, 5, 6,$$
$$\beta_{ij} = \frac{\mu_i - \rho_1}{\mu_i - \mu_{i'}} \cdot b_{ij}, \quad \text{for } i = 1, 2, \quad \text{with } \{i, i'\} = \{1, 2\}, \quad j = 1, 2, 5, 6,$$
$$\gamma_{ij} = \frac{\nu_i - \rho_1}{\nu_i - \nu_{i'}} \cdot c_{ij}, \quad \text{for } i = 1, 2, \quad \text{with } \{i, i'\} = \{1, 2\}, \quad j = 1, 2, 3, 4,$$

$$
\begin{aligned}
a_{13} &= \lambda_1+\mu_2+\nu_1-2\rho_1-\rho_2, & a_{14} &= \lambda_1+\mu_1+\nu_2-2\rho_1-\rho_2,\\
a_{15} &= \lambda_2+\mu_2+\nu_1-2\rho_1-\rho_2, & a_{16} &= \lambda_2+\mu_1+\nu_2-2\rho_1-\rho_2,\\
a_{23} &= \lambda_2+\mu_2+\nu_2-2\rho_1-\rho_2, & a_{24} &= \lambda_2+\mu_1+\nu_1-2\rho_1-\rho_2,\\
a_{25} &= \lambda_2+\mu_2+\nu_2-2\rho_1-\rho_2, & a_{26} &= \lambda_2+\mu_1+\nu_1-2\rho_1-\rho_2,\\
b_{11} &= \lambda_2+\mu_1+\nu_1-2\rho_1-\rho_2, & b_{12} &= \lambda_1+\mu_1+\nu_2-2\rho_1-\rho_2,\\
b_{15} &= \lambda_1+\mu_1+\nu_2-2\rho_1-\rho_2, & b_{16} &= \lambda_1+\mu_2+\nu_2-2\rho_1-\rho_2,\\
b_{21} &= \lambda_2+\mu_2+\nu_2-2\rho_1-\rho_2, & b_{22} &= \lambda_1+\mu_2+\nu_1-2\rho_1-\rho_2,\\
b_{25} &= \lambda_1+\mu_1+\nu_1-2\rho_1-\rho_2, & b_{26} &= \lambda_1+\mu_2+\nu_1-2\rho_1-\rho_2,\\
c_{11} &= \lambda_1+\mu_2+\nu_2-2\rho_1-\rho_2, & c_{12} &= \lambda_1+\mu_2+\nu_1-2\rho_1-\rho_2,\\
c_{13} &= \lambda_1+\mu_2+\nu_1-2\rho_1-\rho_2, & c_{14} &= \lambda_1+\mu_2+\nu_2-2\rho_1-\rho_2,\\
c_{21} &= \lambda_1+\mu_1+\nu_1-2\rho_1-\rho_2, & c_{22} &= \lambda_1+\mu_1+\nu_2-2\rho_1-\rho_2,\\
c_{23} &= \lambda_1+\mu_1+\nu_1-2\rho_1-\rho_2, & c_{24} &= \lambda_1+\mu_1+\nu_2-2\rho_1-\rho_2.
\end{aligned}
$$

The Jordan canonical form of A_{IV^*} is

$$
\begin{pmatrix} \rho_1 I_4 & \\ & \rho_2 I_2 \end{pmatrix}.
$$

The system (2.3.1) is Fuchsian over $\mathbf{P}^1(\mathbf{C})$ with regular singular points at $x = t_1, t_2, t_3, \infty$, and is free from accessory parameters.

We assume

$$
\begin{aligned}
&\rho_1, \rho_2 \notin \mathbf{Z}_{<0}, && \rho_1-\rho_2 \notin \mathbf{Z},\\
&\lambda_1, \lambda_2 \notin \mathbf{Z}, && \lambda_1-\lambda_2 \notin \mathbf{Z},\\
&\mu_1, \mu_2 \notin \mathbf{Z}, && \mu_1-\mu_2 \notin \mathbf{Z},\\
&\nu_1, \nu_2 \notin \mathbf{Z}, && \nu_1-\nu_2 \notin \mathbf{Z}.
\end{aligned} \tag{2.3.2}
$$

Then the system (2.3.1) has 2 singular solutions

$$
y_i(x) = (x-t_1)^{\lambda_i}\left(e_i + O(x-t_1)\right), \qquad i = 1,2
$$

in the domain B_1, 2 singular solutions

$$
z_i(x) = (x-t_2)^{\mu_i}\left(e_{2+i} + O(x-t_2)\right), \qquad i = 1,2
$$

in the domain B_2 and 2 singular solutions

$$
w_i(x) = (x-t_3)^{\nu_i}\left(e_{4+i} + O(x-t_3)\right), \qquad i = 1,2
$$

in the domain B_3. Hence we can define a matrix solution

$$
Y_0(x) = \left(y_1(x)\, y_2(x)\, z_1(x)\, z_2(x)\, w_1(x)\, w_2(x)\right)
$$

for $x \in W = B_1 \cap B_2 \cap B_3$. We see that $Y_0(x)$ is a fundamental matrix solution (cf. Proposition 2).

THEOREM 11. *We assume* (2.3.2) *and*

$$
\begin{aligned}
&\lambda_i-\rho_1 \notin \mathbf{Z} && \text{for } i = 1,2,\\
&\mu_i-\rho_1 \notin \mathbf{Z} && \text{for } i = 1,2,\\
&\nu_i-\rho_1 \notin \mathbf{Z} && \text{for } i = 1,2,\\
&\lambda_i+\mu_j+\nu_k-(2\rho_1+\rho_2) \notin \mathbf{Z} && \text{for } i,j,k = 1,2.
\end{aligned} \tag{2.3.3}
$$

There is a diagonal matrix $D \in \mathrm{GL}(6, \mathbf{C})$ *such that the monodromy representation*

$$R_{\mathrm{IV}^*} : \pi_1 \left(\mathbf{P}^1 (\mathbf{C}) \setminus \{t_1, t_2, t_3, \infty\}, x_0\right) \to \mathrm{GL}\,(6, \mathbf{C})$$

of the system (2.3.1) *with respect to* $Y(x) = Y_0(x)D$ *is given by*

$$R_{\mathrm{IV}^*}([\gamma_1]) = \begin{pmatrix} e(\lambda_1) & & \xi_{11} & \xi_{12} & \xi_{13} & \xi_{14} \\ & e(\lambda_2) & \xi_{21} & \xi_{22} & \xi_{23} & \xi_{24} \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix},$$

$$R_{\mathrm{IV}^*}([\gamma_2]) = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ \eta_{11} & \eta_{12} & e(\mu_1) & & \eta_{13} & \eta_{14} \\ \eta_{21} & \eta_{22} & & e(\mu_2) & \eta_{23} & \eta_{24} \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix},$$

$$R_{\mathrm{IV}^*}([\gamma_3]) = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ \zeta_{11} & \zeta_{12} & \zeta_{13} & \zeta_{14} & e(\nu_1) & \\ \zeta_{21} & \zeta_{22} & \zeta_{23} & \zeta_{24} & & e(\nu_2) \end{pmatrix},$$

where

$$\begin{aligned} \xi_{ij} &= \frac{e(\lambda_i) - e(\rho_1)}{e(\lambda_i) - e(\lambda_{i'})} \cdot x_{ij}, \quad \textit{for } i = 1,2, \quad \{i, i'\} = \{1,2\}, \quad j = 1, \ldots, 4, \\ \eta_{ij} &= \frac{e(\mu_i) - e(\rho_1)}{e(\mu_i) - e(\mu_{i'})} \cdot y_{ij}, \quad \textit{for } i = 1,2, \quad \{i, i'\} = \{1,2\}, \quad j = 1, \ldots, 4, \\ \zeta_{ij} &= \frac{e(\nu_i) - e(\rho_1)}{e(\nu_i) - e(\nu_{i'})} \cdot z_{ij}, \quad \textit{for } i = 1,2, \quad \{i, i'\} = \{1,2\}, \quad j = 1, \ldots, 4, \end{aligned} \tag{2.3.4}$$

$$\begin{aligned} x_{11} &= \frac{[211]}{e(\nu_1 + 2\rho_1 + \rho_2)}, \qquad x_{12} = \frac{[111]}{e(\nu_1 + 2\rho_1 + \rho_2)}, \\ x_{13} &= -\frac{[122]}{e(\lambda_1 + \mu_2 + \nu_1 + 3\rho_1 + \rho_2)}, \qquad x_{14} = \frac{[111]}{e(\lambda_1 + \mu_1 + \nu_1 + 3\rho_1 + \rho_2)}, \\ x_{21} &= \frac{[221]}{e(\nu_1 + 2\rho_1 + \rho_2)}, \qquad x_{22} = \frac{[121]}{e(\nu_1 + 2\rho_1 + \rho_2)}, \\ x_{23} &= -\frac{[212]}{e(\lambda_2 + \mu_1 + \nu_1 + 3\rho_1 + \rho_2)}, \qquad x_{24} = \frac{[221]}{e(\lambda_2 + \mu_1 + \nu_1 + 3\rho_1 + \rho_2)}, \end{aligned}$$

$$y_{11} = \frac{[121]}{e(\rho_1)}, \qquad y_{12} = \frac{[111]}{e(\rho_1)},$$
$$y_{13} = \frac{[212]}{e(\lambda_2 + 3\rho_1 + \rho_2)}, \qquad y_{14} = \frac{[222]}{e(\lambda_2 + 3\rho_1 + \rho_2)},$$
$$y_{21} = \frac{[221]}{e(\rho_1)}, \qquad y_{22} = \frac{[211]}{e(\rho_1)},$$
$$y_{23} = -\frac{[211]}{e(\lambda_1 + \lambda_2 + \mu_1 + \nu_1 + \rho_1)}, \qquad y_{24} = \frac{[221]}{e(\lambda_1 + \lambda_2 + \mu_1 + \nu_1 + \rho_1)},$$
$$z_{11} = -e(\lambda_1 + \nu_1)[221], \qquad z_{12} = -e(\lambda_2 + \nu_1)[111],$$
$$z_{13} = [221], \qquad z_{14} = [111,]$$
$$z_{21} = -\frac{e(2\rho_1 + \rho_2)[121]}{e(\mu_2)}, \qquad z_{22} = -\frac{e(2\rho_1 + \rho_2)[211]}{e(\mu_2)},$$
$$z_{23} = -[211], \qquad z_{24} = \frac{e(2\rho_1 + \rho_2)[212]}{e(\lambda_2 + \mu_2 + \nu_2)}.$$

Here we have set

$$[ijk] = e(\lambda_i + \mu_j + \nu_k) - e(2\rho_1 + \rho_2)$$

for $i, j, k = 1, 2$.

We denote by $\mathcal{M}_{\mathrm{IV}^*,\lambda,\mu,\nu,\rho}$ the image of the monodromy representation R_{IV^*} given in Theorem 11; namely,

$$\mathcal{M}_{\mathrm{IV}^*,\lambda,\mu,\nu,\rho} = \langle R_{\mathrm{IV}^*}([\gamma_1]), R_{\mathrm{IV}^*}([\gamma_2]), R_{\mathrm{IV}^*}([\gamma_3])\rangle.$$

Now we assume that $\lambda_1, \lambda_2, \mu_1, \mu_2, \nu_1, \nu_2, \rho_1, \rho_2$ are *real* numbers, and give the Hermitian forms invariant under $\mathcal{M}_{\mathrm{IV}^*,\lambda,\mu,\nu,\rho}$.

THEOREM 12. *We assume* (2.3.2) *and* (2.3.3). *Let* H *be the Hermitian matrix associated with a Hermitian form* h *invariant under* $\mathcal{M}_{\mathrm{IV}^*,\lambda,\mu,\nu,\rho}$. *Then there is a real number* α *such that*

$$H = \alpha \begin{pmatrix} h_1 & & h_{11} & h_{12} & h_{13} & h_{14} \\ & h_2 & h_{21} & h_{22} & h_{23} & h_{24} \\ \bar{h}_{11} & \bar{h}_{21} & g_1 & & g_{11} & g_{12} \\ \bar{h}_{12} & \bar{h}_{22} & & g_2 & g_{21} & g_{22} \\ \bar{h}_{13} & \bar{h}_{23} & \bar{g}_{11} & \bar{g}_{21} & k_1 & \\ \bar{h}_{14} & \bar{h}_{24} & \bar{g}_{12} & \bar{g}_{22} & & k_2 \end{pmatrix},$$

where

$$h_i = \frac{\sin\pi\lambda_i \cdot \sin\pi(\lambda_i - \lambda_{i'})}{\sin\pi(\lambda_i - \rho_1)} \cdot \prod_{k=1,2} \sin\pi(\lambda_k + \mu_i + \nu_2 - 2\rho_1 - \rho_2),$$
$$g_i = \frac{\sin\pi\mu_i \cdot \sin\pi(\mu_i - \mu_{i'})}{\sin\pi(\mu_i - \rho_1)} \cdot \prod_{k=1,2} \sin\pi(\lambda_i + \mu_k + \nu_2 - 2\rho_1 - \rho_2),$$
$$k_i = \frac{\sin\pi\nu_i \cdot \sin\pi(\nu_i - \nu_{i'})}{\sin\pi(\nu_i - \rho_1)} \cdot \prod_{k=1,2} \sin\pi(\lambda_i + \mu_2 + \nu_k - 2\rho_1 - \rho_2)$$

for $i = 1, 2, \{i, i'\} = \{1, 2\}$,

$$h_{ij} = \frac{\xi_{ij}}{e(\lambda_i) - 1} \cdot h_i, \qquad i = 1, 2, \quad j = 1, \ldots, 4,$$
$$g_{ij} = \frac{\eta_{ij+2}}{e(\mu_i) - 1} \cdot g_i, \qquad i, j = 1, 2,$$

and ξ_{ij}, η_{ij} *are given by* (2.3.4).

Note added in proof. The author has discovered that the system (II) is studied in detail in [ST2].

REFERENCES

[BH] F. BEUKERS AND G. HECKMAN, *Monodromy for the hypergeometric function* ${}_nF_{n-1}$, Invent. Math., 95 (1989), pp. 325–354.

[H1] Y. HARAOKA, *Finite monodromy of Pochhammer equation*, Ann. Inst. Fourier, to appear.

[H2] ———, *Canonical forms of differential equations free from accessory parameters*, SIAM J. Math. Anal., 25 (1994), pp. 1203–1226.

[IKSY] K. IWASAKI, H. KIMURA, S. SHIMOMURA, AND M. YOSHIDA, *From Gauss to Painlevè*, Vieweg, Wiesbaden, 1991.

[L] A. H. M. LEVELT, *Hypergeometric Functions*, Ph.D. Thesis, University of Amsterdam, 1961.

[M] N. MISAKI, *Reducibility condition of Pochhammer's equation*, Master Thesis, Tokyo Univ., 1973. (In Japanese.)

[O] K. OKUBO, *On the group of Fuchsian equations*, Seminar Reports of Tokyo Metropolitan University, 1987.

[OTY] K. OKUBO, K. TAKANO, AND S. YOSHIDA, *A connection problem for the generalized hypergeometric equation*, Funkcial. Ekvac., 31 (1988), pp. 483–485.

[S] T. SASAI, *On a monodromy group and irreducibility conditions of a fourth order Fuchsian differential system of Okubo type*, J. Reine Angew. Math, 299/300 (1978), pp. 38–50.

[ST] T. SASAI AND S. TSUCHIYA, *On a fourth order Fuchsian differential equation of Okubo type*, Funkcial. Ekvac., 34 (1991), pp. 211–221.

[ST2] ———, *On a class of even order Fuchsian equations of Okubo type*, Funkcial. Ekvac., 35 (1992), pp. 505–514.

[TB] K. TAKANO AND E. BANNAI, *A global study of Jordan–Pochhammer differential equations*, Funkcial. Ekvac., 19 (1976), pp. 85–99.

[Y] T. YOKOYAMA, *On an Irreducibility Condition for Hypergeometric Systems*, Funkcial. Ekvac., to appear.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1622–1641, November 1994

011

SMOOTHING EFFECTS FOR DISPERSIVE EQUATIONS VIA A GENERALIZED WIGNER TRANSFORM*

THIERRY COLIN†

Abstract. Generalized Wigner transforms, which are adapted to several linear dispersive equations, are introduced. In applying these transforms some local smoothing effects are recovered, and the estimates on the solutions are found.

