Pacific Journal of Mathematics

Volume 216 No. 2 October 2004

PACIFIC JOURNAL OF MATHEMATICS

http://www.pjmath.org

Founded in 1951 by

E. F. Beckenbach (1906–1982) F. Wolf (1904–1989)

See inside back cover or www.pjmath.org for submission instructions.

Regular subscription rate for 2006: $425.00 a year (10 issues). Special rate: $212.50 a year to individual members of supporting institutions. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163, U.S.A. Prior back issues are obtainable from Periodicals Service Company, 11 Main Street, Germantown, NY 12526-5635. The Pacific Journal of Mathematics is indexed by Mathematical Reviews, Zentralblatt MATH, PASCAL CNRS Index, Referativnyi Zhurnal, Current Mathematical Publications and the Science Citation Index.

The Pacific Journal of Mathematics (ISSN 0030-8730) at the University of California, c/o Department of Mathematics, 969 Evans Hall, Berkeley, CA 94720-3840 is published monthly except July and August. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices. POSTMASTER: send address changes to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163.

PUBLISHED BY PACIFIC JOURNAL OF MATHEMATICS
at the University of California, Berkeley 94720-3840
A NON-PROFIT CORPORATION
Typeset in LaTeX

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

CR APPROXIMATION ON A NONRIGID HYPERSURFACE GRAPH IN $\mathbf{C}^n$

Al Boggess and Roman Dwilewicz

Let M be a hypersurface in $\mathbf{C}^n$ that is the graph over a $(2n-1)$-linear real space. The main result of the paper is that any CR function on M can be uniformly approximated on compact subsets by entire functions on $\mathbf{C}^n$.

1. Introduction

In [**BT**], Baouendi and Treves prove a local CR approximation result that says (in particular) that CR functions on a CR submanifold can be locally approximated by entire functions. The global version of this theorem is false, as seen by the example $M = \{(z,w) \in \mathbf{C}^2;\ |z| = 1\}$ — the function $f(z,w) = 1/z$ is clearly CR on M but cannot be approximated uniformly on compact subsets of M by entire functions. A natural question is whether a global version of the CR approximation theorem holds in cases where there are no topological obstructions. In this work, we establish a partial answer to this question by showing that the global CR approximation theorem holds on any smooth real hypersurface that is globally presented as a graph.

To precisely state our theorem, let $\mathbf{C}^n$ be given coordinates $(z,w) \in \mathbf{C} \times \mathbf{C}^{n-1}$. Let $h : \mathbf{R} \times \mathbf{R}^{n-1} \times \mathbf{R}^{n-1} \to \mathbf{R}$ be a smooth real-valued function and let M be its graph:

$$M = \big\{(z,w) = (h(y,u,v) + iy, u + iv) \in \mathbf{C} \times \mathbf{C}^{n-1}\big\}.$$

Our goal is to prove the following theorem:

Theorem 1. *Given a compact set $K \subset \mathbf{C}^n$, there is a compact set $K' \subset \mathbf{C}^n$ (with $K \subset K'$), such that if f is continuous and* CR *on a neighborhood of $M \cap K'$, then there is a sequence of entire functions on $\mathbf{C}^n$ that converges to f uniformly on $M \cap K$.*

Earlier work (see [**B**]) established this theorem in the case where the graphing function h is rigid (i.e., independent of y) and has polynomial growth. Earlier results on global CR approximation, see for example [**DG**], required additional technical assumptions (including certain convexity restrictions). Global approximation on totally real submanifolds of smooth functions by holomorphic functions has been considered by many authors, including [**HW**] and [**Ch**].

This work only handles the hypersurface case. The analogous theorem for higher codimension is still open and it does not appear that the techniques in this work can be easily modified to handle the case of higher codimension.

2. Outline of proof

As with the local version of the CR approximation theorem, the idea is to first integrate the given CR function $f(z,w)$ against a kernel that approximates the identity. The kernel will be entire in all variables, but is supported along a slice of M that moves as the point (z,w) moves (in a sort of Radon transform style). Since the slices don't depend holomorphically on (z,w), the resulting approximating sequence is not a priori holomorphic. Thus, the next step is to show (with the help of Stokes' Theorem) that the slice can be fixed, independently of the point (z,w). All of this analysis is done locally at first and then pieced together globally with a correction term that involves solving a $\bar\partial$ problem with estimates (in a Cousin-type fashion).

3. Definition of kernel

In [**BT**], an approximation to the identity kernel is used that somewhat looks like a heat kernel in $\mathbf{C}^n$. Our kernel also looks like a heat kernel but with the new feature of involving an extra complex parameter that gets integrated along an infinite ray in $\mathbf{C}$ (somewhat like a complex version of the Laplace transform). This extra parameter will allow us to handle global approximation. More specifically, for a continuous function f on M, we consider the sequence

$$E_\epsilon(f)(z,w) = \frac{C}{\epsilon^n}\int_{(\zeta,\eta)\in M_u}\int_{\alpha\in C_\theta} f(\zeta,\eta)E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n\, d\alpha\, d\zeta d\eta$$

where the kernel E_ϵ is given by

$$E_\epsilon(\zeta,\eta,\alpha,z,w) = \exp\biggl(\alpha^2\Bigl(\frac{\zeta-z}{\epsilon}\Bigr)^2 + \alpha^2\Lambda\Bigl(\frac{w-\eta}{\epsilon}\Bigr)^2 - \alpha^p\biggr).$$