Key words. dispersive equations, smoothing effects, Wigner transform

AMS subject classifications. 35Q20, 35B65, 35A22

1. Introduction.

In a recent work, Lions and Perthame [5] have shown that the classical Wigner transform (see (2) below) can be used in order to recover the smoothing effects of the linear Schrödinger equation. More precisely, consider a solution ψ to

$$i\frac{\partial \psi}{\partial t} + \Delta \psi = 0 \quad \text{in} \quad \mathbb{R}^n \times \mathbb{R}, \tag{1}$$

and set

$$f(x;\xi;t) = \int_{\mathbb{R}} e^{-iy\cdot\xi} \psi(x+y)\,\overline{\psi}(x-y)\,dy, \quad x \in \mathbb{R}^n, \quad \xi \in \mathbb{R}^n. \tag{2}$$

The mapping $\psi \mapsto f$ is known as the Wigner transform; it has the remarkable property that if ψ satisfies (1), then f satisfies the linear transport equation

$$\frac{\partial f}{\partial t} - (\xi \cdot \nabla_x) f = 0 \quad \text{in} \quad \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}. \tag{3}$$

The following question naturally arises. If, instead of (1), ψ is the solution to a more general linear dispersive equation

$$i\frac{\partial \psi}{\partial t} + P(D)\,\psi = 0, \tag{4}$$

where this once $P(D)$ is a general differential operator with real symbol $P(\xi)$, is it possible to recover a similar framework? At least two directions are possible. First, one can keep the transform (2) and try to find an analog of (3). We have not been able to find a "reasonable" equation. Second, one can try a new transform that again leads to (3). Let us denote by $\hat{\psi}$ the Fourier transform of ψ with respect to the space variable x. In this direction, we have the following proposition.

PROPOSITION 1. *Let* $q(\xi,\eta) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ *be such that*

$$\xi \cdot q(\xi,\eta) = P(\eta) - P(\xi - \eta), \tag{5}$$

* Received by the editors December 29, 1992; accepted for publication (in revised form) August 3, 1993.

† Centre de Mathématiques et de Leurs Applications, Ecole Normale Supérieure de Cachan et Centre National de le Recherche Scientifique URA 1611, 61 Avenue du Président Wilson, 94235 Cachan Cedex, France.

and define the generalized Wigner transform (GWT) by

$$GWT(\psi)(x;\xi) = \int_{\mathbb{R}^n} e^{ix\cdot q(\xi;\eta)}\, \hat{\psi}(\eta)\, \overline{\hat{\psi}(\xi-\eta)}\, d\eta.$$

Then $f(x;\xi;t) = GWT(\psi(x;t))$ *solves* (3) *if* ψ *solves* (4). □

Remark 1. (i) If $P(\xi) = |\xi|^2$ then $q(\xi,\eta) = 2\eta - \xi$ satisfies (5) and one recovers the classical Wigner transform (2).

(ii) For the moment, Proposition 1 should be seen as the result of a formal computation since the phase $q(\xi;\eta)$ can be singular. We shall make this more precise later.

2. Statement of the main results.

2.1. The one-dimensional case.

THEOREM 1. *Let* $\psi : \mathbb{R}\times\mathbb{R} \to \mathbb{C}$ *be the solution of* (4) *with initial value* $\psi_0(x)$.
(a) *If* $P(\xi)$ *is a polynomial with odd degree, then*

$$\forall x \in \mathbb{R}, \qquad \int_{-\infty}^{+\infty} \left| \int e^{-ix\cdot\eta}\, \hat{\psi}(\eta)\, \varphi(2\eta)\, |P'(\eta)|^{1/2}\, d\eta \right|^2 dt = 8\pi \int |\hat{\psi}_0(\xi)|^2\, \varphi(2\xi)^2\, d\xi.$$

(b) *If* $P(\xi)$ *is an even polynomial then*

$$\forall x \in \mathbb{R}, \qquad \int_{-\infty}^{+\infty} \left| \int e^{-ix\cdot\eta}\, \hat{\psi}(\eta)\, \varphi(2\eta)^2\, |P'(\eta)|^{1/2}\, d\eta \right|^2 dt = 8\pi \left(\int |\hat{\psi}_0(\xi)|^2\, \varphi(\xi)^4\, d\xi - \int e^{2ix\cdot\xi}\, \hat{\psi}_0(\xi)\, \overline{\hat{\psi}_0(-\xi)}\, \varphi(\xi)^4\, d\xi \right),$$

where φ *is a cut-off function*

$$\begin{cases} \varphi(\xi) \equiv 1 & \text{if } |\xi| \geq 3R, \\ \varphi(\xi) \equiv 0 & \text{if } |\xi| \leq 2R. \end{cases} \qquad \square$$

Remark 2. We recover the local smoothing effect for dispersive equations (see Constantin and Saut [3] and Kenig, Ponce, and Vega [4]).

Indeed, Theorem 1 shows that

$$\left| \mathcal{F}^{-1}\left(\hat{\psi}(\eta)\varphi(2\eta)\, |P'(\eta)|^{1/2} \right) \right|^2_{L^\infty_x(\mathbb{R};L^2_t)} \leq c|\psi_0|^2_{L^2},$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform in the η variable. On the other hand,

$$\mathcal{F}^{-1}\left(\hat{\psi}(\eta)(1-\varphi(2\eta))\, |P'(\eta)|^{1/2} \right) \in L^\infty(\mathbb{R}_x \times \mathbb{R}_t),$$

which leads to

$$\left| (|P'|^{1/2}(D))\, \psi \right|_{L^\infty_x(\mathbb{R};L^2_{\mathrm{loc},t})} \leq c|\psi_0|_{L^2},$$

where $(|P'|^{1/2}(D))\, \psi = \mathcal{F}^{-1}(|P'(\eta)|^{1/2}\hat{\psi}(\eta))(x)$.

Remark 3. If $P(\xi) = |\xi|^{\alpha+1}$ or $P(\xi) = |\xi|^\alpha \xi$, one can take $\varphi \equiv 1$ and one obtains

(i) $\forall x \in \mathbb{R}$,

$$\int_{-\infty}^{+\infty} \left| |D|^{\alpha/2} \psi(x;t) \right|^2 dt = c_\alpha \left(\int |\hat{\psi}_0(\xi)|^2 d\xi - \int e^{2ix\cdot\xi}\, \hat{\psi}_0(\xi) \overline{\hat{\psi}_0(-\xi)}\, d\xi \right)$$
$$\text{for} \quad P(\xi) = |\xi|^{\alpha+1},$$

or

(ii) $\forall x \in \mathbb{R}$,

$$\int_{-\infty}^{+\infty} \left| |D|^{\alpha/2}\, \psi(x;t) \right|^2 dt = c_\alpha \int |\hat{\psi}_0(\xi)|^2\, d\xi, \quad \text{for} \quad P(\xi) = |\xi|^\alpha \xi,$$

where we have used the notation $|D|^\alpha \psi(x) = \mathcal{F}^{-1}(|\xi|^\alpha \hat{\psi}(\xi))(x)$.

For the Airy equation, we obtain the following corollary.

COROLLARY. *Let ψ be the solution to the Airy equation:*

$$\begin{cases} \dfrac{\partial \psi}{\partial t} + \dfrac{\partial^3 \psi}{\partial x^3} = 0, \\ \psi(x;0) = \psi_0(x). \end{cases}$$

One has

$$\forall x \in \mathbb{R}, \qquad \int_{-\infty}^{+\infty} |\psi_x|^2\,(x;t)\, dt = c|\psi_0|^2_{L^2(\mathbb{R})}.$$

We recover the result of [4].

Remark 4. Theorem 1 remains valid for any pseudo-differential operator whose symbol satisfies the conclusions of Lemmas 1 and 2 or Lemma 3 in §3.1 below.

2.2 Generalizations in arbitrary dimensions.

2.2.1. The case of a power of the Laplacian. We consider the equation

$$\begin{cases} i\psi_t + (-\Delta)^\alpha \psi = 0 \quad \text{in} \quad \mathbb{R} \times \mathbb{R}^n, \\ \psi(x;0) = \psi_0(x), \end{cases} \tag{6}$$

where $\alpha \geq 1$.

Our results read as follows.

THEOREM 2. *The solution $\psi(x,t)$ to (6) satisfies the following: $\forall x_0 \in \mathbb{R}^n$,*

• $n = 2$:

$$\int_{-\infty}^{+\infty} \int_{\mathbb{R}^2} \frac{|\nabla(|D|^{\alpha-1}\psi)|^2}{|x-x_0|} - \frac{((x-x_0)\cdot\nabla(|D|^{\alpha-1}\psi))^2}{|x-x_0|^3}\, dx\, dt$$
$$\leq c|\psi_0|^2_{\dot{H}^{1/2}};$$

• $n = 3$:

$$\int_{-\infty}^{+\infty} \left\{ \int_{\mathbb{R}^3} \frac{|\nabla(|D|^{\alpha-1}\psi)|^2}{|x-x_0|} - \frac{((x-x_0)\cdot\nabla(|D|^{\alpha-1}\psi))^2}{|x-x_0|^3}\, dx \right.$$
$$\left. + \left| |D|^{\alpha-1}\psi(x_0;t) \right|^2 \right\} dt = c|\psi_0|^2_{\dot{H}^{1/2}};$$

• $n \geq 4$:

$$\int_{-\infty}^{+\infty} \int_{\mathbb{R}^n} \left\{ \frac{|\nabla(|D|^{\alpha-1}\psi)|^2}{|x-x_0|} - \frac{((x-x_0)\cdot\nabla(|D|^{\alpha-1}\psi))^2}{|x-x_0|^3} + \frac{||D|^{\alpha-1}\psi|^2}{|x-x_0|^3}\,\alpha_n \right\} dx\,dt = c|\psi_0|^2_{\dot{H}^{1/2}}. \qquad \square$$

THEOREM 3. *The solution ψ to* (6) *satisfies*

$$\int_{-\infty}^{+\infty} \int \frac{\left|\nabla(|D|^{\alpha-1}\psi)(y)\right|^2}{\left|1+|y|^{\delta}\right|^{1+1/\delta}}\,dy\,dt \leq c|\psi_0|_{\dot{H}^{1/2}}$$

for $0 < \delta \leq 1$. $\square$

Theorems 2 and 3 are the results corresponding to those of Lions and Perthame [5] in the case where $P(D)$ is the Laplacian. They improve the known results on the local smoothing effects (see [3] and [4]). Theorem 3 implies, for example, that

$$\left||D|^{(\alpha-1/2)}\psi\right|_{L^2_t(\mathbb{R};\,L^2_{\mathrm{loc},x}(\mathbb{R}^n))} \leq c(\alpha,\delta)\,|\psi_0|_{L^2(\mathbb{R}^n)}.$$

The estimate in Theorem 3 is analogous to the one given by Ben-Artzi and Devinatz in [1].

2.2.2 Case of a tensorial symbol. We consider the equation

$$\text{(7)} \qquad \begin{cases} i\dfrac{\partial\psi}{\partial t} + \displaystyle\sum_{j=1}^{n} P_j(D_j)\psi = 0 \quad \text{in} \quad \mathbb{R}\times\mathbb{R}^n, \\ \psi(x;0) = \psi_0(x), \end{cases}$$

where the symbols $P_j(\xi_j)$ of $P_j(D_j)$ are real and even. We impose the requirement that P_j' be a bijection from $\mathbb{R}$ onto $\mathbb{R}$. We have

$$P_j(\eta_j) - P_j(\xi_j - \eta_j) = -\xi_j \int_0^1 P_j'(\eta_j + t\xi_j)\,dt,$$

and since P_j' is odd

$$\text{(8)} \qquad -\int_0^1 P_j'(\eta_j - t\xi_j)dt = (2\eta_j - \xi_j)\,g_j(\xi_j,\eta_j),$$

where g_j is a regular function.

THEOREM 4. *The function ψ solution to* (7) *satisfies*

$$\forall y_1 \in \mathbb{R}, \qquad \int_{-\infty}^{+\infty} \int_{\mathbb{R}^{n-1}} \left| \left(|P_1'|^{1/2} \left(\frac{\partial}{\partial y_1} \right) \psi \right) \right|^2 dy_2 \ldots dy_n\,dt = c \int \left(|\hat{\psi}_0(\xi)|^2 - e^{-2iy_1\xi_1}\,\hat{\psi}_0(\xi)\,\overline{\hat{\psi}_0(-\xi)} \right) d\xi. \quad \square$$

The results of this paper were announced in [2].

3. One-dimensional case.

3.1. Technical results. As in (5), we introduce the following function:

$$q(\xi,\eta)=\frac{P(\eta)-P(\xi-\eta)}{\xi}.$$

The aim of this section is to prove some results that we will use in the course of the proof of Theorem 1.

LEMMA 1. *If P is an even polynomial or if the degree of P is odd, then $\exists R_1\geq 0$ such that if $|\xi|\geq R_1$, then $\eta\mapsto q(\xi;\eta)$ is a diffeomorphism from $\mathbb{R}$ to $\mathbb{R}$.* □

Proof. First we notice that

$$\frac{\partial q}{\partial \eta}=\frac{P'(\eta)+P'(\xi-\eta)}{\xi}.$$

• If P is even, then P' is odd and

$$\frac{\partial q}{\partial \eta}=\frac{P'(\eta)+P'(\eta-\xi)}{\xi}.$$

Now $\eta\mapsto P'(\eta)$ and $\eta\mapsto P'(\eta-\xi)$ are two translated odd polynomials; therefore, if $|\xi|$ is sufficiently large, they do not have any intersection point and $\partial q/\partial\eta\neq 0\quad\forall\eta\in\mathbb{R}$.

• If P is a polynomial with odd degree, then P' is a polynomial with even degree and it is clear that if $|\xi|$ is large enough, then

$$\eta\mapsto\frac{P'(\eta)+P'(\xi-\eta)}{\xi}$$

cannot vanish.

Now, in both cases, it is easy to verify that $\partial q/\partial\eta$ is a polynomial in the η variable with odd degree. The lemma follows. □

LEMMA 2. *If P is even, then $\exists R_2>0$ such that $\forall\xi\in\mathbb{R}$, $\eta\mapsto q(\xi;\eta)$ is a diffeomorphism from $\mathbb{R}\setminus[-R_2;R_2]$ onto its range.* □

LEMMA 3. *If the degree of P is odd, then $\exists R_3>0$ such that $\forall\xi\in\mathbb{R}$, $\eta\mapsto P(\eta)-P(\xi-\eta)$ is a diffeomorphism from $\mathbb{R}\setminus[-R_3;R_3]$ onto its range.* □.

Proofs. • If P is even, then the coefficient of the dominant term of $\eta\mapsto q(\xi;\eta)$ does not depend on ξ; hence Lemma 2 follows from Lemma 1 and from the continuity of the roots of a polynomial with respect to the coefficients.

• If the degree of P is odd, we make the same remark with $\eta\mapsto P(\eta)-P(\xi-\eta)$ instead of $\eta\mapsto q(\xi,\eta)$. □

3.2. Proof of Theorem 1 in the case where the degree of P is odd. Let P be a polynomial with odd degree, let q be given by

$$q(\xi,\eta)=\frac{P(\eta)-P(\xi-\eta)}{\xi},$$

and let R_1,R_3 be associated with P according to Lemmas 1 and 3. We introduce the following function:

$$\begin{cases}\varphi\equiv 1 & \text{if}\quad |\xi|\geq 3R,\\ \varphi\equiv 0 & \text{if}\quad |\xi|\leq 2R,\end{cases}$$

with $0\leq\varphi\leq 1$, φ even, and $R=\max(R_1,R_3)$.

Let $\psi \in \mathcal{S}(\mathbb{R})$; define the generalized Wigner transform of ψ as in Proposition 1:

$$\mathrm{GWT1}(\psi)(x;\xi) = \varphi(\xi) \int \hat{\psi}(\eta)\, \overline{\hat{\psi}(\xi-\eta)} e^{ixq(\xi,\eta)}\, d\eta.$$

We now consider the case where ψ depends on the time t and satisfies (4), i.e.,

$$\begin{cases} i\dfrac{\partial \psi}{\partial t} + P(D)\psi = 0, & x,t \in \mathbb{R}\times\mathbb{R}, \\ \psi(x;0) = \psi_0(x), & x \in \mathbb{R}, \end{cases}$$

and we introduce

$$f(x;\xi;t) = \mathrm{GWT1}(\psi(\cdot,t)). \tag{9}$$

It follows that ψ is given by $\hat{\psi}(\xi;t) = \hat{\psi}_0(\xi)\, e^{iP(\xi)t}$. Hence f satisfies the transport equation (3), i.e.,

$$\frac{\partial f}{\partial t} - \xi \frac{\partial f}{\partial x} = 0, \qquad t,\xi,x \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}.$$

We multiply (3) by $\frac{x}{|x|}\,\xi$ and we integrate on $\mathbb{R}\times\mathbb{R}$ with respect to x,ξ, and on $[-T;T]$ with respect to t. One obtains

$$\begin{aligned} I(T) &\equiv \iint f(x;\xi;T)\,\frac{x}{|x|}\,\xi\, d\xi\, dx - \iint f(x;\xi;-T)\,\frac{x}{|x|}\,\xi\, d\xi\, dx \\ &= \int_{-T}^{T} \iint \frac{\partial f}{\partial x}(x;\xi;t)\,\frac{x}{|x|}\,\xi^2\, d\xi\, dx\, dt \equiv II(T). \end{aligned} \tag{10}$$

At this stage, we let $T \to +\infty$ and we will identify the limits of $I(T)$ and $II(T)$.

(i) *Limit of $I(T)$ as $T \to +\infty$.* Since f satisfies (3), we have an explicit formula,

$$f(x;\xi;T) = f_0(x+\xi T\,;\,\xi),$$

where $f_0(y;\xi) = f(y;\xi;0)$.