Here, $\alpha,\zeta\in\mathbf{C}$ and $\eta\in\mathbf{C}^{n-1}$ will be the variables of integration; (z,w) is a point in $\mathbf{C}\times\mathbf{C}^{n-1}=\mathbf{C}^n$; ϵ is a real parameter that will later converge to zero; and Λ is a fixed positive real number (chosen below). The term $(w-\eta)^2$ stands for the sum of the squares of the components of $w-\eta\in\mathbf{C}^{n-1}$ (without any conjugates). The power p will be a real number slightly larger than 2 and will be chosen below. The domain of integration involves M_u, which is the intersection of M with the slice $\{(\zeta,\eta)\in\mathbf{C}\times\mathbf{C}^{n-1};\ \operatorname{Re}\eta = u\}$ where u is the real part of w. The other component of the domain of integration is the ray $C_\theta=\{\alpha = re^{i\theta};\ r>0\}$, where the angle θ will be between $-\frac{\pi}{4}$ and $\frac{\pi}{4}$ and its choice below will be determined by the desire to make the real part of the exponent of the kernel as negative as possible (so that the

resulting integral will converge). One angle will not work for all (ζ,η) and (z,w) and so a localization argument with many angles must be used.

4. Main estimate

In order to obtain convergence results, we will need an estimate on the real part of the exponent of our kernel. Various constants will emerge in the statement and proof of this estimate that depend on the compact set K or K'. We denote the dependency of a constant, such as δ, on a set A by writing δ_A. We also use the following coordinates: the variable of integration (ζ,η) will always lie on M and will be written as

$$(\zeta,\eta) = (h(s,u',t) + is,\ u' + it) \in M \quad \text{with} \quad s \in \mathbf{R},\ u',\ t \in \mathbf{R}^{n-1}.$$

The point (z,w) lies in $\mathbf{C}\times\mathbf{C}^{n-1}$ (not necessarily in M) and will be written as

$$(z,w) = (h(y,u,v) + q + iy,\ u + iv) \quad \text{with} \quad y,\ q \in \mathbf{R},\quad u,\ v \in \mathbf{R}^{n-1}.$$

Note that q represents the vertical distance from the point (z,w) to M. In particular (z,w) belongs to M if and only if $q = 0$.

We now state the main estimate on the exponent of our kernel.

Lemma 1 (Main estimate)**.** *Suppose $K \subset K'$ are given compact sets in $\mathbf{C}^n$. There exist constants $\Lambda_{K'}, C_{K'}, d_K, \delta_K,\ \delta_{K'},\ C_K > 0$ and $p = p_{K'}$ with $2 < p < 2/(1-\delta_{K'})$ and an open cover Q_j $(j = 1,\dots,N)$ of K and an open cover Q'_k $(k = 1,\dots,N')$ of K' with diameters less than $d_K/2$ and angles $\theta_{j,k}$ such that for all $\alpha \in C_{\theta_{j,k}} = \{re^{i\theta_{j,k}};\ r > 0\}$ and for all*

$$(\zeta,\eta) = (h(s,u',t) + is,\ u' + it) \in Q'_k \cap M \qquad \text{and}$$
$$(z,w) = (h(y,u,v) + q + iy,\ u + iv) \in Q_j$$

(with $u,\ u',\ t,\ v \in \mathbf{R}^{n-1}$ and $s,\ y,\ q \in \mathbf{R}$) we have:

1) *If $|s-y| < 2d_K$ for all $(z,w) \in Q_j$ and $(\zeta,\eta) \in Q'_k \cap M$, then $\theta_{j,k}$ can be chosen with $|\theta_{j,k}| < \frac{\pi}{4} - \delta_K$ and*

$$\text{(1)}\quad \operatorname{Re}\Big(\alpha^2\Big(\frac{\zeta - z}{\epsilon}\Big)^2 + \Lambda_{K'}\alpha^2\Big(\frac{w-\eta}{\epsilon}\Big)^2 - \alpha^p\Big)$$
$$\le -\delta_K\Big(\Big|\frac{s-y}{\epsilon}\Big|^2 r^2 + \Big|\frac{t-v}{\epsilon}\Big|^2 r^2 + r^p\Big) + C_{K'}\Big|\frac{u-u'}{\epsilon}\Big|^2 r^2 + C_K\Big(\frac{q}{\epsilon}\Big)^2 r^2.$$

2) *If $|s-y| \ge d_K$ for all $(z,w) \in Q_j$ and $(\zeta,\eta) \in Q'_k \cap M$, then $\theta_{j,k}$ can be chosen with $|\theta_{j,k}| < \frac{\pi}{4} - \delta_{K'}$ and*

$$\text{(2)}\qquad \operatorname{Re}\Big(\alpha^2\Big(\frac{\zeta - z}{\epsilon}\Big)^2 + \Lambda_{K'}\alpha^2\Big(\frac{w-\eta}{\epsilon}\Big)^2 - \alpha^p\Big)$$
$$\le -\delta_{K'}\Big(\Big|\frac{s-y}{\epsilon}\Big|^2 r^2 + \Big|\frac{t-v}{\epsilon}\Big|^2 r^2 + r^p\Big) + C_{K'}\Big|\frac{u-u'}{\epsilon}\Big|^2 r^2.$$

Remark. Since the diameters of Q_j and $Q_k^{'}$ are less than $d_K/2$, each Q_j and $Q_k^{'}$ falls under either case 1 or case 2. In the second case, the constant, $\delta_{K'}$, depends on K' (as opposed to K for the constant in the first case). Since $K \subset K'$, $\delta_{K'}$ will generally be smaller than δ_K. Also note that the estimate in the second case does not involve q^2.