Hence

$$I(T) = \iint f_0(x+\xi T,\xi)\,\frac{x}{|x|}\,\xi\, dx\, d\xi - \iint f_0(x-\xi T;\xi)\,\frac{x}{|x|}\,\xi\, dx\, d\xi.$$

We make the changes of variables $y = x+\xi T$ and $z = x-\xi T$ in the above expression, and letting $T \to +\infty$ we get

$$I(T) \underset{T\to+\infty}{\to} -2\iint f_0(y,\xi)\,|\xi|\,\varphi(\xi)\, d\xi\, dy. \tag{11}$$

Using (9) in (11) leads to

$$\lim_{T\to+\infty} I(T) = -2\iiint \hat{\psi}_0(\eta)\,\overline{\hat{\psi}_0(\xi-\eta)}\,|\xi|\, e^{iyq(\xi,\eta)}\,\varphi(\xi)\, d\xi\, d\eta\, dy.$$

Thanks to Lemma 1, $\eta \mapsto z = q(\xi,\eta)$ is a diffeomorphism for $|\xi| \geq R$; hence, letting $\eta = f(\xi,z)$, we obtain

$$\lim_{T\to+\infty} I(T) = -2\iiint \frac{\hat{\psi}_0(f(\xi;z))|\xi|\,\overline{\hat{\psi}_0(\xi - f(\xi;z))}\,\varphi(\xi)\, e^{iyz}}{\left|\frac{\partial q}{\partial \eta}(\xi;f(\xi;z))\right|}\, dy\, dz\, d\xi. \tag{12}$$

In (12), we use the formula $\int \hat{\varphi}(\xi)\,d\xi = 2\pi\,\varphi(0)$ and we are led to

$$\lim_{T\to+\infty} I(T) = -4\pi \int \frac{\hat{\psi}_0(f(\xi;0))|\xi|\,\overline{\hat{\psi}_0(\xi - f(\xi;0))}\,\varphi(\xi)}{\left|\frac{\partial q}{\partial \eta}(\xi; f(\xi;0))\right|}\,d\xi.$$

We remark that $f(\xi;0) = \xi/2$ and that

$$\frac{\partial q}{\partial \eta}(\xi;\xi/2) = \frac{2P'(\xi/2)}{\xi},$$

which implies

$$\lim_{T\to+\infty} I(T) = -4\pi \int \hat{\psi}_0(\xi/2)\,\overline{\hat{\psi}_0(\xi/2)}\,\xi^2 \frac{\varphi(\xi)}{|P'(\xi/2)|}\,d\xi;$$

hence

$$\lim_{T\to+\infty} I(T) = -32\pi \int |\hat{\psi}_0(\xi)|^2\,\xi^2 \frac{\varphi(2\xi)}{|P'(\xi)|}\,d\xi. \tag{13}$$

(ii) *Limit of $II(T)$ as $T \to +\infty$.* Let us recall that

$$II(T) = \int_{-T}^{T} \iint \frac{\partial f}{\partial x}(x;\xi;t)\,\frac{x}{|x|}\,\xi^2\,d\xi\,dx\,dt.$$

We integrate by parts in the x variable:

$$\begin{aligned} II(T) &= -\int_{-T}^{T} \int f(0;\xi,t)\xi^2\,d\xi\,dt, \\ &= -\int_{-T}^{T} \iint \hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,\xi^2\,\varphi(\xi)\,d\xi\,d\eta\,dt, \end{aligned}$$

by the definition of f (see (9)).

Let us now estimate the difference:

$$\begin{aligned} II'(T) \equiv \int_{-T}^{T} \iint \hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,[\xi^2\varphi(\xi) \\ - 4|\xi-\eta|\,\varphi(2(\xi-\eta))^{1/2}|\eta|\,\varphi(2\eta)^{1/2}]\,d\eta\,d\xi\,dt. \end{aligned}$$

Since ψ is given by $\hat{\psi}(\xi;t) = \hat{\psi}_0(\xi)\,e^{iP(\xi)t}$, one has

$$\begin{aligned} II'(T) = \iint \Big\{ \hat{\psi}_0(\eta)\,\overline{\hat{\psi}_0(\xi-\eta)}\,[\xi^2\varphi(\xi) - 4|\xi-\eta|\,\varphi(2(\xi-\eta))^{1/2}|\eta| \\ \cdot \varphi(2\eta)^{1/2}] \int_{-T}^{T} e^{i\xi q(\xi;\eta)t}\,dt \Big\}\,d\xi\,d\eta. \end{aligned}$$

Now, Lemmas 1 and 3 enable us to make the change of variables $z = \xi q(\xi;\eta) \Leftrightarrow \eta = g(\xi;z)$.

One obtains

$$\begin{aligned} II'(T) = \iint \Big\{ \hat{\psi}_0(\eta)\,\overline{\hat{\psi}_0(\xi-\eta)} \\ \cdot \frac{[\xi^2\varphi(\xi) - 4|\xi-\eta|\,\varphi(2(\xi-\eta))^{1/2}|\eta|\,\varphi(2\eta)^{1/2}]}{P'(\eta) + P'(\xi-\eta)} \int_{-T}^{T} e^{izt}\,dt \Big\}\,dz\,d\xi. \end{aligned} \tag{14}$$

First we remark that $\int_{-T}^{T} e^{izt}\,dt \rightharpoonup c\,\delta_0$ in the sense of measures and $g(\xi;0) = \xi/2$.

On the other hand,

$$\xi \mapsto \int \hat{\psi}_0(\eta)\,\overline{\hat{\psi}_0(\xi-\eta)}\,[\xi^2\varphi(\xi) - 4|\xi-\eta|\,\varphi(2(\xi-\eta))^{1/2}|\eta|\,\varphi(2\eta)^{1/2}] \cdot \int_{-T}^{T} e^{i\xi q(\xi;\eta)t}\,dt\,d\eta$$

is dominated almost everywhere by

$$\xi \mapsto \int |\hat{\psi}_0(\eta)|\,|\hat{\psi}_0(\xi-\eta)|\,\frac{|\xi^2\varphi(\xi) - 4|\xi-\eta|\,\varphi(2(\xi-\eta))^{1/2}|\eta|\,\varphi(2\eta)^{1/2}|}{P'(\eta)+P'(\xi-\eta)}\,dz,$$

which belongs to L^1 by Fubini's theorem.

Therefore, we can pass to the limit as $T \to +\infty$ in (14) and we obtain

$$\lim_{T\to+\infty} II'(T) = 0.$$

Hence, we deduce that

$$II(T) \underset{T\to+\infty}{\to} -4\int_{-\infty}^{+\infty} \iint \hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,|\xi-\eta|\,\varphi(2(\xi-\eta))^{1/2} \cdot |\eta|\,\varphi(2\eta)^{1/2}\,d\eta\,d\xi\,dt,$$

whence

$$\lim_{T\to+\infty} II(T) = -4\int_{-\infty}^{+\infty} \left| \int \hat{\psi}(\eta)\,|\eta|\,(\varphi(2\eta))^{1/2}\,d\eta \right|^2 dt. \tag{15}$$

(10), (13), and (15) yield

$$\int_{-\infty}^{+\infty} \left| \int \hat{\psi}(\eta)\,|\eta|\,\varphi(2\eta)^{1/2}\,d\eta \right|^2 dt = 8\pi \int_{-\infty}^{+\infty} |\hat{\psi}_0(\xi)|^2\,\frac{\xi^2\varphi(\xi)^2}{|P'(\xi)|}\,d\xi.$$

We apply this result to the function $\tilde{\psi}$ defined by

$$\widehat{\tilde{\psi}}(\xi;t) = \hat{\psi}(\xi;t)\,\frac{|P'(\xi)|^{1/2}}{|\xi|}\,\varphi(2\xi)^{1/2}.$$

This leads to the result of Theorem 1(a) with $x = 0$. The invariance under the translations ends the proof of Theorem 1(a). □

3.3. Proof of Theorem 1 in the case where P is an even polynomial. In this section, we consider the case where P is an even polynomial; thanks to Lemmas 1 and 2, we can find three nonnegative functions $K_1(\xi)$, $K_2(\xi)$, and $K_3(\xi)$ such that $\eta \mapsto P(\eta) - P(\xi-\eta)/\xi$ is a diffeomorphism from $\mathbb{R}\setminus]-K_1(\xi), K_1(\xi)[$ onto $\mathbb{R}\setminus]-K_2(\xi); K_3(\xi)[$ and $K_1 \equiv K_2 \equiv K_3 \equiv 0$ for $|\xi| \geq R_1$. We take $R = \max(R_1, R_2)$ and we introduce φ as in §3.2. We define a slightly different generalized Wigner transform for $\psi \in \mathcal{S}(\mathbb{R})$:

$$\mathrm{GWT2}(\psi)(x;\xi) = \int \hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,e^{ix\cdot q(\xi,\eta)}\,\varphi(\eta)\,\varphi(\xi-\eta)\,d\eta.$$

If ψ satisfies (4),

$$\begin{cases} i\,\dfrac{\partial \psi}{\partial t} + P(D)\psi = 0, & t, x \in \mathbb{R} \times \mathbb{R}, \\ \psi(x;0) = \psi_0(x), & x \in \mathbb{R}, \end{cases}$$

we introduce

$$f(x;\xi;t) = \mathrm{GWT2}(\psi(\cdot;t)), \tag{16}$$

and f satisfies the linear transport equation (3). Arguing as in §3.2, one obtains the equality (10).

(i) *Limit of $I(T)$ as $T \to +\infty$.* As in the previous section, we get

$$\begin{aligned} \lim_{T\to+\infty} I(T) &= -2 \iint f_0(y;\xi)\,|\xi|\,d\xi\,dy \\ &= -2 \iiint |\xi|\overline{\hat\psi_0(\eta)\,\hat\psi_0(\xi-\eta)}\,e^{iyq(\xi,\eta)}\,\varphi(\eta)\,\varphi(\xi-\eta)\,d\eta\,d\xi\,dy, \end{aligned}$$

by the definition of f (see (16)). Lemmas 1 and 2 enable us to make the change of variables $\eta \mapsto z = q(\xi,\eta)$ and $\eta = f(\xi,z)$. We obtain

$$\begin{aligned} &\lim_{T\to+\infty} I(T) \\ &= -2 \iiint \frac{\hat\psi_0(f(\xi;z))|\xi|\overline{\hat\psi_0(\xi-f(\xi,z))}\,e^{iyz}\varphi(f(\xi,z))\varphi(\xi-f(\xi,z))}{\left|\frac{\partial q}{\partial \eta}\,(\xi;f(\xi,z))\right|} \\ &\qquad \cdot \mathbf{1}_{\{\mathbb{R}\setminus]-K_2(\xi);K_3(\xi)[\}}(z)\,dy\,d\xi\,dz. \end{aligned}$$

Using the formula $\int \hat h(\xi)\,d\xi = 2\pi h(0)$, we obtain

$$\begin{aligned} &\lim_{T\to+\infty} I(T) \\ &= -4\pi \int \frac{\hat\psi_0(f(\xi;0))|\xi|\overline{\hat\psi_0(\xi-f(\xi,0))}\,\varphi(f(\xi,0))\varphi(\xi-f(\xi,0))}{\left|\frac{\partial q}{\partial \eta}\,(\xi;f(\xi,0))\right|} \\ &\qquad \cdot \mathbf{1}_{\{\mathbb{R}\setminus]-K_2(\xi);K_3(\xi)[\}}(0)\,d\xi. \end{aligned} \tag{17}$$

- If $0 \in\,]-K_2(\xi);K_3(\xi)[$, then the integrand in (17) is zero.
- If $0 \notin\,]-K_2(\xi);K_3(\xi)[$, then $f(\xi;0) = \xi/2$. Whence (17) yields

$$\lim_{T\to+\infty} I(T) = -4\pi \int |\hat\psi_0(\xi/2)|^2\,\frac{|\xi|\,\varphi(\xi/2)^2}{\left|\frac{\partial q}{\partial \eta}(\xi;\xi/2)\right|}\,d\xi,$$

and we obtain

$$\lim_{T\to+\infty} I(T) = -32\pi \int |\hat\psi_0(\xi)|^2\,\frac{\xi^2}{|P'(\xi)|}\,\varphi(\xi)^2\,d\xi. \tag{18}$$

(ii) *Limit of $II(T)$ as $T \to +\infty$.* As in §3.2, we obtain

$$II(T) = -\int_{-T}^{T} \iint \hat\psi(\eta)\,\overline{\hat\psi(\xi-\eta)}\,\xi^2\,\varphi(\eta)\,\varphi(\xi-\eta)\,d\eta\,d\xi\,dt. \tag{19}$$

The identity $\xi^2 = (\xi - 2\eta)^2 + 4\eta(\xi - \eta)$ and the change of variables $(\xi, \eta) \to (z = 2\eta - \xi, \eta)$ in (19) lead to

$$
\begin{aligned}
II(T) = &- \int_{-T}^{T} \iint \hat{\psi}(\eta)\, \overline{\hat{\psi}(\eta - z)}\, z^2\, \varphi(\eta)\, \varphi(\eta - z)\, d\eta\, dz\, dt \\
&- 4 \int_{-T}^{T} \iint \widehat{\psi'}(\eta)\, \overline{\widehat{\psi'}(\xi - \eta)}\, \varphi(\eta)\, \varphi(\xi - \eta)\, d\eta\, d\xi\, dt.
\end{aligned} \tag{20}
$$

Let us set $II'(T) = -\int_{-T}^{T} \iint z^2 \hat{\psi}(\eta)\, \overline{\hat{\psi}(\eta - z)}\, \varphi(\eta)\, \varphi(\eta - z)\, d\eta\, dz\, dt$. We can estimate this term by defining a Wigner-like transform

$$
\tilde{f}(x; \xi; t) = \int e^{ixq(\xi,\eta)}\, \hat{\psi}(\eta)\, \overline{\hat{\psi}(\eta - \xi)}\, \varphi(\eta)\, \varphi(\eta - \xi)\, d\eta. \tag{21}
$$

Since P is even, $\tilde{f}$ still satisfies the linear transport equation. The function $\tilde{f}(x; \xi; t)$ is "almost" the generalized Wigner transform defined by (16); therefore, arguing as in the beginning of this section, one obtains

$$
\begin{aligned}
&- 2 \iint \tilde{f}_0(y, \xi) |\xi|\, d\xi\, dy \\
&\qquad = - \int_{-\infty}^{+\infty} \iint z^2 \hat{\psi}(\eta)\, \overline{\hat{\psi}(\eta - z)}\, \varphi(\eta)\, \varphi(\eta - z)\, d\eta\, dz\, dt.
\end{aligned} \tag{22}
$$

As in §3.3(i), a change of variable in the left-hand side of (22) yields

$$
\iint \tilde{f}_0(y, \xi) |\xi|\, d\xi\, dy = 16\pi \int \hat{\psi}_0(\xi)\, \overline{\hat{\psi}_0(-\xi)}\, \xi^2\, \frac{\varphi(\xi)^2}{|P'(\xi)|}\, d\xi,
$$

which implies

$$
\lim_{T \to +\infty} II'(T) = -32\pi \int \hat{\psi}_0(\xi)\, \overline{\hat{\psi}_0(-\xi)}\, |\xi|^2\, \frac{|\varphi(\xi)|^2}{|P'(\xi)|}\, d\xi. \tag{23}
$$

Together with (20), (23) leads to

$$
\begin{aligned}
\lim_{T \to +\infty} II(T) = &- 4 \int_{-\infty}^{+\infty} \left| \int \widehat{\psi'}(\eta)\, \varphi(\eta)\, d\eta \right|^2 dt \\
&- 32\pi \int \hat{\psi}_0(\xi)\, \overline{\hat{\psi}_0(-\xi)}\, \frac{|\xi|^2 \varphi(\xi)^2}{|P'(\xi)|}\, d\xi.
\end{aligned} \tag{24}
$$

Now (10), (18), and (24) yield

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} \left| \int \widehat{\psi'}(\eta)\, \varphi(\eta)\, d\eta \right|^2 dt \\
&= 8\pi \left\{ \int |\hat{\psi}_0(\xi)|^2\, \frac{\xi\, \varphi(\xi)^2}{|P'(\xi)|}\, d\xi - \int \hat{\psi}_0(\xi)\, \overline{\hat{\psi}_0(-\xi)}\, \frac{|\xi|^2 |\varphi(\xi)|^2}{|P'(\xi)|}\, d\xi \right\}.
\end{aligned} \tag{25}
$$

We apply (25) to the function $\tilde{\psi}$ defined by

$$
\widehat{\tilde{\psi}} = \hat{\psi}\, \frac{|P'(\xi)|^{1/2}}{|\xi|}\, \varphi(\xi),
$$

and we obtain Theorem 1(b) by translation in the x variable. □

4. Multidimensional case.

4.1. Technical results. Let us denote by ψ the solution to (6):

$$\begin{cases} i\dfrac{\partial \psi}{\partial t} + (-\Delta)^{\alpha}\psi = 0, & t, x \in \mathbb{R}\times\mathbb{R}^n, \\ \psi(x;0) = \psi_0(x), & x \in \mathbb{R}^n, \end{cases}$$

where $\alpha \geq 1$.