Proof of the lemma. If $\widehat{W}$ is a complex number with $|\mathrm{Arg}\,(\widehat{W})| \geq 3\delta_1 > 0$, there is an angle θ with $|\theta| < \frac{\pi}{4} - \delta_1$ such that $\mathrm{Re}\,\{e^{2i\theta}\widehat{W}\} < 0$. If we let $C_\theta = \{re^{i\theta};\ r > 0\}$, there is a $\delta_2 > 0$ such that $\mathrm{Re}\,\{\alpha^2\widehat{W}\} \leq -\delta_2|\widehat{W}||\alpha|^2$ for all $\alpha \in C_\theta$. This observation will be used below to choose the angles $\theta_{j,k}$.

We first examine the term $(\zeta - z)^2$ that appears in the exponent of $E_\epsilon(\zeta, \eta, \alpha, z, w)$. We write

$$(\zeta, \eta) = H(s, u', t) = (h(s, u', t) + is,\ u' + it),$$
$$(z, w) = H(y, u, v) + (q, 0) = (h(y, u, v) + iy + q,\ u + iv).$$

Note that (ζ, η) belongs to M; but (z, w) belongs to M if and only if $q = 0$.

We wish to rotate $(\zeta - z)^2$ so that its real part is negative. For technical reasons that will be clear later, the rotation angle must be less than $\frac{\pi}{2}$. Thus, the troublesome case to handle is when $\zeta - z$ is a positive real number (which would then require a rotation angle greater than $\frac{\pi}{2}$ to make its real part negative). Therefore, we will handle separately the cases when $s - y = \mathrm{Im}\{\zeta - z\}$ is small and not so small.

First, observe that there is a $\widetilde{\delta}_K > 0$ such that

$$|\mathrm{Arg}\,(h_y(y, u, v) + i)\,| \geq \widetilde{\delta}_K \quad \text{for all } (y, u, v) \in K.$$

By choosing $d_K > 0$ small enough and by shrinking $\widetilde{\delta}_K > 0$, we can arrange

$$\left|\mathrm{Arg}\left(\frac{h(s, u, v) - h(y, u, v)}{s - y} + i\right)\right| \geq \widetilde{\delta}_K \tag{3}$$
$$\text{for all } (y, u, v) \in K, \text{ and } |s - y| \leq 2d_K.$$

Choose a smooth function $\phi : \mathbf{R} \to [0, 1]$ such that $\phi(s - y) = 0$ for $|s - y| \leq d_K/2$ and $\phi(s - y) = 1$ for $|s - y| \geq d_K$.

After adding and subtracting terms, we obtain

$$\begin{aligned}
\zeta - z &= h(s, u', t) - h(y, u, v) - q + i(s - y) \\
&= \left(\frac{h(s, u, v) - h(y, u, v)}{s - y} - \frac{\phi(s - y)q}{s - y} + i\right)(s - y) \qquad (4) \\
&\quad + \big(h(s, u', v) - h(s, u, v)\big) + \big(h(s, u', t) - h(s, u', v)\big) \\
&\quad - (1 - \phi(s - y))q \\
&= W \cdot (s - y) + \big(h(s, u', v) - h(s, u, v)\big) \\
&\quad + \big(h(s, u', t) - h(s, u', v)\big) - \big(1 - \phi(s - y)\big)\, q,
\end{aligned}$$

where

$$W = \frac{h(s,u,v) - h(y,u,v)}{s-y} - \frac{\phi(s-y)q}{s-y} + i.$$

In view of (3) and the fact that $\phi(s-y) = 0$ for $|s-y| \leq d_K/2$, $\widetilde{\delta}_K$ can be shrunk, if necessary, so that $|\operatorname{Arg} W| \geq \widetilde{\delta}_K$ for $(y,u,v) \in K$ and $|s-y| \leq 2d_K$. If $|s-y| \geq d_K$ with $H(s,u,v)$ belonging to the compact set K', then the inequality above holds with a constant $\widetilde{\delta}_{K'}$ depending on K'. We can then find an open cover (in $\mathbf{C}^n$) $Q_1, \ldots, Q_N$ of K and an open cover $Q'_1, \ldots, Q'_{N'}$ of $K' \cap M$ with diameter smaller than $d_K/2$ and angles $\theta_{j,k}$ and constants $\delta_K > 0$ and $\delta_{K'} > 0$ such that:

Case 1. If $H(y,u,v)+(q,0) \in Q_j$ and $H(s,u,v) \in Q'_k \cap M$ with $|s-y| \leq 2d_K$, then $|\theta_{j,k}| \leq \frac{\pi}{4} - \delta_K$ and

$$\operatorname{Re}\{e^{2i\theta_{j,k}} W^2\} \leq -\delta_K |W|^2 \leq -\delta_K. \tag{5}$$

Case 2. If $H(y,u,v)+(q,0) \in Q_j$ and $H(s,u,v) \in Q'_k \cap M$ with $|s-y| \geq d_K$, then $|\theta_{j,k}| \leq \frac{\pi}{4} - \delta_{K'}$ and

$$\operatorname{Re}\{e^{2i\theta_{j,k}} W^2\} \leq -\delta_{K'} |W|^2 \leq -\delta_{K'}. \tag{6}$$