We denote by $P(\xi) = |\xi|^{2\alpha}$ the symbol of $(-\Delta)^{\alpha}$ and we compute

$$\begin{aligned} P(\eta) - P(\xi-\eta) &= |\eta|^{2\alpha} - |\xi-\eta|^{2\alpha} \\ &= \xi\cdot[2\eta-\xi]\,\frac{|\eta|^{2\alpha}-|\eta-\xi|^{2\alpha}}{|\eta|^2-|\eta-\xi|^2}. \end{aligned}$$

Let $\omega_\alpha(a;b) = a^\alpha - b^\alpha/a - b$; ω_α is a regular ($\mathcal{C}^1$) function. It follows that $P(\eta) - P(\xi-\eta) = \xi\cdot(2\eta-\xi)\,\omega_\alpha(|\eta|^2;|\eta-\xi|^2)$. We introduce

$$q(\xi;\eta) = (2\eta-\xi)\,\omega_\alpha(|\eta|^2;|\eta-\xi|^2). \tag{26}$$

We have the following lemma.

LEMMA 4. *If ξ is fixed in $\mathbb{R}^n$, then $\eta \mapsto q(\xi,\eta)$ is one-to-one from $\mathbb{R}^n$ onto $\mathbb{R}^n$.*

Proof. Let $z \in \mathbb{R}^n$, $\xi \in \mathbb{R}^n$. Let us solve the equation

$$z = (2\eta-\xi)\,\omega_\alpha(|\eta|^2\,;\,|\eta-\xi|^2). \tag{27}$$

The vector η must be in the one-dimensional affine space $\xi/2 + \mathbb{R}z$. We search η under the form $\eta = \xi/2 + \lambda z$. λ satisfies

$$2\lambda z = \frac{z}{\omega_\alpha\big(|\xi/2+\lambda z|^2;|\lambda z-\xi/2|^2\big)},$$

or equivalently

$$2\lambda = \frac{1}{\omega_\alpha\big(|\xi/2+\lambda z|^2;|\lambda z-\xi/2|^2\big)},$$

which is equivalent to

$$2\lambda = \frac{2\lambda(\xi\cdot z)}{|\xi/2+\lambda z|^{2\alpha}-|\lambda z-\xi/2|^{2\alpha}}. \tag{28}$$

- *If $\xi\cdot z \neq 0$,* (28) means

$$2(\xi\cdot z) = |\xi/2+\lambda z|^{2\alpha} - |\lambda z-\xi/2|^{2\alpha},$$

and we have to prove that $\lambda \mapsto |\xi/2+\lambda z|^{2\alpha} - |\lambda z-\xi/2|^{2\alpha}$ is one-to-one. By a change of variable, it is sufficient to prove that $\mu \mapsto (1+\mu^2+\beta\mu)^\alpha - (1+\mu^2-\beta\mu)^\alpha$ is one-to-one. This can be verified after derivation by a straightforward calculation.

- *If $\xi\cdot z = 0$,* then $2\eta\cdot\xi = |\xi|^2$ and $|\eta-\xi|^2 = |\eta|^2$, (28) becomes $z = 2\lambda\, z\,\omega_\alpha(|\eta|^2;|\eta|^2)$, and it is easy to conclude. □

For $\psi \in \mathcal{S}(\mathbb{R}^n)$, we define the generalized Wigner transform of ψ by

$$\mathrm{GWT3}(\psi) = \int e^{ix\cdot q(\xi,\eta)}\,\hat\psi(\eta)\,\overline{\hat\psi(\xi-\eta)}\,\omega_\alpha(|\eta|^2;|\xi-\eta|^2)^{n-1}\,d\eta. \tag{29}$$

If ψ solves (6), then $f(x;\xi;t) \equiv \mathrm{GWT3}(\psi(\cdot,t))$ satisfies the linear transport equation.

4.2. Proof of Theorem 3. We define an operator $\mathcal{L}$ as follows. Let $h(\xi, z) \in \mathcal{S}(\mathbb{R}^n \times \mathbb{R}^n)$ and $H(x;\xi)$ be the function given by

$$H(x,\xi) = \int e^{ix\cdot z}\, h(\xi; z)\, dz; \tag{30}$$

then

$$\mathcal{L}(H) = \int e^{ix\cdot z}\, \frac{h(\xi,z)\, x\cdot\xi\omega_\alpha(|\eta^2|;|\xi-\eta|^2)}{\big(1+\omega_\alpha(|\eta|^2;|\xi-\eta|^2)^\delta\,|x|^\delta\big)^{1/\delta}}\, dz, \tag{31}$$

where $z = q(\xi;\eta)$ and $\eta = g(z,\xi)$, i.e., g is the inverse function of $q(\cdot,\xi)$ which exists by Lemma 4.

We will apply $\mathcal{L}$ to the transport equation.

From now on, we suppose that

$$\hat{\psi}\,|\xi|^{-(\alpha-1)} \in L^2(\mathbb{R}^n). \tag{32}$$

We perform the change of variables $z = q(\xi;\eta)$ in the definition of f (see (29)):

$$f(x;\xi;t) = \int e^{ix\cdot z}\, \hat{\psi}(g(z;\xi))\, \overline{\hat{\psi}(\xi - g(z;\xi))}\, \frac{\omega_\alpha(|g|^2,|\xi-g|^2)^{n-1}}{J(z;\xi)}\, dz, \tag{33}$$

where $J(z;\xi) = \big|\det\left(\partial q/\partial\eta\right)(\xi; g(z;\xi))\big|$. Hypothesis (32) implies that (33) is well defined.

We apply $\mathcal{L}$ to the transport equation and then we integrate $\int_{-T}^{T}\iint dx\, d\xi\, dt$:

$$\int_{-T}^{T}\!\!\iint \mathcal{L}\left(\frac{\partial f}{\partial t}\right) dx\, d\xi\, dt = \int_{-T}^{T}\!\!\iint \mathcal{L}(\xi\cdot\nabla_x f)\, dx\, d\xi\, dt. \tag{34}$$

(i) *Left-hand side of* (34). First we remark that $\mathcal{L}$ commutes with $\partial/\partial t$. On the other hand, since f solves (3), $f(x;\xi;t) = f_0(x+\xi t;\xi)$. Therefore, by (33),

$$f(x;\xi;t) = \int e^{i(x+\xi t)\cdot z}\, \hat{\psi}_0(g(z;\xi))\, \frac{\overline{\hat{\psi}_0(\xi - g(z;\xi))}\, \omega_\alpha(|g|^2;|\xi-g|^2)^{n-1}}{J}\, dz.$$

Whence f is in the form (30), and by (31)

$$\mathcal{L}(f) = \int e^{i(x+\xi t)\cdot z}\, \frac{\hat{\psi}_0(g)\,\overline{\hat{\psi}_0(\xi-g)}\omega_\alpha^n\, x\cdot\xi\,\omega_\alpha}{J\big(1+\omega_\alpha(|g|^2;|\xi-g|^2)^\delta\,|x|^\delta\big)^{1/\delta}} dz.$$

Hence

$$\begin{aligned} I(T) &\equiv \int_{-T}^{T}\!\!\iint \frac{\partial\mathcal{L}(f)}{\partial t}\, dx\, d\xi\, dt, \\ &= \iiint e^{i(x+\xi T)\cdot z}\, \frac{\hat{\psi}_0(g)\,\overline{\hat{\psi}_0(\xi-g)}\omega_\alpha^n\, x\cdot\xi}{J\big(1+\omega_\alpha^\delta|x|^\delta\big)^{1/\delta}} dz\, d\xi\, dx \\ &\quad - \iiint e^{i(x-\xi T)\cdot z}\, \frac{\hat{\psi}_0(g)\,\overline{\hat{\psi}_0(\xi-g)}\omega_\alpha^n\, x\cdot\xi}{J\big(1+\omega_\alpha^\delta|x|^\delta\big)^{1/\delta}} dz\, d\xi\, dx. \end{aligned}$$

We make the changes of variables $y = x + \xi T$ (respectively, $y = x - \xi T$) in this expression and letting $T \to +\infty$, we get

$$\lim_{T\to+\infty} I(T) = -2 \iiint e^{iy\cdot z} \frac{\hat\psi_0(g)\,\overline{\hat\psi_0(\xi-g)}\,\omega_\alpha^{n-1}\,|\xi|}{J(\xi;z)}\,dz\,d\xi\,dy. \tag{35}$$

In (35) we use the formula $\int \hat h(\xi)\,d\xi = ch(0)$ and we obtain

$$\lim_{T\to+\infty} I(T) = -c\int \hat\psi_0(g(0;\xi))\,\overline{\hat\psi_0(\xi-g(0;\xi))} \cdot \frac{\omega_\alpha^{n-1}\left(|g(0;\xi)|^2;|\xi-g(0;\xi)|^2\right)|\xi|}{J(\xi;0)}\,d\xi.$$

We remark that $g(0;\xi) = \xi/2$ and $J(\xi;0) = \omega_\alpha(|\xi|^2/4;|\xi|^2/4)^n$. This leads to

$$\lim_{T\to+\infty} I(T) = -c\int |\hat\psi_0(\xi)|^2\,|\xi|^{1-2(\alpha-1)}\,d\xi. \tag{36}$$

(ii) *The right-hand side of* (34). Notice that

$$\xi\cdot\nabla_x f = \int \xi\cdot\nabla_x(e^{ix\cdot z})\,\frac{\hat\psi(g)\,\overline{\hat\psi(\xi-g)}}{J}\,\omega_\alpha^{n-1}\,dz = \int i\xi\cdot z\,e^{ix\cdot z}\,\frac{\hat\psi(g)\,\overline{\hat\psi(\xi-g)}}{J}\,\omega_\alpha^{n-1}\,dz.$$

It follows that $\xi\cdot\nabla_x f$ is under the form (30), and the definition (31) of $\mathcal{L}$ implies

$$\mathcal{L}(\xi\cdot\nabla_x f) = \int i\xi\cdot z e^{ix\cdot z}\,\hat\psi(g)\,\overline{\hat\psi(\xi-g)} \cdot \frac{\omega_\alpha(|g^2|;|\xi-g|^2)^{n-1}}{J}\,\frac{(x\omega_\alpha)\cdot\xi}{(1+\omega_\alpha^\delta|x|^\delta)^{1/\delta}}\,dz.$$

This leads to

$$II(T) \equiv \int_{-T}^{T}\iint \mathcal{L}(\xi\cdot\nabla_x f)\,dx\,d\xi\,dt = \int_{-T}^{T}\iiint (\xi\cdot\nabla_x)(e^{ix\cdot z})\,\frac{\hat\psi(g)\,\overline{\hat\psi(\xi-g)}\,\omega_\alpha^{n-1}\,x\cdot\xi\,\omega_\alpha}{J(1+\omega_\alpha^\delta|x|^\delta)^{1/\delta}}\,dz\,d\xi\,dx\,dt.$$

We integrate by parts in the x variable:

$$II(T) = -\int_{-T}^{T}\iiint \xi_j\,\frac{\hat\psi(g)\,\overline{\hat\psi(\xi-g)}}{J}\,e^{ix\cdot z}\,\omega_\alpha^{n-1} \cdot\frac{\partial}{\partial x_j}\left[\frac{x\cdot\xi\omega_\alpha}{(1+\omega_\alpha^\delta|x|^\delta)^{1/\delta}}\right]dz\,d\xi\,dx\,dt$$

$$= -\int_{-T}^{T}\iiint \hat\psi(g)\,\frac{\overline{\hat\psi(\xi-g)}}{J}\,e^{ix\cdot z}\,\omega_\alpha^{n-1}\left[\frac{\xi^2\omega_\alpha}{(1+\omega_\alpha^\delta|x|^\delta)^{1/\delta}} - \frac{(\xi\cdot x)^2\,|x|^{\delta-2}\,\omega_\alpha^{\delta+1}}{(1+\omega_\alpha^\delta|x|^\delta)^{1+1/\delta}}\right]dz\,d\xi\,dx\,dt.$$

In order to carry on the computation, we return to the variable η:

$$II(T) = -\int_{-T}^{T} \iiint \hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,e^{ix\cdot q(\xi;\eta)}\,\omega_\alpha^n \left[\frac{\xi^2}{(1+\omega_\alpha^\delta|x|^\delta)^{1/\delta}} - \frac{(\xi\cdot x)^2\,|x|^{\delta-2}\,\omega_\alpha^\delta}{(1+\omega_\alpha^\delta|x|^\delta)^{1+1/\delta}}\right] d\eta\,d\xi\,dx\,dt.$$

Recall that $q(\xi;\eta) = (2\eta-\xi)\,\omega_\alpha(|\eta|^2;|\xi-\eta|^2)$ and set $y = x\,\omega_\alpha(|\eta|^2;|\xi-\eta|^2)$. $II(T)$ becomes

$$II(T) = -\int_{-T}^{T} \iiint \hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,e^{iy\cdot(2\eta-\xi)} \left[\frac{\xi^2}{(1+|y|^\delta)^{1/\delta}} - \frac{(\xi\cdot y)^2\,|y|^{\delta-2}}{(1+|y|^\delta)^{1+1/\delta}}\right] d\eta\,d\xi\,dy\,dt. \tag{37}$$

At this stage, we are in the same situation as in [5]. Let us recall how to conclude. We have

$$\frac{\xi^2}{(1+|y|^\delta)^{1/\delta}} - \frac{(\xi\cdot y)^2\,|y|^{\delta-2}}{(1+|y|^\delta)^{1+1/\delta}} = \xi_i\xi_j\,\frac{\partial}{\partial y_j}\left(\frac{y_i}{(1+|y|^\delta)^{1/\delta}}\right).$$

We write $\xi_i = (\xi_i-\eta_i)+\eta_i$ in (37) and we obtain

$$II(T) = c\int_{-T}^{T}\int \left[\frac{\overline{\partial^2\psi}}{\partial y_i\,\partial y_j}\,\psi(y) - \frac{\partial\psi}{\partial y_j}\,\frac{\partial\overline{\psi}}{\partial y_i} - \frac{\partial\psi}{\partial y_i}\,\frac{\partial\overline{\psi}}{\partial y_j} + \frac{\partial^2\psi}{\partial y_i\,\partial y_j}\,\overline{\psi}\right] \cdot\frac{\partial}{\partial y_j}\left(\frac{y_i}{(1+|y|^\delta)^{1/\delta}}\right) dy\,dt,$$

$$= c\int_{-T}^{T}\int \left(\frac{\partial^2|\psi|^2}{\partial y_i\partial y_j} - 2\frac{\partial\psi}{\partial y_j}\,\frac{\partial\overline{\psi}}{\partial y_i} - 2\frac{\partial\psi}{\partial y_i}\,\frac{\partial\overline{\psi}}{\partial y_j}\right)\frac{\partial}{\partial y_j}\left(\frac{y_i}{(1+|y|^\delta)^{1/\delta}}\right) dy dt,$$

$II(T)$

$$= -2c\int_{-T}^{T}\left\{\left(\frac{\partial\psi}{\partial y_j}\,\frac{\partial\overline{\psi}}{\partial y_i} + \frac{\partial\psi}{\partial y_i}\,\frac{\partial\overline{\psi}}{\partial y_j}\right)\left[\frac{\delta_{ij}}{(1+|y|^\delta)^{1/\delta}} - \frac{y_iy_j|y|^{\delta-2}}{(1+|y|^\delta)^{1+1/\delta}}\right] dy\right.$$

$$\left. -c\int_{-T}^{T}\int |\psi|^2\,\frac{\partial^3}{\partial y_i\,\partial y_j\,\partial y_j}\left(\frac{y_i}{(1+|y|^\delta)^{1/\delta}}\right) dy\right\} dt. \tag{38}$$

The equalities (34), (36), and (38) give

$$c\int |\hat{\psi}_0(\xi)|^2|\xi|^{1-2(\alpha-1)}\,d\xi$$

$$= 4\int_{-\infty}^{+\infty}\left\{\int\left(\frac{|\nabla\psi|^2}{(1+|y|^\delta)^{1/\delta}} - \frac{(y\cdot\nabla\psi)^2|y|^{\delta-2}}{(1+|y|^\delta)^{1+1/\delta}}\right) dy + \frac{1}{2}\int|\psi|^2\,\Delta\left(\frac{\partial}{\partial y_i}\left(\frac{y_i}{(1+|y|^\delta)^{1/\delta}}\right)\right) dy\right\} dt. \tag{39}$$

Now we remark that the Cauchy–Schwarz inequality yields, for any v in $\mathbb{R}^n$,

$$\frac{v^2}{(1+|y|^\delta)^{1/\delta}} - \frac{(y\cdot v)^2\,|y|^{\delta-2}}{(1+|y|^\delta)^{1+1/\delta}} \geq v^2\left[\frac{1}{(1+|y|^\delta)^{1/\delta}} - \frac{|y|^\delta}{(1+|y|^\delta)^{1+1/\delta}},\right]$$
$$= \frac{v^2}{(1+|y|^\delta)^{1+1/\delta}}.$$

On the other hand, we have the following lemma.

LEMMA 5. *If $\delta \leq 1$, then*

$$\Delta \operatorname{div}\left(\frac{y}{(1+|y|^\delta)^{1/\delta}}\right) \geq 0.$$

Assuming this lemma, (39) gives

$$\int_{-\infty}^{+\infty}\int \frac{|\nabla\psi|^2}{(1+|y|^\delta)^{1+1/\delta}}\,dy \leq c\int |\hat{\psi}_0(\xi)|^2\,|\xi|^{1-2(\alpha-1)}\,d\xi.$$

We apply this equality to the function $\tilde{\psi}$ defined by

$$\widehat{\tilde{\psi}}(\xi) = |\xi|^{\alpha-1}\,\hat{\psi}(\xi)$$

(in particular, hypothesis (32) is satisfied). We get

$$\int_{-\infty}^{+\infty}\int \frac{|\nabla(|D|^{\alpha-1}\psi)|^2}{(1+|y|^\delta)^{1+1/\delta}}\,dy\,dt \leq c|\psi_0|^2_{\dot{H}^{1/2}}. \qquad \square$$

We still must prove Lemma 5. One has

$$\operatorname{div}\frac{y}{(1+|y|^\delta)^{1/\delta}} = \frac{n+(n-1)|y|^\delta}{(1+|y|^\delta)^{1+1/\delta}}.$$

We introduce

$$k(v) = \frac{n+(n-1)u}{(1+u)^{1+1/\delta}}.$$

Then, one computes

$$k'(u) = \frac{n-1}{(1+u)^{1+1/\delta}} - \frac{(n+(n-1)u)}{(1+u)^{2+1/\delta}}\,(1+1/\delta)$$
$$= \frac{-(1+n/\delta) - \frac{(n-1)u}{\delta}}{(1+u)^{2+1/\delta}} < 0 \quad \text{for} \quad u \geq 0,$$
$$k''(u) = [1+1/\delta]\left(\frac{(2+n/\delta)+\frac{n-1}{\delta}\,u}{(1+u)^{3+1/\delta}}\right) > 0 \quad \text{for} \quad u \geq 0.$$

Hence k is convex and decreasing. Finally, since $y \mapsto |y|^\delta$ is concave for $0 < \delta \leq 1$, the mapping

$$y \mapsto \operatorname{div}\frac{y}{(1+|y|^\delta)^{1/\delta}}$$

is convex and the lemma follows. $\square$

4.3. Proof of Theorem 2. In order to prove Theorem 2, we proceed as in the previous section, but instead of applying $\mathcal{L}$ in (3), we multiply (3) by $(x\cdot\xi)/|x|$ and we integrate over $\mathbb{R}^n\times\mathbb{R}^n$ with respect to x and ξ and on $[-T;T]$ with respect to t. We get

$$\begin{aligned} I(T) &\equiv \iint f(x;\xi;T)\,\frac{x\cdot\xi}{|x|}\,dx\,d\xi - \iint f(x;\xi;-T)\,\frac{x\cdot\xi}{|x|}\,dx\,d\xi,\\ &= \int_{-T}^{T}\iint \xi\cdot\nabla_x f\,\frac{x\cdot\xi}{|x|}\,dx\,d\xi\,dt \equiv II(T). \end{aligned} \tag{40}$$

As previously, one has

$$f(x;\xi;T) = f_0(x+\xi T,\xi).$$

From now on, we suppose

$$\hat{\psi}_0(\xi)|\xi|^{-(\alpha-1)} \in L^2(\mathbb{R}^n). \tag{41}$$

(i) *Limit of $I(T)$ as $T\to+\infty$.* Making the changes of variables $y = x+\xi T$, $y = x-\xi T$ in each term of the expression of $I(T)$ in (40) and letting $T\to+\infty$ lead to

$$I(T) \underset{T\to+\infty}{\to} -2\iint |\xi|\,f_0(z;\xi)\,dz\,d\xi. \tag{42}$$

We argue as in §4.2(i), and we get

$$\lim_{T\to+\infty} I(T) = c_\alpha \int |\hat{\psi}_0(\xi)|^2\,|\xi|^{1-2(\alpha-1)}\,d\xi. \tag{43}$$

(ii) *Limit of $II(T)$ as $T\to+\infty$.* An integration by parts in $II(T)$ with respect to x yields

$$II(T) = -\int_{-T}^{T}\iint f\,\xi_i\xi_j\,\frac{\partial}{\partial x_i}\left(\frac{x_j}{|x|}\right)dx\,d\xi\,dt$$

and

$$\frac{\partial}{\partial x_i}\left(\frac{x_j}{|x|}\right) = \frac{\delta_{ij}}{|x|} - \frac{x_i\,x_j}{|x|^3},$$

and we get

$$\begin{aligned} II(T) = \int_{-T}^{T}\iiint \xi_i\xi_j\left(\frac{\delta_{ij}}{|x|} - \frac{x_i\,x_j}{|x|^3}\right)e^{ix\cdot q(\xi;\eta)}\\ \cdot\hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,\omega_\alpha^{n-1}\,dx\,d\eta\,d\xi\,dt. \end{aligned}$$

Now since $q(\xi;\eta) = (2\eta-\xi)\,\omega_\alpha(|\eta|^2;|\xi-\eta|^2)$, we perform the following change of variables:

$$y = x\,\omega_\alpha(|\eta|^2;|\xi-\eta|^2),$$

and we get

$$\begin{aligned} II(T) = -\int_{-T}^{T}\iiint \xi_i\xi_j\left(\frac{\delta_{ij}}{|y|} - \frac{y_i\,y_j}{|y|^3}\right)e^{iy[2\eta-\xi]}\\ \cdot\hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,d\eta\,d\xi\,dy\,dt. \end{aligned} \tag{44}$$

We now conclude as in [5].

(44), (43), and (40) applied to the function $\tilde{\psi}$ defined by

$$\widehat{\tilde{\psi}} = |\xi|^{(\alpha-1)}\,\hat{\psi}$$

(in particular, $\tilde{\psi}$ satisfies (41)) imply the theorem for $x_0 = 0$. A translation on the initial data gives the theorem for any $x_0 \in \mathbb{R}^N$. □

4.4. Proof of Theorem 4. Let ψ be a solution of (7)

$$\begin{cases} i\dfrac{\partial \psi}{\partial t} + \displaystyle\sum_{j=1}^{n} P_j(D_j)\psi = 0 \quad \text{in} \quad \mathbb{R}\times\mathbb{R}^n, \\ \psi(x;0) = \psi_0(x). \end{cases}$$

We define g_j as in (8), i.e.,

$$g_j(\xi_j;\eta_j) = -\frac{P_j(\eta_j) - P_j(\xi_j - \eta_j)}{\xi_j(2\eta_j - \xi_j)}, \tag{45}$$

and we denote by $q_j(\xi_j;\eta_j)$ the following expression:

$$q_j(\xi_j;\eta_j) = (2\eta_j - \xi_j)\, g_j(\xi_j;\eta_j). \tag{46}$$

We now introduce the generalized Wigner transform of ψ by

$$\begin{aligned} f(x;\xi;t) &= \mathrm{GWT4}(\psi) \\ &\equiv \int e^{ix\cdot q(\xi;\eta)}\,\hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,\prod_{j\geq 2}|g_j(\xi_j;\eta_j)|\,d\eta, \end{aligned}$$

where $q(\xi;\eta)$ is the vector formed by $(q_j(\xi_j;\eta_j))_{j=1 \text{ to } n}$. Then f solves the linear transport equation (3)

$$\frac{\partial f}{\partial t} - \xi\cdot\nabla_x f = 0.$$

As in the proof of Theorem 1, we multiply (3) by $x_1\xi_1/|x_1|$ and we integrate $\int_{-T}^{T}\iint dx\,d\xi\,dt$. We get

$$\begin{aligned} I(T) &\equiv \iint f(x;\xi;T)\,\frac{x_1\xi_1}{|x_1|}\,dx\,d\xi - \iint f(x;\xi;-T)\,\frac{x_1\xi_1}{|x_1|}\,dx\,d\xi \\ &= \int_{-T}^{T}\iint \xi_j\,\frac{\partial f}{\partial x_j}\,\frac{x_1\xi_1}{|x_1|}\,dx\,d\xi\,dt \equiv II(T). \end{aligned} \tag{47}$$

- *Calculation of* $I(T)$. We impose that

$$|\xi_1|\,\frac{|\hat{\psi}(\xi)|^2}{|P_1'(\xi_1)|^{1/2}} \in L^1(\mathbb{R}^n). \tag{48}$$

Since f satisfies (1), $f(x;\xi;T) = f_0(x+\xi T;\xi)$. Therefore,

$$I(T) = \iint f_0(x+\xi T;\xi)\,\frac{x_1\xi_1}{|x_1|}\,dx\,d\xi - \iint f_0(x-\xi T;\xi)\,\frac{x_1\xi_1}{|x_1|}\,dx\,d\xi.$$

Using the change of variable $y = x+\xi T$ (respectively, $y = x-\xi T$) and letting $T \to +\infty$ leads to

$$\lim_{T\to+\infty} I(T) = -2 \iint f_0(y;\xi)\,|\xi_1|\,dx\,d\xi$$
$$= -2 \iiint e^{iy\cdot q}|\xi_1|\,\hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)} \prod_{j\geq 2} |g_j|\,d\eta\,d\xi\,dy.$$

We now make the change of variables $z = q(\xi;\eta)$ and we denote by $g(\xi;z)$ the inverse function. Thus we obtain

$$\lim_{T\to+\infty} I(T) = -2 \iiint e^{iy\cdot z}|\xi_1|\,\hat{\psi}(g(\xi;z))\,\overline{\hat{\psi}(\xi-g(\xi;z))}$$
$$\cdot \frac{\prod_{j\geq 2}|g_j|}{J}\,dz\,d\xi\,dy,$$

where J is the Jacobian of the transformation. Making use of the formula $\int \hat{h}(\xi)\,d\xi = ch(0)$ in the above expression of $\lim I(T)$ leads to

$$\lim_{T\to+\infty} I(T) = -2c \int |\xi_1|\,\frac{\hat{\psi}(\xi/2)\,\overline{\hat{\psi}(\xi/2)}}{|g_1(\xi;\xi/2)|}\,d\xi,$$

or equivalently,

$$\lim_{T\to+\infty} I(T) = -c \int \frac{|\xi_1|^2}{|P_1'(\xi_1/2)|}\,|\hat{\psi}(\xi/2)|^2\,d\xi. \tag{49}$$

• *Calculation of* $II(T)$. We integrate by parts with respect to the x variable in the expression that defines $II(T)$:

$$II(T) = -\int_{-T}^{T} \iint \xi_1^2\, f(x_1 = 0\,,\,\xi;t)\,dx_2\ldots dx_n\,d\xi\,dt$$
$$= -\int_{-T}^{T} \iiint \xi_1^2\,\hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,e^{i\begin{pmatrix}0\\x_2\\\vdots\\x_n\end{pmatrix}\cdot q} \prod_{j\geq 2}|g_j|\,dx_2\ldots dx_n\,d\xi\,d\eta\,dt.$$

For $j \geq 2$, let $y_j = x_j\,g_j(\xi_j;\eta_j)$; thus we get

$$II(T) = -\int_{-T}^{T} \iiint \xi_1^2\,\hat{\psi}(\eta)\,\overline{\hat{\psi}(\xi-\eta)}\,e^{i\begin{pmatrix}0\\x_2\\\vdots\\x_n\end{pmatrix}\cdot(2\eta-\xi)} dy_2\ldots dy_n\,d\xi\,d\eta\,dt.$$

Let us remark that $\xi_1^2 = (\xi_1 - 2\eta_1)^2 + 4(\xi_1-\eta_1)\eta_1$. After the change of variables

$$\begin{cases} z = 2\eta - \xi, \\ \eta = \eta, \end{cases}$$

$II(T)$ becomes

$$II(T) = -\int_{-T}^{T} \iiint z_1^2\, \hat{\psi}(\eta)\, \overline{\hat{\psi}(\eta - z)}\, e^{i \begin{pmatrix} 0 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \cdot z}\, dy_2 \dots dy_n\, dz\, d\eta\, dt$$

$$- 4 \int_{-T}^{T} \iiint (\xi_1 - \eta_1)\eta_1\, \hat{\psi}(\eta)\, \overline{\hat{\psi}(\xi - \eta)}\, e^{i \begin{pmatrix} 0 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \cdot (2\eta - \xi)}\, dy_2 \dots dy_n\, d\xi\, d\eta\, dt$$

$$\equiv II_1(T) + II_2(T).$$

In order to estimate $II_1(T)$, we let

$$\tilde{f}(x;\xi;t) = \int e^{ix\cdot q(\xi;\eta)}\, \hat{\psi}(\eta)\, \overline{\hat{\psi}(\eta - \xi)} \prod_{j\geq 2} |g_j(\xi_j;\eta_j)|\, d\eta.$$

$\tilde{f}$ satisfies (1), and proceeding in the same way as in the beginning of this section, we obtain

(50)

$$- c \int \frac{|\xi_1|^2}{|P_1'(\xi_1/2)|}\, \hat{\psi}(\xi/2)\, \overline{\hat{\psi}(-\xi/2)}\, d\xi$$

$$= - \int_{-T}^{T} \iiint |\xi_1|^2\, \hat{\psi}(\eta)\, \overline{\hat{\psi}(\xi - \eta)}\, e^{i \begin{pmatrix} 0 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \cdot q} \prod_{j\geq 2} |g_j|\, dx_2 \dots dx_n\, d\xi\, d\eta\, dt.$$

On the other hand,

$$II_2(T) = -c \int_{-T}^{T} \int \left| \frac{\partial \psi}{\partial y_1} \right|^2 (0, y_2, \dots, y_n, t)\, dy_2 \dots dy_n\, dt. \tag{51}$$

Equations (50), (51), (49), and (47) lead to

$$\int_{-\infty}^{+\infty} \int \left| \frac{\partial \psi}{\partial y_1} \right|^2 (0; y_2, \dots, y_n)\, dy_2 \dots dy_n\, dt$$

$$= c \left(\int \frac{|\xi_1|^2\, |\hat{\psi}(\xi/2)|^2}{|P_1'(\xi_1/2)|}\, d\xi - \int \frac{|\xi_1|^2}{|P_1'(\xi_1/2)|}\, \hat{\psi}(\xi/2)\, \overline{\hat{\psi}(-\xi/2)}\, d\xi \right).$$

Applying this formula to the function defined by

$$\widehat{\tilde{\psi}}(\xi) = \frac{|P_1'(\xi_1)|^{1/2}}{|\xi_1|}\, \hat{\psi}(\xi)$$

(in particular, (48) is satisfied) leads to Theorem 4 with $y_1 = 0$. The general case follows by the invariance under the translations. □

REFERENCES

[1] M. BEN-ARTZI AND A. DEVINATZ, *Local smoothing and convergence properties of Schrödinger type equations*, J. Funct. Anal., 101 (1991), pp. 231–234.

[2] T. COLIN, *Effets régularisants pour des équations dispersives obtenus par une transformée de Wigner généralisée*, C. R. Acad. Sci. Paris, 317 (1993), pp. 673–676.

[3] P. CONSTANTIN AND J.-C. SAUT, *Local smoothing properties of dispersive equations*, J. Amer. Math. Soc., 1 (1988), pp. 413–446.

[4] C. KENIG, G. PONCE, AND L. VEGA, *Oscillatory integrals and regularity of dispersive equations*, Indiana Univ. Math. J., 40 (1991), pp. 33–69.

[5] P. L. LIONS AND B. PERTHAME, *Lemme de moments, de moyenne et de dispersion*, C.R. Acad. Sci. Paris I, t. 314 (1992), pp. 801–806.

SIAM J. MATH. ANAL.
Vol. 25, No. 6, pp. 1642–1655, November 1994

012

CHARACTERIZATION OF THE SMOOTHEST INTERPOLANT*

BORISLAV BOJANOV†

Abstract. This paper characterizes the smoothest function φ from $W_p^r[a,b]$ of a given shape. The shape is prescribed by the condition that φ should take given values $\{f_i\}_0^{n+1}$ consecutively on $[a,b]$ and that its derivatives up to certain orders $\{\nu_i\}_0^{n+1}$, respectively, have to vanish at the interpolation points.

Key words. interpolation, perfect spline, interpolant, Sobolev space

AMS subject classifications. 41A05, 65D10

1. Introduction. This paper is devoted to an extremal problem studied by A. Pinkus in [10]. For given $[a,b]$ and interpolation values $\{f_k\}_0^{N+1}$ satisfying the conditions

$$(f_{k+1}-f_k)(f_k-f_{k-1})<0, \quad k=1,\dots,N, \tag{1}$$

he characterized the function φ from the Sobolev space

$$W_p^r[a,b] := \{f \in AC^{r-1}[a,b] : \| f^{(r)} \|_p < \infty\},$$

which takes on the values $f_0, f_1, \dots, f_{N+1}$ consecutively on $[a,b]$ at some points $a = x_0 < x_1 < \cdots < x_{N+1} = b$ (free to vary) and has a minimal L_p-norm of its rth derivative. We consider here a generalization of this problem allowing equalities in (1). This leads to certain osculatory conditions on the smoothest interpolant. In addition, we extend the study to a set of classes (defined by a condition of the form $| f^{(r)}(t) | \le \sigma(t)$ on $[a,b]$) that includes $W_\infty^r[a,b]$ and can be used to obtain a generalization of the Pinkus result for $W_p^r[a,b], 1<p<\infty$, in an indirect, simple way.

Let us state the problem.