For shorthand in the next set of calculations, we let $\theta = \theta_{j,k}$. From (4), we obtain

$$\begin{aligned} \operatorname{Re}\big(e^{2i\theta}(\zeta - z)^2\big) = \operatorname{Re}\Big(& e^{2i\theta} W^2 (s-y)^2 + e^{2i\theta}\big(h(s,u',v) - h(s,u,v)\big)^2 \\ & + e^{2i\theta}\big(h(s,u',t) - h(s,u',v)\big)^2 + e^{2i\theta}(1-\phi)^2 q^2 + \text{cross terms}\Big). \end{aligned} \tag{7}$$

For $(z,w) = H(y,u,v) + (q,0) \in Q_j$ and $(\zeta,\eta) = H(s,u',t) \in Q'_k \cap M$ with $|s-y| \leq 2d_K$, we have

$$\begin{aligned} |h(s,u',v) - h(s,u,v)|^2 &\leq C_{K'} |u-u'|^2, \\ |h(s,u',t) - h(s,u',v)|^2 &\leq C_{K'} |t-v|^2, \end{aligned}$$

for some constant $C_{K'}$ depending only on K'. Therefore, in view of (5),

$$\operatorname{Re}\big(e^{2i\theta}(\zeta - z)^2\big) \leq -\frac{\delta_K}{2} |W|^2 |s-y|^2 + C_{K'}\left(|u-u'|^2 + |t-v|^2\right) + C_K q^2. \tag{8}$$

In this estimate we have handled the cross terms in the usual manner. For example, the cross term $W(s-y)(1-\phi)q$ is estimated as follows:

$$|W(s-y)(1-\phi)q| = \sqrt{\frac{\delta_K}{4}}\, W\, |s-y| \cdot \sqrt{\frac{4}{\delta_K}}\, |q|\, |1-\phi| \leq \frac{\delta_K}{4} |W|^2 |s-y|^2 + \frac{4}{\delta_K} q^2.$$

The first term on the right can be absorbed by the $-\delta_K |W|^2$ term in (5). The second term on the right contributes to $C_K q^2$ appearing in (8). Other cross terms are handled similarly.

In the case when $|s-y| \geq d_K$, we have $\phi(s-y) = 1$, so the term in (7) involving $(1-\phi)q$ is zero. Therefore the preceding analysis can be repeated (using (6)) but without the term on the right involving q; that is, if $(z,w) = H(y,u,v) + (q,0) \in Q_j$ and $(\zeta,\eta) = H(s,u',t) \in Q'_k \cap M$ with $|s-y| \geq d_K$ then

$$\text{(9)} \qquad \operatorname{Re}\big(e^{2i\theta}(\zeta - z)^2\big) \leq \frac{-\delta_{K'}}{2}|W|^2|s-y|^2 + C_{K'}\left(|u-u'|^2 + |t-v|^2\right).$$

In (8), the angle θ is θ_{jk}, satisfying $|\theta| = |\theta_{j,k}| \leq \frac{\pi}{4} - \delta_K$. In (9), the angle $\theta = \theta_{jk}$ satisfies $|\theta| \leq \frac{\pi}{4} - \delta_{K'}$ (depending on the larger set K').

We now turn our attention to the terms involving $\eta - w$ in the exponent $E_\epsilon(\zeta,\eta,\alpha,z,w)$. If $|\theta| \leq \frac{\pi}{4} - \delta$, then

$$\begin{aligned}
\text{(10)}\quad & \operatorname{Re}\left(e^{2i\theta}(\eta - w)^2\right) \\
&= \operatorname{Re}\left(e^{2i\theta}(u' - u + i(t-v))^2\right) \\
&\leq -\cos(\tfrac{\pi}{2} - 2\delta)(t-v)^2 + (u-u')^2 + 2\sin(\tfrac{\pi}{2} - 2\delta)\,|u-u'|\,|t-v| \\
&\leq -(\delta)(t-v)^2 + (u-u')^2 + \frac{2\sqrt{2}\,|u-u'|}{\sqrt{\delta}}\,\frac{|t-v|\sqrt{\delta}}{\sqrt{2}} \\
&\leq -\frac{\delta}{2}|t-v|^2 + \Big(1 + \frac{2}{\delta}\Big)|u-u'|^2 \quad (\text{using } \ 2\,|a|\,|b| \leq a^2 + b^2)
\end{aligned}$$

for all η and w. In the case $|s-y| \leq 2d_K$ the constant $\delta = \delta_K$ depends on K, whereas in the case $|s-y| \geq d_K$ the constant $\delta = \delta_{K'}$ depends on the larger set K'.

Now choose a p with $2 < p < 2/(1-\delta)$, where δ is either δ_K or $\delta_{K'}$. Then since $|\theta| \leq \frac{\pi}{4} - \delta$, we have $p|\theta| < \frac{\pi}{2}$; so there is a $\widetilde{\delta} > 0$ such that

$$\text{(11)} \qquad \text{if } \alpha = re^{i\theta}, \text{ with } r \geq 0, \quad \text{then} \quad \operatorname{Re}\alpha^p \geq \widetilde{\delta} r^p.$$

Here, $\widetilde{\delta}$ depends either on K, in the case $|s-y| \leq 2d_K$, or on K', in the case $|s-y| \geq d_K$.