Suppose that Ω is a given subspace of $AC^{r-1}[a,b]$ supplied with a certain seminorm $\| f^{(r)} \|$. Let $\nu_0, \nu_1, \dots, \nu_{n+1}$ be fixed natural numbers not exceeding r. Suppose further that $\mathbf{y} = \{y_k\}_0^{n+1}$ are given values such that

$$(-1)^{\nu_k}(y_{k+1}-y_k)(y_k-y_{k-1})>0, \quad k=1,\dots,n.$$

For every set of points $\mathbf{t} = \{t_k\}_0^{n+1}, a = t_0 < t_1 < \cdots < t_{n+1} = b$, $F(\mathbf{t},\mathbf{y})$ will denote the class of all functions f from Ω that satisfy the interpolation conditions

$$\begin{aligned} f(t_k) &= y_k, \quad k=0,\dots,n+1, \\ f^{(j)}(t_k) &= 0, \quad k=0,\dots,n+1,\ j=1,\dots,\nu_k-1. \end{aligned}$$

The aim of this paper is to describe the extremal function in the problem

$$\inf_{a=t_0<t_1<\cdots<t_{n+1}=b} \ \inf_{f\in F(\mathbf{t},\mathbf{y})} \| f^{(r)} \| \tag{2}$$

with respect to certain norms, including L_p, $1<p\le\infty$.

* Received by the editors October 1, 1992; accepted for publication (in revised form) July 19, 1993. This work was supported by grant MM-15 of the Bulgarian Ministry of Sciences.

† Department of Mathematics, University of Sofia, Blvd. J. Boucher 5, 1126 Sofia, Bulgaria (bor@bgearn.bitnet).

2. σ-perfect splines. Assume that $\sigma(t)$ is a nonnegative function from $L_1[a,b]$. A *σ-perfect spline* on $[a,b]$ of degree r with knots $\bar{\xi} = \{\xi_i\}_1^n$ ($a =: \xi_0 < \xi_1 < \cdots < \xi_n < \xi_{n+1} := b$) is any expression of the form

$$p(x) = \sum_{i=1}^{r} a_i x^{i-1} + d \int_a^b \frac{(x-t)_+^{r-1}}{(r-1)} \psi(\bar{\xi}; t)\, dt$$

where $\{a_i\}$ and d are real parameters and

$$\psi(\bar{\xi}; t) := (-1)^{n-j}\sigma(t) \quad \text{for } \xi_j < t < \xi_{j+1},$$

$j = 0, 1, \ldots, n$. Functions of this kind were introduced and used in [2], [4] to solve certain extremal problems.

The function $\sigma(t)$ gives rise to the Minkowski seminorm

$$\| f^{(r)} \|_\sigma := \inf\{M : |f^{(r)}(t)| \le M\sigma(t) \text{ a.e. on } [a,b]\}$$

in $AC^{r-1}[a,b]$. The σ-perfect splines are an important tool for the study of (2) in the space

$$\Omega := \{f \in AC^{r-1}[a,b] : \| f^{(r)} \|_\sigma < \infty\} \tag{3}$$

with respect to $\| \cdot \| = \| \cdot \|_\sigma$. They preserve most of the properties of the ordinary polynomial perfect splines. We recall below some of them, which will be used in the proof of our main result.

THEOREM A. *Suppose that $\sigma(t)$ is an integrable nonnegative function that does not vanish on subintervals. For any given set of multiplicities $\{\nu_k\}_0^{n+1}$, $1 \le \nu_k \le r$, points $\mathbf{x} = \{x_k\}_0^{n+1}$, $a = x_0 < x_1 < \cdots < x_{n+1} = b$, and values $\mathbf{y} = \{y_{ij},\ i = 0, \ldots, n+1,\ j = 0, \ldots, \nu_i - 1\}$, there exists a σ-perfect spline p of degree r with no more than $N - r - 1$ knots ($N := \nu_0 + \cdots + \nu_{n+1}$) that satisfies the interpolation conditions*

$$p^{(j)}(x_i) = y_{ij}, \quad i = 0, \ldots, n+1,\ j = 0, \ldots, \nu_i - 1. \tag{4}$$

Moreover, every such σ-perfect spline p possesses the extremal property

$$\| p^{(r)} \|_\sigma = \inf \| f^{(r)} \|_\sigma$$

over the set of all functions $f \in AC^{r-1}[a,b]$ satisfying (4).

The proof, even of a more general statement including Birkhoff's interpolation conditions, can be found in [4]. Notice here that Theorem A does not hold in the case when $\sigma(t)$ vanishes on subintervals. Indeed, assume that $\sigma(t) \equiv 0$ on $[\alpha, \beta] \subset [a,b]$ and $\alpha < x_k < \cdots < x_{k+r} < \beta$. Let $\nu_k = \cdots = \nu_{k+r} = 1$ and $y_{k+i} = (-1)^i$, $i = 0, \ldots, r$. Suppose that there is a solution p of the corresponding interpolation problem in Theorem A. Since $\sigma(t) \equiv 0$ on (α, β), p coincides with a polynomial of degree $r-1$ on (α, β). On the other hand, because of the choice of $y_k, \ldots, y_{k+r}$, p should change sign at least r times on (α, β), a contradiction.

The next theorem, the so-called Fundamental Theorem of Algebra for σ-perfect splines, can be derived as a particular case of Theorem A (see [4]).

THEOREM B. *Suppose that $\sigma(t)$ is an integrable nonnegative function that does not vanish on subintervals. Given the points $\mathbf{t} = \{\mathbf{t}_i\}_0^{n+1}$, $a = t_0 < t_1 < \cdots < t_{n+1} = b$, and the multiplicities $\lambda_0, \lambda_1, \ldots, \lambda_{n+1}$ such that*

$$0 \le \lambda_0 \le r,\ 0 \le \lambda_{n+1} \le r,\ 1 \le \lambda_k \le r,\ k = 1, \ldots, n, \quad N := \lambda_0 + \lambda_1 + \ldots + \lambda_{n+1},$$

there exists a unique (up to multiplication by -1) σ-perfect spline $p(\mathbf{t};x)$ of degree r with no more than $N-r$ knots, which satisfies the conditions

$$p^{(j)}(\mathbf{t};t_i)=0, \quad i=0,\dots,n+1,\ j=0,\dots,\lambda_i-1,$$

(no condition is imposed if $\lambda_i=0$), and

$$|p^{(r)}(\mathbf{t};t)|=\sigma(t) \text{ a.e. on } [a,b].$$

Moreover, p has exactly $N-r$ knots.

3. Main result. We extend here a result from [1] concerning polynomial perfect splines and apply it to get the complete characterization of the extremal function of the problem (2) for Ω, defined as in (3), with $\|\cdot\|=\|\cdot\|_\sigma$.

Let $[a,b]$ be a given finite interval. The multiplicities $\lambda_0,\dots,\lambda_{n+1}$ and r will stay fixed,

$$N:=\lambda_0+\cdots+\lambda_{n+1}\geq r, \quad 1\leq\lambda_k<r, \quad k=1,\dots,n, \quad 0\leq\lambda_0\leq r, \quad 0\leq\lambda_{n+1}\leq r.$$

Denote by $\mathcal{P}_r(\lambda_0,\dots,\lambda_{n+1})$ the set of all σ-perfect splines p of degree r with $N-r$ knots, which satisfy the boundary conditions $p^{(j)}(a)=0$, $j=0,\dots,\lambda_0-1$, $p^{(j)}(b)=0$, $j=0,\dots,\lambda_{n+1}-1$, and have n freely chosen zeros $t_1<\cdots<t_n$ in (a,b) of multiplicities $\lambda_1,\dots,\lambda_n$, respectively. Assume that they are normalized by the conditions $p(t)>0$ on (t_n,b) and

$$|p^{(r)}(t)|=\sigma(t) \text{ a.e. on } [a,b].$$

According to Theorem B, p is defined uniquely by its zeros $t_1,\dots,t_n$.

THEOREM 3.1. *Suppose that $\sigma(t)$ is a continuous, positive function on $[a,b]$. Then for any given set of numbers $e_0>0,\dots,e_n>0$, there exists a unique σ-perfect spline p from $\mathcal{P}_r(\lambda_0,\dots,\lambda_{n+1})$ and a constant $c>0$ such that*

$$c\left|\int_{t_k}^{t_{k+1}}p(t)\,dt\right|=e_k, \quad k=0,\dots,n, \tag{5}$$

where $\{t_i\}_1^n$ are the interior zeros of p, $a=:t_0<t_1<\cdots<t_{n+1}:=b$. Moreover, c is a continuous, strictly increasing function of $e_0,e_1,\dots,e_n$ in the domain $e_0>0,\dots,e_n>0$.

Proof. In order to prove the existence of a particular p that satisfies (5) we start from an arbitrary $p_0\in\mathcal{P}_r(\lambda_0,\dots,\lambda_{n+1})$ and $c^0=1$ and get p by a continuous deformation of p_0 described by a system of nonlinear equations. The parameters $\{a_i^0\}_1^r,\{\xi_i^0\}_1^{N-r}$ of $p_0(t)$ and $c^0=1$ are taken as initial conditions. Let $a=t_0^0<t_1^0<\cdots<t_n^0<t_{n+1}^0=b$ be the zeros of p_0. Denote

$$e_k^0:=\left|\int_{t_k^0}^{t_{k+1}^0}p_0(t)\,dt\right|, \quad k=0,\dots,n.$$

For every $s\in[0,1]$ define the quantities

$$e_k(s):=e_k^0+s(e_k-e_k^0), \quad k=0,\dots,n.$$

Clearly $e_k(0) = e_k^0$, $e_k(1) = e_k$ and hence $e_k(s) > 0$ on $[0,1]$. Our goal is to construct a σ-perfect spline $p(s;t) \in \mathcal{P}_r(\lambda_0, \dots, \lambda_{n+1})$ with parameters $a_i(s), \xi_i(s), t_i(s)$ and a function $c(s)$ such that

$$(6) \qquad \begin{cases} p^{(j)}(s; t_i(s)) = 0, \quad i = 1, \dots, n,\ j = 0, \dots, \lambda_i - 1, \\ c(s) \int_{t_k(s)}^{t_{k+1}(s)} p(s;\tau_k)\, d\tau_k - (-1)^{l_k} e_k(s) = 0, \quad k = 0, \dots, n, \end{cases}$$

where $l_n := 0$, $l_k := \lambda_{k+1} + \cdots + \lambda_n$, $k = 0, \dots, n-1$. Evidently system (6) has a solution $p(s;t) = p_0(t)$, $c(s) = 1$ for $s = 0$. The case $s = 1$ corresponds to the problem stated in Theorem 3.1.

System (6) consists of $N + n + 1$ equations in unknowns

$$\mathbf{t} := (t_1, \dots, t_n, c, a_1, \dots, a_r, \xi_1, \dots, \xi_{N-r}).$$

Denote by $J(s)$ the Jacobian matrix of (6) with respect to $\mathbf{t}$. Assume for the sake of convenience, that the equations in (6) are ordered in the following manner:

$$p^{(\lambda_i - 1)},\ i = 1, \dots, n, \quad p(a), p'(a), \dots, p^{(\lambda_0 - 1)}(a),\ c \int_a^{t_1} p\, dt,$$

$$p(t_i), p'(t_i), \dots, p^{(\lambda_i - 2)}(t_i),\ c \int_{t_i}^{t_{i+1}} p\, d\tau_i, \quad i = 1, \dots, n,$$

$$p(b), p'(b), \dots, p^{(\lambda_{n+1} - 1)}(b).$$

Here only the characterization parts of the equations are mentioned. Then $J(s)$ has the form

	$t_1 \ \dots \ t_n$	c	$a_1 \ \dots \ a_r \ \xi_1 \ \dots \ \xi_{N-r}$
1 $\vdots$ n	D		
$n+1$ $\vdots$ $n+\lambda_0$	O	0 $\vdots$ 0	
$n+\lambda_0+1$ $\vdots$ $n+\lambda_0+\lambda_1+1$ $\vdots$ $n+\sum_0^n \lambda_i + 1$		$\int_a^{t_1} p\, d\tau_0$ $\vdots$ $\int_{t_1}^{t_2} p\, d\tau_1$ $\vdots$ $\int_{t_n}^{b} p\, d\tau_n$	
$\vdots$ $n+N+1$		$\vdots$ 0	

where the block marked by O consists of zero elements and

$$D := \operatorname{diag}\{p^{(\lambda_1)}(s; t_1(s)), \dots, p^{(\lambda_n)}(s; t_n(s))\}.$$

Therefore

$$\det J(s) = \prod_{k=1}^{n} p^{(\lambda_k)}(s; t_k(s)).\det \Delta(s),$$

where the matrix $\Delta(s)$ is obtained from $J(s)$ by deletion of the first n rows and columns. Now unfolding $\det \Delta(s)$ along the elements of the first column we get

$$\det \Delta(s) = \sum_{i=0}^{n} (-1)^{\lambda_0+\cdots+\lambda_i} \int_{t_i(s)}^{t_{i+1}(s)} p(s;\tau_i)\, d\tau_i.\det \Delta_i(s),$$

where $\det \Delta_i(s)$ is the adjoint minor of the $(\lambda_0 + \cdots + \lambda_i + 1, 1)$-element of $\Delta(s)$. In order to describe $\Delta_i(s)$ precisely we need an additional notation.

For any set of points $y_1 \le \cdots \le y_N$ and smooth functions $u_1(t), \ldots, u_N(t)$ we define

$$\begin{bmatrix} u_1(t), & \ldots & , u_N(t) \\ y_1, & \ldots & , y_N \end{bmatrix} := \det \begin{bmatrix} u_1(y_1) & \cdots & u_N(y_1) \\ u_1(y_2) & \cdots & u_N(y_2) \\ . & \cdots & . \\ u_1(y_N) & \cdots & u_N(y_N) \end{bmatrix}$$

subject to the usual convention that for coincident y's the repeated rows are replaced by appropriate derivatives of $u_1, \ldots, u_N$. Let us choose

$$\{u_1(t), \ldots, u_N(t)\} \equiv \{1, t, \ldots, t^{r-1}, (\xi_1 - t)_+^{r-1}, \ldots, (\xi_{N-r} - t)_+^{r-1}\}$$

and

$$\{y_1(i), \ldots, y_N(i)\} \equiv \{(a, \lambda_0), \tau_0, (t_1, \lambda_1 - 1), \tau_1, \ldots, (t_n, \lambda_n - 1), \tau_n, (b, \lambda_{n+1})\} \backslash \tau_i,$$

where $\{\tau_i\}_0^n$ are parameters, $\tau_i \in (t_i, t_{i+1})$, and (t, λ) means that the point t is repeated λ times in the sequence. Now taking into account that

$$\begin{aligned} \frac{\partial}{\partial \xi_j} p(x) &= \frac{\partial}{\partial \xi_j} \sum_{k=0}^{N-r} \int_{\xi_k}^{\xi_{k+1}} \frac{(x-t)_+^{r-1}}{(r-1)} (-1)^{n-k} \sigma(t)\, dt \\ &= 2(-1)^{n-j-1} \sigma(\xi_j) \frac{(\xi_j - t)_+^{r-1}}{(r-1)} \end{aligned}$$

and setting $\sigma_j := 2(-1)^{n-j-1}\sigma(\xi_j)$, $j = 1, \ldots, N-r$, we see that

$$\det \Delta_i(s) = [c(s)]^n \left(\prod_{j=1}^{N-r} \sigma_j \right) . \det \delta_i(s)$$

with

$$\det \delta_i(s) := \int_{t_0}^{t_1} .\backslash^i/. \int_{t_n}^{t_{n+1}} \begin{bmatrix} u_1(t), \ldots, u_N(t) \\ y_1(i), \ldots, y_N(i) \end{bmatrix} d\tau_0 .\backslash^i/. d\tau_n,$$

where $\backslash^i/$ means that the symbol corresponding to i is skipped.

It follows from the total positivity of the truncated power kernel (see [7], [8]) that

$$\alpha \det \delta_i(s) \geq 0 \tag{7}$$

with some $\alpha = 1$ or $\alpha = -1$, independent of i. It is also seen that the points $y_1(i), \ldots, y_N(i)$ and the knots $\xi_1, \ldots, \xi_{N-r}$ of the σ-perfect spline $p(s;t)$ satisfy the *interlacing conditions*

$$y_k(i) < \xi_k < y_{k+r}(i), \quad k = 1, \ldots, N-r \tag{8}$$

for some $\tau_j \in (t_j, t_{j+1})$, $j = 1, \ldots, n$, provided $p(s;t)$ is a solution of system (6). Indeed, the σ-perfect spline $p(s;t)$ vanishes at

$$\{\theta_1, \ldots, \theta_N\} \equiv \{(t_0, \lambda_0), (t_1, \lambda_1), \ldots, (t_{n+1}, \lambda_{n+1})\}$$

and $p^{(r)}(s;t)$ changes sign at $\xi_1, \ldots, \xi_{N-r}$. Then, by Rolle's theorem, $\theta_k < \xi_k < \theta_{k+r}$. But $y_1(i), \ldots, y_N(i)$ becomes very close to $\theta_1, \ldots, \theta_N$ for some $\tau_0, \ldots, \tau_n$ and thus (8) holds too. This yields (on the basis of the total positivity of the truncated power kernel) strict inequality in (7). Therefore $\det \delta_i(s) \neq 0$. Finally, noticing that

$$c(s) \int_{t_i(s)}^{t_{i+1}(s)} p(s;\tau_i) d\tau_i = (-1)^{l_i} e_i(s)$$

and $\lambda_0 + \cdots + \lambda_i + l_i = N - \lambda_{n+1} =: N_0$, we get

$$\begin{aligned} \det J(s) &= [c(s)]^{n-1} \prod_{k=1}^{n} p^{(\lambda_k)}(s; t_k(s)) \sum_{i=0}^{n} (-1)^{N_0} e_i(s). \det \delta_i(s) \\ &= (-1)^{N_0} [c(s)]^{n-1} \alpha \prod_{j=1}^{N-r} \sigma_j \prod_{k=1}^{n} p^{(\lambda_k)}(s; t_k(s)) \sum_{i=0}^{n} e_i(s). |\det \delta_i(s)| \end{aligned}$$

and hence $\det J(s) \neq 0$, if (6) admits a solution $p(s;t), c(s)$.

Now we are ready to prove the existence of a solution of (6) for each $s \in [0,1]$. As we mentioned earlier $p_0(t)$ and $c^0 = 1$ is a solution for $s = 0$. Hence, by the result we just obtained, $\det J(0) \neq 0$. Then, by the implicit function theorem, there exists a unique set of continuous functions $a_i(s), \xi_i(s), t_i(s), c(s)$, which satisfy (6) in a neighborhood of 0. In other words, there exists a solution $p(s;t), c(s)$ of (6) for all s from some interval $[0, \beta)$ and $p(0;t) = p_0(t)$, $c(0) = 1$. So, if $\beta > 1$, our aim is achieved.

Assume that $\beta \leq 1$ for the maximal β. Letting $s \to \beta$ we shall define a σ-perfect spline $p(\beta;t)$ and $c(\beta)$ that satisfies (6) for $s = \beta$. In order to do this, note that

$$\| p^{(j)} \|_\infty \leq (b-a)^{r-j} \| \sigma \|_\infty, \quad j = 0, 1, \ldots, r-1 \tag{9}$$

for every function $p \in AC^{r-1}[a,b]$ that has at least $r-1$ zeros and $|p^{(r)}(t)| = \sigma(t)$ almost everywhere in $[a,b]$. This implies the equicontinuity of $p^{(j)}(s;t)$ on $[a,b]$ for $s \in [0, \beta)$. Since

$$|a_i(s)| = |p^{(i-1)}(s;a)|/(i-1),$$

(9) also yields the existence of a constant $M_1 > 0$ such that

$$| a_i(s) | \leq M_1 \quad \forall s \in [0, \beta).$$

Note further that $\| p(s;.) \|_\infty$ is not less than the best uniform approximation of the function

$$\int_a^b \frac{(x-t)_+^{r-1}}{(r-1)} \psi(\bar{\xi}(s);t)\, dt$$

by algebraic polynomials of degree $r-1$ in $[a,b]$. Thus there exists a constant $M_2 > 0$ such that

$$\| p(s;.) \|_\infty \geq M_2 \quad \forall s \in [0,\beta).$$

This implies

$$\int_a^b |p(s;t)|\, dt \geq M_3 > 0$$

because $p'(s;t)$ is bounded. Since

$$0 < \min\{e_k^0, e_k\} < e_k(s) < \max\{e_k^0, e_k\}, \quad k = 0,\dots,n$$

for $s \in [0,1]$ and

$$c(s) \int_a^b |p(s;t)|\, dt = e_0(s) + \cdots + e_n(s),$$

we get $0 < M_4 \leq c(s) \leq M_5$ with some constants M_4 and M_5. Finally, it follows from the equicontinuity of $p(s;t)$ and the boundedness of $c(s)$ and $e_k(s)$ that there is a $\delta > 0$ such that

$$\delta < t_{k+1}(s) - t_k(s) \quad \text{on } [0,\beta).$$

All these uniform estimates assure the existence of convergent subsequences of $a_i(s)$, $\xi_i(s), t_i(s)$, and $c(s)$ as $s \to \beta$ and the limit values define a σ-perfect spline $p(\beta;t) \in \mathcal{P}_r(\lambda_0,\dots,\lambda_{n+1})$ and a constant $c(\beta)$ that satisfy (6) for $s = \beta$. Then, by the result we proved in the beginning, $\det J(\beta) \neq 0$ and the solution $p(s;t), c(s)$ can be extended beyond β. This contradicts the assumption that β is maximal. Thus $\beta > 1$ and the existence is proved.

The differential equation approach of Fitzgerald and Schumaker [6] to problems of a similar kind can be used here to show the uniqueness in an elegant way. In order to do this, note that if we differentiate the equations in (6) with respect to s, we shall get a linear system of differential equations. Denote it by (S). Observe that the coefficient matrix of (S) with respect to

$$t_1'(s),\dots,t_n'(s), c'(s), a_1'(s),\dots,a_r'(s), \xi_1'(s),\dots,\xi_{N-r}'(s)$$

coincides with $J(s)$.

Consider the set Y of all points

$$\mathbf{y} = \{t_1,\dots,t_n,a_1,\dots,a_r,\xi_1,\dots,\xi_{N-r}\},$$

where t_j, a_j and ξ_j are the parameters of p when p runs over $\mathcal{P}_r(\lambda_0,\dots,\lambda_{n+1})$. According to Theorem B, $\{a_j\}$ and $\{\xi_j\}$ are defined uniquely by the zeros $a < t_1 < \cdots < t_n < b$ of p. Moreover, the system of equations

$$p^{(j)}(t_k) = 0, \quad k = 0,\dots,n+1,\ j = 0,\dots,\lambda_k - 1$$

with respect to $a_1, \dots, a_r, \xi_1, \dots, \xi_{N-r}$ has a nonzero Jacobian (since the zeros and the knots of every $p \in \mathcal{P}_r(\lambda_0, \dots, \lambda_{n+1})$ satisfy the interlacing conditions). Then, by the implicit function theorem, $\{a_j\}$ and $\{\xi_j\}$ are continuous functions of $t_1, \dots, t_n$. This shows that Y is a connected set because the set $\{(t_1, \dots, t_n) : a < t_1 < \cdots < t_n < b\}$ is connected. We have proved already that for each choice of the initial values $\mathbf{y} \in Y$ and $c(0) = 1$, system (6), and consequently the system of differential equations (S), has a unique solution $p(\mathbf{y}, s; t), c(\mathbf{y}, s)$ in $[0, 1]$. Define the mapping $\Phi : Y \to (Y, \mathbf{R})$ in the following way:

$$\Phi(\mathbf{y}) = \{a_1(\mathbf{y}, 1), \dots, a_r(\mathbf{y}, 1), \xi_1(\mathbf{y}, 1), \dots, \xi_{N-r}(\mathbf{y}, 1), c(\mathbf{y}, 1)\},$$

where $\{a_j(\mathbf{y}, s)\}, \{\xi_j(\mathbf{y}, s)\}$, and $c(\mathbf{y}, s)$ is the solution of the system (S) corresponding to the initial conditions defined by $\mathbf{y}$. As we mentioned already, the matrix of (S) coincides with $J(s)$ and thus it is nonsingular. Then, by virtue of a well-known result in the theory of differential equations, the solution of (S) depends continuously on the initial values $\mathbf{y}$. Therefore

(a) Φ is a continuous mapping.

Furthermore, the result $\det J(\mathbf{y}, s) \neq 0$ for $s = 1$ and the implicit function theorem show that system (6) has only isolated solutions for $s = 1$. Thus

(b) $\Phi(Y)$ consists of isolated points.

Observation (a), together with the fact that Y is connected, implies that $\Phi(Y)$ is a connected set. Then it follows from (b) that $\Phi(Y)$ must consist of only one point. Thus, starting from any $\mathbf{y} \in Y$ and $c = 1$, the described procedure leads to one and the same point. Therefore system (6) has a unique solution for $s = 1$. The uniqueness part of the proof is completed.

It remains to show the monotone dependence of c on e_k. Recall that $c = c(e_0, \dots, e_n)$ is a solution of (6) for $s = 1$. Since $\det J(1) \neq 0$, c is a differentiable function of $e_0, \dots, e_n$. Moreover,

$$\frac{dc}{de_k} = -\frac{\det J_k}{\det J(1)},$$

where J_k differs from $J(1)$ by its $(n+1)$th column, which has only one nonzero element, namely $(-1)^{l_k+1} e_k$, on position $n + \lambda_0 + \cdots + \lambda_k + 1$. Unfolding $\det J_k$ according to this column, we get

$$\begin{aligned} \det J_k &= (-1)^{\lambda_0 + \cdots + \lambda_k} (-1)^{l_k+1} e_k \det D \det \Delta_k(1) \\ &= -(-1)^{N_0} \det D \det \Delta_k(1) e_k. \end{aligned}$$

Therefore

$$\frac{dc}{de_k} = c \frac{|\det \delta_k(1)| e_k}{\sum_{i=0}^{n} e_i |\det \delta_i(1)|} > 0,$$

which was to be shown. The theorem is proved. □

In the study of problem (2) we shall actually use the following immediate consequence of Theorem 3.1.

COROLLARY 3.2. *Suppose that $\nu_0, \dots, \nu_{n+1}$ are fixed multiplicities such that $0 \leq \nu_k \leq r$, $k = 0, \dots, n+1$, $N := \nu_0 + \nu_1 + \cdots + \nu_{n+1} \geq r$. Then for each set of real numbers $y_0, \dots, y_{n+1}$ satisfying the requirement*

$$(-1)^{\nu_k}(y_{k+1} - y_k)(y_k - y_{k-1}) > 0, \quad k = 1, \dots, n, \tag{10}$$

there exists a unique σ-perfect spline φ of degree r with $N-r-1$ knots and points $a=x_0<x_1<\cdots<x_{n+1}=b$ satisfying the condition

$$(11)\qquad \begin{cases} \varphi(x_k)=y_k, & k=1,\dots,n,\\ \varphi^{(j)}(x_k)=0, & k=1,\dots,n,\ j=1,\dots,\nu_k,\\ \varphi^{(j)}(x_k)=0 & \text{for } k=0 \text{ and } k=n,\ j=1,\dots,\nu_k-1. \end{cases}$$

Moreover, the quantity $\| \varphi^{(r)} \|_\sigma$ is a strictly increasing function of $e_k := | y_{k+1}-y_k |$, $k=0,\dots,n$.

Proof. It suffices to observe that $\varphi'(t)$ is the solution of a problem considered in Theorem 3.1 and $c=\| \varphi^{(r)} \|_\sigma$. □

Now we are prepared to present the main result concerning the extremal problem (2).

Let us recall that here $\| \cdot \| = \| \cdot \|_\sigma$ and $F(\mathbf{x},\mathbf{y})$ are defined for the set Ω from (3).

THEOREM 3.3. *Let $\sigma(t)$ be an arbitrary positive, continuous function on $[a,b]$. Then for every given $\mathbf{y}=\{y_k\}_0^{n+1}$ satisfying requirement (10) there exists a unique σ-perfect spline φ of degree r with $N-r-1$ knots for which*

$$\| \varphi^{(r)} \| = \inf_{a=x_0<x_1<\cdots<x_{n+1}=b} \ \inf_{f\in F(\mathbf{x},\mathbf{y})} \| f^{(r)} \|_\sigma .$$

φ is uniquely characterized by the condition

$$\varphi^{(\nu_k)}(x_k^*)=0, \quad k=1,\dots,n,$$

at the extremal points $a=x_0^*<x_1^*<\cdots<x_{n+1}^*=b$. Moreover, φ is the unique extremal function to (2).

Proof. By Corollary 3.2 there exists a unique set of points $\mathbf{x}^*=\{x_k^*\}_0^{n+1}$ and a unique σ-perfect spline $\varphi_*(t)$ of degree r with $N-r-1$ knots that satisfy (11). We shall show that φ_* is the wanted extremal function to (2). To do this, recall that by Theorem A, for any given $\mathbf{x}=\{x_k\}_0^{n+1}$ there exists a σ-perfect spline $\varphi(\mathbf{x};t)\in F(\mathbf{x},\mathbf{y})$ of degree r with $N-r-1$ knots such that

$$\| \varphi^{(r)}(\mathbf{x};.) \|_\sigma = \inf\{ \| g^{(r)} \|_\sigma :\ g\in F(\mathbf{x},\mathbf{y})\}.$$

Since $\mathbf{y}$ satisfies (10), $\varphi(\mathbf{x};t)$ has an additional local extremum in (x_{k-1},x_{k+1}) at some point t_k, for $k=1,\dots,n$. Then $\varphi(\mathbf{x};t)$ may be considered as the solution of a problem like that studied in Corollary 3.2 with multiplicities $V:=\{\nu_0,1,\nu_1-1,1,\nu_2-1,1,\dots,\nu_{n+1}\}$, in an order corresponding to the order of the points $\{(x_k)_0^{n+1},(t_k)_1^n\}$ in $[a,b]$, and values $Y=\{y_o,\tilde{y}_1,y_1,\tilde{y}_2,y_2,\dots,y_{n+1}\}$, where $\tilde{y}_k=\varphi(\mathbf{x};t_k)$. But φ_* is the solution of the problem with the same multiplicities V and values Y with $\tilde{y}_k=y_k, k=1,\dots,n$. Since for each $\mathbf{x}\neq\mathbf{x}^*$, $e_k(\varphi):=|\tilde{y}_{k+1}-\tilde{y}_k|>|y_{k+1}-y_k|=e_k(\varphi_*)$, the monotone dependence of $\| \varphi^{(r)} \|_\sigma$ on $e_k(\varphi)$ shows that $\| \varphi_*^{(r)} \|_\sigma < \| \varphi^{(r)}(\mathbf{x};.) \|_\sigma$. Thus φ_* is an extremal function to problem (2).

Assume that f is another extremal function, different from φ_*. It follows from Theorem A that

$$\| \varphi_*^{(r)} \|_\sigma < \| \varphi^{(r)}(\mathbf{x};.) \|_\sigma \le \| g^{(r)} \|_\sigma$$

for each $\mathbf{x}\neq\mathbf{x}^*$ and $g\in F(\mathbf{x},\mathbf{y})$. Therefore $f\in F(\mathbf{x}^*,\mathbf{y})$.

It is seen also that $f^{(\nu_k)}(x_k^*)=0$ for $k=1,\dots,n$. Otherwise f would have additional local extremum at some point t_k from (x_{k-1},x_{k+1}) and then, as in the

reasoning above, we can show that the σ-perfect spline $\varphi_0(t)$ that interpolates f at the points $(x_0, t_1, x_1, t_2, x_2, \ldots, t_n, x_n, x_{n+1})$ with multiplicities $\{\nu_0, 1, \nu_1 - 1, 1, \nu_2 - 1, \ldots, 1, \nu_n - 1, \nu_{n+1})$ would satisfy

$$\| \varphi_*^{(r)} \|_\sigma < \| \varphi_0^{(r)} \|_\sigma \le \| f^{(r)} \|_\sigma,$$

which is a contradiction. Therefore $f^{(\nu_k)}(x_k^*) = 0$ for $k = 1, \ldots, n$.

Next the proof proceeds as in the simple node case (considered in [10]). If $f \neq \varphi_*$ on some subinterval (x_{k-1}^*, x_k^*), then $\varphi'(t) - f'(t)$ changes its sign on (x_{k-1}^*, x_k^*). Thus, for sufficiently small $\epsilon > 0$ the function $\varphi_*'(t) - (1-\epsilon)f'(t)$ would change sign too in (x_{k-1}^*, x_k^*) and hence would have at least $1 + \nu_0 - 1 + \nu_1 + \nu_2 + \cdots + \nu_n + \nu_{n+1} - 1 = N - 1$ zeros in $[a, b]$ counting the multiplicities. Then, by the Rolle theorem, $\varphi_*^{(r)}(t) - (1-\epsilon)f^{(r)}(t)$ would have at least $N - r$ sign changes while $\varphi_*(t)$ has only $N - r - 1$ knots, a contradiction. The proof is completed. □

Remark 1. The extremal function φ_* is a σ-perfect spline of degree r with $N-r-1$ knots. Since $\varphi_*'(t)$ has $N - 2$ zeros (prescribed by (11)), it is seen that $\varphi_*(t)$ has no other local extrema except those eventually at $x_1^*, \ldots, x_n^*$. Therefore $\varphi_*(t)$ is a strictly monotone function between x_k^* and x_{k+1}^* for $k = 0, \ldots, n$.

We promised in the beginning of this paper to characterize the smoothest interpolant in the case some of the values $\{f_k\}_0^{N+1}$ in (1) are equal. We are ready to do this now.