In the case $|s-y| \leq 2d_K$, we can combine the estimates given in (8) (noting that $|W| \geq 1$), (10), and (11), to obtain, with $\alpha = re^{i\theta}$:

$$\begin{aligned}
&\operatorname{Re}\left(\alpha^2\Big(\frac{\zeta - z}{\epsilon}\Big)^2 + \Lambda\alpha^2\Big(\frac{w-\eta}{\epsilon}\Big)^2 - \alpha^p\right) \\
&\quad \leq -\frac{\delta_K}{2}\Big|\frac{s-y}{\epsilon}\Big|^2 r^2 + C_{K'}\left(\Big|\frac{u-u'}{\epsilon}\Big|^2 + \Big|\frac{t-v}{\epsilon}\Big|^2\right)r^2 \\
&\qquad + C_K\Big(\frac{q}{\epsilon}\Big)^2 r^2 - \frac{\delta_K}{2}\Lambda\Big|\frac{t-v}{\epsilon}\Big|^2 r^2 + \Lambda\left(1 + \frac{2}{\delta_K}\right)\Big|\frac{u-u'}{\epsilon}\Big|^2 r^2 - \widetilde{\delta}_K r^p.
\end{aligned}$$

Choosing Λ $(= \Lambda_{K'})$ large enough so that $\Lambda\delta_K/2 > C_{K'} + \delta_K/2$ we obtain (1) after relabeling δ_K to be the smaller of $\delta_K/2$ and $\widetilde{\delta}_K$ and relabeling as $C_{K'}$ the quantity $C_{K'} + \Lambda(1 + 2/\delta_K)$.

For the case when $|s-y| \geq d_K$, we use estimate (9) instead of (8) and the constant $\delta = \delta_{K'}$ now depends on the larger set K'. The only difference in the estimates above is that the right side no longer involves the term q^2 (compare (9) with (8)). Thus, (2) is obtained and this completes the proof of the main estimate given in the lemma.

5. Approximation to the identity

We restate the definition of our basic kernel:

$$E_\epsilon(\zeta,\eta,\alpha,z,w) = \exp\biggl(\alpha^2\Bigl(\frac{\zeta-z}{\epsilon}\Bigr)^2 + \alpha^2\Lambda_{K'}\Bigl(\frac{w-\eta}{\epsilon}\Bigr)^2 - \alpha^p\biggr).$$

Let $Q = Q_j$ and $Q' = Q'_k$ be the open sets and let $\theta = \theta_{j,k}$ be the angle provided by Lemma 1. Suppose f is continuous with compact support in $Q' \cap M$. Define

$$E_\epsilon(f)(z,w) = \frac{C}{\epsilon^n}\int_{(\zeta,\eta)\in M_u}\int_{\alpha\in C_\theta} f(\zeta,\eta)E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n\,d\alpha\,d\zeta d\eta,$$

where C is a normalizing constant to be chosen later, $u = \operatorname{Re} w$ and

$$M_u = \{(\zeta, u+it) \in M\} = \{\operatorname{Re}\eta = u\} \cap M$$

is a totally real n-dimensional submanifold of M parameterized by

$$(s,t) \in \mathbf{R}\times\mathbf{R}^{n-1} \mapsto (h(s,u,t)+is,\, u+it).$$

Here and below, $d\eta = d\eta_1 \wedge \cdots \wedge d\eta_{n-1}$. The domain of integration in $E_\epsilon(f)(z,w)$ is M_u, which depends on $\operatorname{Re} w = u$. Since $p > 2$ in the main estimate (Lemma 1), and since f has compact support in $Q' \cap M$, the expression $E_\epsilon(f)(z,w)$ is well-defined for $(z,w) \in Q$. This main estimate will also allow us to prove the following approximation result:

Lemma 2. *Suppose f is continuous with compact support in $Q' \cap M$. Then*

$$E_\epsilon(f) \to f \qquad as \quad \epsilon \to 0$$

uniformly on $Q \cap M$ (here, $Q = Q_j$ and $Q' = Q'_k$ are the open sets as in Lemma 1).

Proof. We analyze the exponent

$$Z = \alpha^2\Bigl(\frac{\zeta-z}{\epsilon}\Bigr)^2 + \alpha^2\Lambda_{K'}\Bigl(\frac{w-\eta}{\epsilon}\Bigr)^2 - \alpha^p$$

in the case where

$$(\zeta,\eta) = (h(s,u,t)+is,\, u+it) \in M_u$$

(in particular, $u' = u$ on M_u) and

$$(z,w) = (h(y,u,v)+iy,\, u+iv)$$

(i.e., (z,w) belongs to M and so $q=0$). We obtain

$$\begin{aligned}\zeta - z &= h(s,u,t) - h(y,u,v) + i(s-y) \in \mathbf{C},\\ \eta - w &= i(t-v) \in \mathbf{C}^{n-1}.\end{aligned}$$

With $\alpha = re^{i\theta}$, the exponent of our kernel is

$$\begin{aligned}Z &= \alpha^2\Big(\frac{\zeta - z}{\epsilon}\Big)^2 + \alpha^2\Lambda_{K'}\Big(\frac{w-\eta}{\epsilon}\Big)^2 - \alpha^p\\ &= e^{2i\theta}r^2\left(\frac{h(s,u,t)-h(y,u,v)+i(s-y)}{\epsilon}\right)^2 - \Lambda_{K'}\Big(\frac{t-v}{\epsilon}\Big)^2 e^{2i\theta}r^2 - e^{ip\theta}r^p.\end{aligned}$$

Now let

$$\hat{s} = \frac{s-y}{\epsilon} \in \mathbf{R}, \qquad \hat{t} = \frac{t-v}{\epsilon} \in \mathbf{R}^{n-1}.$$

Estimates (1) and (2) with $u = u'$ and $q = 0$ essentially reduce to the same estimate except that the δ-constant depends on K in the first estimate and on K' in the second estimate. Since $K \subset K'$, we will assume that $\delta_K \geq \delta_{K'}$. In the hat-variables, these estimates (with $u = u'$ and $q = 0$) become

$$\operatorname{Re} Z \leq -\delta_{K'}\big(|\hat{s}|^2r^2 + |\hat{t}|^2r^2 + r^p\big).$$

The function

$$e(\hat{s},\hat{t},r) = r^n \exp\big(-\delta_{K'}\left(|\hat{s}|^2r^2 + |\hat{t}|^2r^2 + r^p\right)\big)$$

is an integrable function on the set $\{(\hat{s},\hat{t},r) \in \mathbf{R}\times\mathbf{R}^{n-1}\times\mathbf{R}_+;\ r \geq 0\}$ (to see this, integrate $\hat{s}$ and $\hat{t}$ and then integrate r).

The Dominated Convergence Theorem now allows us to let $\epsilon \to 0$ in the integrand. The resulting exponent becomes

$$Z = e^{2i\theta}r^2\left(\left(\frac{\partial h(y,u,v)}{\partial y} + i\right)\hat{s} + \frac{\partial h}{\partial v}(y,u,v)\cdot\hat{t}\right)^2 - \Lambda_{K'}\hat{t}^2r^2e^{2i\theta} - r^pe^{ip\theta}.$$

In addition, $\zeta = (h(y+\epsilon\hat{s},u,v+\epsilon\hat{t})+i(y+\epsilon\hat{s}))$ converges to $h(y,u,v)+iy = z$ uniformly for $(z,w) \in K$ as $\epsilon \to 0$. Likewise, $\alpha^n\epsilon^{-n}d\zeta\wedge d\eta\wedge d\alpha$ approaches $i^{n-1}r^ne^{i(n+1)\theta}(h_y(y,u,v)+i)\,d\hat{s}\,d\hat{t}\,dr$.

Now we rewrite the first part of the exponent involving $\hat{s}$ as

$$\begin{aligned}e^{2i\theta}r^2\left(\left(\frac{\partial h(y,u,v)}{\partial y} + i\right)\hat{s} + \frac{\partial h}{\partial v}(y,u,v)\cdot\hat{t}\right)^2&\\ = -r^2\Big((-i)e^{i\theta}W\hat{s} + (-i)e^{i\theta}\frac{\partial h}{\partial v}(y,u,v)\cdot\hat{t}\Big)^2,&\end{aligned}$$

where $W = \dfrac{\partial h(y,u,v)}{\partial y} + i \in \mathbf{C}$. We have

$$\lim_{\epsilon\to 0} E_\epsilon(f)(z,w) = e^{i(n+1)\theta} Cf(z,w)$$
$$\times \int_{\hat{s},\hat{t}} \int_{r\geq 0} \exp\Big(-r^2\Big(-ie^{i\theta} W\hat{s} - ie^{i\theta}\frac{\partial h}{\partial v}(y,u,v)\,\hat{t}\Big)^2 - \Lambda_{K'}\hat{t}^2 r^2 e^{2i\theta} - r^p e^{ip\theta}\Big)$$
$$\times\, r^n i^{n-1}(h_y(y,u,v)+i)\,d\hat{s}\,d\hat{t}\,dr.$$

Using the limiting case as $s \to y$ in (5), we obtain $\mathrm{Re}\big(e^{2i\theta}(W)^2\big) < 0$. Therefore

$$\mathrm{Re}\big(-ie^{i\theta} W\hat{s}\big)^2 > 0. \tag{12}$$

The idea now is to view the $\hat{s}$-integral in $E_\epsilon(f)$ as a contour integral over the contour C given by

$$\hat{s} \mapsto z = -ie^{i\theta} W\hat{s} \quad \text{for } \hat{s} \in \mathbf{R},$$
$$\text{with } dz = -ie^{i\theta}(h_y(y,u,v)+i)\,d\hat{s}.$$

The integral is $\int_C e^{-r^2(z+c)^2}\,dz$, where $c = -ie^{i\theta} h_v(y,u,v)\,\hat{t}$. In view of (12), $e^{-r^2(z+c)^2}$ is negatively exponentially decreasing on C. Cauchy's theorem implies that this contour integral is the same as $\int_{-\infty}^{\infty} e^{-r^2(s+c)^2}\,ds$. Therefore,

$$\lim_{\epsilon\to 0} E_\epsilon(f)(z,w) = -i^{n-2} e^{in\theta} Cf(z,w)$$
$$\times \int_{s,\hat{t}} \int_{r\geq 0} \exp\Big(-r^2\Big(s - ie^{i\theta}\frac{\partial h}{\partial v}(y,u,v)\,\hat{t}\Big)^2 - \Lambda_{K'}\hat{t}^2 r^2 e^{2i\theta} - r^p e^{ip\theta}\Big)$$
$$r^n\,ds\,d\hat{t}\,dr.$$

Now we make a change of variables:

$$\widetilde{s} = s + a\,\hat{t}, \quad \text{with } a = (-i)e^{i\theta}\frac{\partial h}{\partial v}(y,u,v) \in \mathbf{C}^{n-1},$$
$$\widetilde{t} = \hat{t} \in \mathbf{R}^{n-1}.$$