THEOREM 3.4. *Let $\{f_k\}_0^{N+1}$ be given values satisfying the requirement*

$$(f_{k+1} - f_k)(f_k - f_{k-1}) \le 0, \quad k = 1, \ldots, N. \tag{12}$$

Assume that $N := \nu_1 + \cdots + \nu_n \ge r$ and the first ν_1 values in the sequence $f_1, \ldots, f_N$ are equal to $y_1 \neq y_0 := f_0$, the next ν_2 are equal to $y_2 \neq y_1$, and so on, with the last ν_n values equal to $y_n \neq y_{n+1} := f_{N+1}$. Then, for any fixed positive, piecewise continuous $\sigma(t)$ on $[a, b]$, the extremal problem

$$\inf_{a=x_0 \le x_1 \le \cdots \le x_{N+1}=b} \quad \inf_{g \in \Omega,\, g(x_k)=g_k,\, k=0,\ldots,N+1} \| g^{(r)} \|_\sigma \tag{13}$$

has a unique solution $(\mathbf{x}^*, \varphi)$. The sequence $\mathbf{x}^* = \{x_i^*\}_0^{N+1}$ contains only $n+1$ distinct points $a = t_0^* < t_1^* < \cdots < t_{n+1}^* = b$ and φ is the σ-perfect spline of degree r with $N - r + 1$ knots, which satisfies the conditions

$$\varphi(t_k^*) = y_k, \quad k = 0, \ldots, n+1,$$

$$\varphi^{(j)}(t_k^*) = 0, \quad k = 1, \ldots, n,\ j = 1, \ldots, \nu_k.$$

Proof. In other words, the solution of (13) coincides with the unique solution of the problem

$$\inf_{a=t_0<t_1<\cdots<t_{n+1}=b} \quad \inf_{g \in F(\mathbf{t}, \mathbf{y})} \| g^{(r)} \|_\sigma,$$

where $F(\mathbf{t}, \mathbf{y})$ is defined with respect to the multiplicities $1, \nu_1, \ldots, \nu_n, 1$.

Assume that φ is an extremal function to the problem (12) and $\mathbf{x}^*$ is an extremal set of points. Clearly φ is a σ-perfect spline of degree r with no more than $N - r + 1$

knots. Assume further that $\mathbf{x}^*$ contains more than $n+1$ distinct points. Then, because of (13), φ would satisfy interpolation conditions of the form

$$\varphi(x_k) = \hat{f}_k, \quad k = 0, \dots, N+1,$$

$$\varphi^{(j)}(x_{k+j-1}) = 0 \quad \text{if } x_{k+j-1} = x_k,\ 1 \le k+j-1 \le N,$$

with some $\{x_k\}$ and $\{\hat{f}_k\}$ such that $\mid \hat{f}_{k+1} - \hat{f}_k \mid \geq \mid f_{k+1} - f_k \mid$, $k = 0, \dots, N$. Moreover, at least one of the last inequalities is strict. Then, by the monotonicity of $\| \varphi^{(r)} \|_\sigma$ on $\mid \hat{f}_{k+1} - \hat{f}_k \mid$ (see Theorem 3.3), we can find a σ-perfect spline φ_1 with $\| \varphi_1^{(r)} \|_\sigma < \| \varphi^{(r)} \|_\sigma$ and still take consecutively the values $f_0, f_1, \dots, f_{N+1}$. This is a contradiction to the extremality of φ. Thus the extremal set of points $\mathbf{x}^*$ consists of $n+1$ groups of coinciding points. Using the equicontinuity of the σ-perfect splines, one sees that the required Lagrangean interpolation conditions

$$\varphi(x_k) = f_k, \quad k = 0, \dots, N+1,$$

transforms into the corresponding Hermitian conditions for $\mathbf{x} = \mathbf{x}^*$. The theorem is proved. □

4. Applications: $\mathbf{W}_p^r$ case. We shall apply the results of the previous section to study problem (2) for $\Omega = W_p^r[a,b]$ and $\| \cdot \| = \| \cdot \|_p$. As usual, set

$$\| f \|_p := \left\{ \int_a^b |f(t)|^p \, dt \right\}^{1/p} \qquad \text{for } 1 < p < \infty,$$

$$\| f \|_\infty := \text{supvrai } \{ |f(t)| : t \in [a,b] \}.$$

Clearly the particular choice $\sigma(t) \equiv 1$ of the function σ leads to the case $p = \infty$. We concentrate here on $1 < p < \infty$.

Our main observation is that the characterization of the extremal function to problem (2) in $W_p^r[a,b] (1 < p < \infty)$ can be derived from an extension of Theorems 3.1 and 3.3, which allows functions σ that could vanish on subintervals. Only for the sake of simplicity, we gave a detailed proof of these results in the case of strictly positive $\sigma(t)$. Now we shall see that Theorem 3.1 (and consequently, Theorems 3.3 and 3.4), holds also for any integrable nonnegative function on $[a,b]$, which is strictly positive on a subset of $[a,b]$ of positive measure. Let $\sigma(t)$ be such a function. For every small $\epsilon > 0$, define the function $\sigma_\epsilon(t)$ as a positive continuous function such that $|\sigma_\epsilon(t) - \sigma(t)| \le \epsilon$ almost everywhere on $[a,b]$. By Theorem 3.1, there is a unique σ_ϵ-perfect spline p_ϵ from $\mathcal{P}_r(\lambda_0, \dots, \lambda_{n+1})$ and a constant $c(\epsilon) > 0$ that satisfy (5). Since estimation (9) holds for σ_ϵ , we see (following the reasoning after (9)) that the functions $\{p_\epsilon(t)\}$ are equicontinuous on $[a,b]$. Thus, letting $\epsilon \to 0$ we can find a subsequence of $\{p_\epsilon(t)\}$, which tends uniformly to a σ-perfect spline $p_0(t)$. The critical point in the proof is to show that the sequence $\{c(\epsilon)\}$ is also bounded as ϵ approaches 0. The assumptian that $\sigma(t) > 0$ on a subset of measure zero is used here. Let us prove the boundedness of $c(\epsilon)$. Denote by $\bar{\xi}(\epsilon) = \{\xi_i(\epsilon)\}$ the knots of p_ϵ. Clearly the best uniform approximation of the function

$$s(x) := \int_a^b \frac{(x-t)_+^{r-1}}{(r-1)} \psi(\bar{\xi}(\epsilon); t) \, dt$$

on $[a,b]$ by algebraic polynomials of degree $r-1$ is distinct from zero (since $s^{(r)}(t) \equiv \sigma(t)$ and $\sigma(t) > 0$ on a subinterval). Then there is a constant $M_2 > 0$ such that $\| p_\epsilon \| > M_2$ for all sufficiently small $\epsilon > 0$. Now following the same reasoning as that in Theorem 3.1 (the part after (9)), we deduce that the estimate $0 < M_4 < c(\epsilon) < M_5$ holds for all small $\epsilon > 0$. Thus $\{c(\epsilon)\}$ is bounded and going to a subsequence if necessary, $c(\epsilon) \to c(0)$ as $\epsilon \to 0$. The existence of a solution to the problem in Theorem 3.1 is proved.

Next we sketch the proof of the uniqueness. Let $a_1(\epsilon), \dots, a_r(\epsilon), \xi_1(\epsilon) < \cdots < \xi_{N-r}(\epsilon)$ be the parameters of $p_\epsilon(t)$ $(\epsilon \geq 0)$. By the construction of $p_0(t)$,

$$\xi_i(\epsilon) \to \xi_i(0), \quad a_i(\epsilon) \to a_i(0). \tag{14}$$

Assume that there is another solution $\hat{c}, \hat{p}(t)$ of (5), with parameters $\{\hat{a}_i\}, \{\hat{\xi}_i\}$ and zeros $\{\hat{t}_i\}$. Since $N \geq r$, we have $\hat{c} > 0$. Furthermore, for small $\epsilon > 0$, denote by $\hat{p}_\epsilon(t)$ the σ_ϵ-perfect spline with the same parameters $\{\hat{a}_i\}, \{\hat{\xi}_i\}$ as $\hat{p}(t)$. Clearly $\hat{c}$ and $\hat{p}_\epsilon(t)$ satisfy a system of equations

$$\begin{cases} \hat{p}_\epsilon^{(j)} = h_{ij}(\epsilon), & i = 0, \dots, n+1,\ j = 0, \dots, \lambda_j - 1, \\ \hat{c} \int_{\hat{t}_k}^{\hat{t}_{k+1}} \hat{p}_\epsilon(t)\, dt = e_k(\epsilon), & k = 0, \dots, n, \end{cases} \tag{15}$$

where $H(\epsilon) := \{\{h_{ij}(\epsilon)\}, \{e_k(\epsilon)\}\}$ and $H(\epsilon) \to H(0)$ as $\epsilon \to 0$. Note that $h_{ij}(0) = 0$ and $e_k(0) = e_k$. Consider now (15) with a right-hand side $\alpha H(\epsilon) + (1-\alpha)H(0)$ for $\alpha \in [0,1]$. Denote by $J(\epsilon;\alpha)$ the Jacobian matrix of (15) with respect to $\{t_1, \dots, t_n, c, a_1, \dots, a_r, \xi_1, \dots, \xi_{N-r}\}$. Let $\det J(\epsilon;\alpha)$ be the determinant evaluated at the solution of (15). Since (15) admits a solution for $\alpha = 0$, namely $(c(\epsilon), p_\epsilon(t))$, we see that as in the proof of Theorem 3.1, $\det J(\epsilon;0) = \beta(\epsilon)B \neq 0$, where

$$\beta(\epsilon) := 2 \prod_{j=1}^{N-r} \sigma_\epsilon(t_j)$$

and B is a determinant expression, distinct from zero. Similarly,

$$\det J(\epsilon;\alpha) = \beta(\epsilon)(B + \gamma(\epsilon;\alpha))$$

with some γ such that $|\gamma(\epsilon;\alpha)| \to 0$ as $\epsilon \to 0$ for each $\alpha \in [0,1]$. It is seen that $\det J(\epsilon;\alpha) \neq 0$ for sufficiently small ϵ and every $\alpha \in [0,1]$. Thus, starting from $(c(\epsilon), p_\epsilon(t))$ one can get by continuous deformations the solution $(\hat{c}, \hat{p}_\epsilon(t))$ of (15) moving α from 0 to 1. By the Implicit Function Theorem

$$\frac{d\xi_j}{d\alpha} = -\frac{\det J_j(\epsilon;\alpha)}{\det J(\epsilon;\alpha)},$$

where $J_j(\epsilon;\alpha)$ is obtained from $J(\epsilon;\alpha)$ by replacing the column corresponding to ξ_j by the column $(H(\epsilon) - H(0))$. Since $\mid H(\epsilon) - H(0) \mid$ tends to 0 as $\epsilon \to 0$, we get $\det J_j(\epsilon;\alpha) = O(\epsilon\beta(\epsilon))$ and therefore $\mid d\xi_j/d\alpha \mid = O(\epsilon)$. Now the assumption that $\xi_j(\epsilon) \neq \hat{\xi}_j$ would lead to contradiction for small ϵ. Thus $\xi_j(\epsilon) = \hat{\xi}_j$ for all $j = 1, \dots, N-r$ and sufficiently small $\epsilon > 0$. This is possible only if $\xi_j(0) = \hat{\xi}_j$ for all j. Similarly one shows that $c(0) = \hat{c}$, $a_i(0) = \hat{a}_i$, i.e., that (5) has a unique solution.

In order to prove the monotone dependence of c on e_k, consider the function $c(\epsilon; e_k)$, which is the solution of the problem (5) for $\sigma_\epsilon(t)$ and $e_0, \dots, e_n$ (e_k being chosen as a parameter). We have to show that

$$c(0; e_k) < c(0; e_k + h) \tag{16}$$

for each $h > 0$. It follows from the expression of dc/de_k that

$$\left|\frac{dc(\epsilon; e_k)}{de_k}\right| > \text{ const.}$$

for each sufficiently small $\epsilon > 0$. Thus

$$c(\epsilon; e_k) + K.h \leq c(\epsilon; e_k + h)$$

with some $K > 0$. Now the desired inequality (16) follows from the fact that $c(\epsilon; e_k) \to c(0; e_k)$ and $c(\epsilon; e_k + h) \to c(0; e_k + h)$ as ϵ tends to 0. Thus we proved the following remark.

Remark 2. Theorems 3.1, 3.3, and 3.4 hold for every integrable nonnegative function $\sigma(t)$, which is distinct from zero on a subset of positive measure.

After this remark we can prove our characterization result.

THEOREM 4.1. *Let $1 < p < \infty$. Suppose that φ is an extremal function to the problem*

$$\inf_{a=t_0<t_1<\cdots<t_{n+1}=b} \quad \inf_{f\in F(\mathbf{t},\mathbf{y})} \| f^{(r)} \|_p,$$

where $F(\mathbf{t}, \mathbf{y})$ is defined with respect to $\Omega = W_p^r[a, b]$, multiplicities $\nu_0, \nu_1, \dots, \nu_{n+1}$, and values $\{y_k\}_0^{n+1}$ such that

$$(-1)^{\nu_k}(y_{k+1} - y_k)(y_k - y_{k-1}) > 0, \quad k = 1, \dots, n.$$

Then φ satisfies the conditions

$$\varphi^{(\nu_k)}(t_k^*) = 0, \quad k = 1, \dots, n,$$

at the extremal points $\mathbf{t}^* = \{t_k^*\}_0^{n+1}$ and $\varphi(t)$ is strictly monotone on (t_k^*, t_{k+1}^*), $k = 0, \dots, n$.

Proof. Any function f from $F(\mathbf{t}, \mathbf{y})$ satisfies the inequality

$$\| f^{(r-1)} \|_{C[a,b]} \leq \int_a^b | f^{(r)}(t) | \, dt \leq (b-a)^{1/q} \| f^{(r)} \|_p,$$

where $1/q + 1/p = 1$. Then $\| f' \|_{C[a,b]} \leq$ const.$\| f^{(r)} \|_p$ and this, together with the assumption $y_{k+1} - y_k > 0$, implies the existence of noncoincident extremal points (i.e., such that $a = t_0^* < t_1^* < \cdots < t_{n+1}^* = b$). Next, we know from [5] that for fixed $\mathbf{t}^*$ there is a unique function φ from $F(\mathbf{t}^*, \mathbf{y})$ of smallest L_p-norm of its rth derivative. Moreover,

$$\varphi^{(r)}(t) = \text{ const. } | \psi(t) | \operatorname{sign}\psi(t),$$

where $\psi(t)$ is a function of the form

$$\psi(t) = \sum_{i=1}^{N-r} \alpha_i B_i(t), \quad N := \nu_0 + \cdots + \nu_{n+1},$$

and B_i are B-splines. Thus, $\varphi^{(r)}(t)$ is a piecewise continuous function. By the variation diminishing property of the B-spline sequences (see [7] or [3]), $\psi(t)$ has at most $N-r-1$ sign changes. On the other hand, it follows from the interpolation conditions on φ that $\varphi^{(r)}(t)$, and consequently $\psi(t)$, has at least $N-r-1$ sign changes. Thus, $\psi(t)$ has exactly $N-r-1$ sign changes. Then all coefficients α_i must be distinct from zero and hence, $\psi(t)$ cannot vanish on subintervals. (Similar reasoning was used before in the proof of Theorem 2 from [2].) Thus the function $\sigma(t) := \mid \varphi^{(r)}(t) \mid$ is positive almost everywhere on $[a,b]$. Clearly φ is an extremal function to the problem considered in Theorem 3.3 with this σ. Then the characterization of φ follows from Theorem 3.3, taking into account Remark 2. The proof is completed. □

Finally, note that the same reasoning also shows that Theorem 3.4 holds in the space $W_p^r[a,b]$.

The uniqueness of the extremal function in $W_p^r[a,b]$ is still an open problem. Uniqueness results are known only for $p=2$ in case $r=2$ [9] and $r=3$ [11].

REFERENCES

[1] B. D. BOJANOV, *Perfect splines of least uniform norm*, Anal. Math., 6(1980), pp. 185–197.

[2] ———, *σ-Perfect splines and their application to optimal recovery problems*, J. Complexity, 3(1987), pp. 429–450.

[3] ———, *B-splines with Birkhoff knots*, Constr. Approx., 4(1988), pp. 147–156.

[4] ———, *Optimal recovery of differentiable functions*, Mat. Sb., 181 (1990), pp. 334–353. (In Russian.) English transl.: Math. USSR Sb., 69 (1991), pp. 357–377.

[5] C. DE BOOR, *On "best" interpolation*, J. Approx. Theory, 16(1976), pp. 28–42.

[6] C. H. FITZGERALD AND L. L. SCHUMAKER, *A differential equation approach to interpolation at extremal points*, J. Analyse Math., 22(1969), pp. 117–134.

[7] S. KARLIN, *Total Positivity*, Vol. 1, Stanford University Press, Stanford, CA, 1968.

[8] ———, *Total positivity, interpolation by splines and Green's functions of differential operators*, J. Approx. Theory, 4(1971), pp. 91–112.

[9] S. MARIN, *An approach to data parameterization in parametric cubic spline interpolation problems*, J. Approx. Theory, 41(1984), pp. 66–86.

[10] A. PINKUS, *On smoothest interpolation*, SIAM J. Math. Anal., 19 (1988), pp. 1431–1441.

[11] R. ULUCHEV, *Smoothest interpolation with free knots in W_p^r*, in Progress in Approximation Theory, P. Nevai and A. Pinkus, eds., Academic Press, Inc., San Diego, 1991, pp. 787–896.