The Jacobian of the derivative of this change of variables is 1. If a were a real number, we could change variables in the usual manner and obtain (after dropping the tilde)

$$\int_{s,\hat{t}} e^{-r^2(s+a\hat{t})^2 - r^2\Lambda_{K'}\hat{t}^2 e^{2i\theta}}\,ds\,d\hat{t} = \int_{s,t} e^{-r^2 s^2 - r^2\Lambda_{K'} t^2 e^{2i\theta}}\,ds\,dt.$$

Increasing $\Lambda_{K'}$ if necessary, the left side is an analytic function of a in a neighborhood of the ball $|a| \leq R$, where $R = \sup|h_v|$ over K. Therefore the preceding equality holds (by the identity theorem for analytic functions) for

complex a as well. We thus obtain

$$\lim_{\epsilon\to 0} E_\epsilon(f)(z,w)$$
$$= -i^{n-2}e^{in\theta}Cf(z,w)\int_{s,t}\int_{r\geq 0}\exp\bigl(-r^2s^2-\Lambda_{K'}t^2r^2e^{2i\theta}-r^pe^{ip\theta}\bigr)r^n\,ds\,dt\,dr.$$

We can now replace $e^{i\theta}t\in\mathbf{C}^{n-1}$ by $\widetilde{t}\in\mathbf{R}^{n-1}$ and $e^{i(n-1)\theta}dt$ by $d\widetilde{t}$. This change of variables is justified, again, by Cauchy's Theorem and the fact that $|\theta|<\frac{\pi}{4}$. The result is (dropping the tilde)

$$\lim_{\epsilon\to 0} E_\epsilon(f)(z,w)$$
$$= -i^{n-2}e^{i\theta}Cf(z,w)\int_{s,t}\int_{r\geq 0}\exp\bigl(-r^2s^2-\Lambda_{K'}t^2r^2-r^pe^{ip\theta}\bigr)r^n\,ds\,dt\,dr.$$

We change variables by letting $\hat{t}=rt\in\mathbf{R}^{n-1}$ and $\hat{s}=rs\in\mathbf{R}$ (with $r^n\,ds\,dt=d\hat{s}\,d\hat{t}$). After integrating $\hat{s}$ and $\hat{t}$, the right side becomes

$$Cf(z,w)\cdot\mathrm{Const}_{\Lambda_{K'}}\int_0^\infty\exp\bigl(-r^pe^{ip\theta}\bigr)e^{i\theta}\,dr$$

where $\mathrm{Const}_{\Lambda_{K'}}\in\mathbf{C}$ is a constant depending only on $\Lambda_{K'}$. We can view the preceding integral as an integral over the contour

$$C_\theta=\{r\mapsto z=re^{i\theta}\},\quad\text{with}\quad dz=e^{i\theta}dr.$$

Since $|\theta|<\frac{\pi}{4}$, this contour lies in the right half-plane. Using the principal branch of z^p, the contour integral becomes

$$Cf(z,w)\cdot\mathrm{Const}_{\Lambda_{K'}}\int_{z\in C_\theta}e^{-z^p}\,dz.$$

Since the integrand is analytic and rapidly decreasing at infinity (recall that $|p\theta|<\frac{\pi}{2}$), Cauchy's Theorem can be used to transform this integral into the following one over the positive real axis:

$$Cf(z,w)\cdot\mathrm{Const}_{\Lambda_{K'}}\int_0^\infty e^{-\hat{r}^p}\,d\hat{r}.$$

This is now independent of θ. Since the $\hat{r}$-integral converges, we may choose C (as a normalizing constant) so that the expression above becomes $f(z,w)$. This completes the proof of Lemma 2.

6. Fixing the slice M_u

So far, the function f is only required to be continuous (not CR). However, the domain of integration in $E_\epsilon(f)(z,w)$ is $M_u=\{\mathrm{Re}\,\eta=u\}$, which depends on $u=\mathrm{Re}\,w$. So despite the fact that the kernel is entire in (z,w), the expression $E_\epsilon(f)(z,w)$ is not holomorphic in w. The next major step is to show that if f is CR, then the domain can be fixed at a particular slice M_{u_0} (independent of w). Then, the resulting integral will be holomorphic in both

z and w. Since f is assumed to only have support on a small neighborhood of a fixed point (see Lemma 1), a localization argument with a partition of unity is needed. This adds some complicating facets since a partition of unity is not CR.

To get started with the next step, we first cover $M \cap K'$ with an open cover Q'_k $(k = 1, 2, \ldots, N')$, where each Q'_k satisfies the properties given in Lemma 1. Let ϕ_k be a partition of unity subordinate to this cover. Fix any $(z, w) \in K$ and let Q_j be an open set that satisfies Lemma 1 and is of the form $Q_j = I \times J$, where $I \subset \mathbf{R}^{n-1}$ is an open set containing $\operatorname{Re} w$ and J is an open set in $\mathbf{C} \times \mathbf{R}^{n-1}$ containing $(z, \operatorname{Im} \mathrm{w})$. Let $\theta_{j,k}$ be the angle given in Lemma 1 corresponding the open sets Q'_k in the coordinates (ζ, η) and Q_j in the coordinates (z, w). For the moment, the index j will be fixed and k will vary. Therefore, we will suppress the index j and write $Q = Q_j$ and $\theta_{Q,k} = \theta_{j,k}$. For any $u_0 \in I$, define

$$F_\epsilon^{u_0,Q}(f)(z,w) = \sum_{k=1}^{N'} \int_{(\zeta,\eta)\in M_{u_0}} \int_{\alpha \in C_{\theta_{Q,k}}} (\phi_k f)(\zeta,\eta) E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n \, d\zeta \, d\eta \, d\alpha.$$

From Lemma 2, $F_\epsilon^{u_0,Q}(f) \to f$ as $\epsilon \to 0$ uniformly on the set $Q \cap M_{u_0}$.

The next lemma contains the key step in fixing the domain of integration independently of $\operatorname{Re} w$. In this lemma, the size of $K' \cap M$ (the set on which f is CR) will be determined based on the size of the original set K. Increasing the size of K' will add terms to the sum above (i.e., additional open sets, Q'_k, and additional partition of unity functions ϕ_k may be required). However as long as K' is compact (which it will be), the sum will be finite.

Lemma 3. *Suppose K is a compact set in $\mathbf{C}^n$; then $K' \subset \mathbf{C}^n$ can be chosen large enough (containing K) so that the following holds: suppose f is* CR *on $K' \cap M$. Let $Q = I \times J$ be the open set described above. There exists $\Delta_{K'} > 0$ such that if the diameter of I is less than $\Delta_{K'}$ and if $u_0, u_1 \in I$, then $\left|(F_\epsilon^{u_1,Q} - F_\epsilon^{u_0,Q})(f)(z,w)\right| \to 0$ as $\epsilon \to 0$ uniformly for $(z,w) \in Q = I \times J$.*

Proof. Assume that K' is a closed ball of radius $R' + 1$ in $\mathbf{C}^n$. Since f is CR on $K' \cap M$, we can multiply f by an appropriate cut-off function and assume that f has compact support in $K' \cap M$ and is $\bar{\partial}$-closed on the intersection of M with the ball of radius R'.

M_{u_0} and M_{u_1} are totally real n-dimensional slices of M that naturally form the boundary of a real $n+1$-dimensional submanifold $\widetilde{M}_{0,1} \subset M$. In fact, just connect u_0 and u_1 by a real line segment and let $\widetilde{M}_{0,1}$ be the graph of h over the $n+1$ real-dimensional strip spanned by this line segment and the n-dimensional subspace $\{\operatorname{Re} \eta = u_0\}$ of $\mathbf{R}^{2n-1}$.

Since E_ϵ is holomorphic, Stokes' Theorem implies

$$(F_\epsilon^{u_1,Q} - F_\epsilon^{u_0,Q})(f)(z,w) = S_1 + S_2$$

with

$$S_1 = \sum_{k=1}^{N'} \int_{(\zeta,\eta)\in \widetilde{M}_{01}} \int_{\alpha\in C_{\theta_{Q,k}}} (\bar\partial\phi_k f)(\zeta,\eta) E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n d\zeta\, d\eta\, d\alpha, \tag{13}$$

$$S_2 = \sum_{k=1}^{N'} \int_{(\zeta,\eta)\in \widetilde{M}_{01}} \int_{\alpha\in C_{\theta_{Q,k}}} (\phi_k\bar\partial f)(\zeta,\eta) E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n d\zeta\, d\eta\, d\alpha. \tag{14}$$

We first show that the sum in (14) converges to zero as $\epsilon \to 0$. The point $(\zeta,\eta) = (h(s,u',t) + is, u' + it)$ must lie in the support of $\bar\partial f$, which lies outside the ball of radius $R' \approx \text{radius}(K')$. Also note that u' belongs to the line segment connecting u_0 and u_1, which in turn belongs to K (or rather, the projection of K onto the u-axis). The point $(z,w) = (h(y,u,v)+q+iy, u+iv)$ lies in the smaller compact set K. We are at liberty to take R' as large as we please (relative to the diameter of K) in order to make the integrand in (14) converge to zero as $\epsilon \to 0$. Since (z,w) and u, u' belong to K, we may choose R' large enough so that either

$$|s-y| \geq \frac{R'}{2} \quad \text{or} \quad |t-v| \geq \frac{R'}{2} \tag{15}$$

for all (ζ,η) belonging to the support of $\bar\partial f$.

We have two cases to consider on the exponent

$$Z = \alpha^2\Big(\frac{\zeta - z}{\epsilon}\Big)^2 + \alpha^2\Lambda_{K'}\Big(\frac{w-\eta}{\epsilon}\Big)^2 - \alpha^p.$$

The first is $|s-y| \leq 2d_K$, in which case we repeat the estimate (1):

$$\operatorname{Re} Z \leq -\delta_K\left(\Big|\frac{s-y}{\epsilon}\Big|^2 r^2 + \Big|\frac{t-v}{\epsilon}\Big|^2 r^2 + r^p\right) + C_{K'}\Big|\frac{u-u'}{\epsilon}\Big|^2 r^2 + C_K\Big(\frac{q}{\epsilon}\Big)^2 r^2.$$

Now u, u' belong to I and we will require the diameter of I to be less than $\Delta_{K'}$ (where $\Delta_{K'}$ will be chosen below). In view of the inequality above and (15), we have

$$\operatorname{Re} Z \leq \frac{-\delta_K (R'/2)^2 r^2 + C_{K'}\Delta_{K'}^2 r^2 + C_K q^2 r^2}{\epsilon^2}.$$

Choose R' large enough that $-\delta_K(R'/2)^2 + C_K q^2 < -2$ for all q in K. This choice of R' fixes the constant $C_{K'}$. Now choose $\Delta_{K'}$ small enough that $C_{K'}\Delta_{K'}^2 < 1$. Then the real part of exponent is at most $-r^2/(\epsilon^2)$.

In the case where $|s-y| \geq d_K$, (2) implies

$$\operatorname{Re} Z \leq -\delta_{K'}\left(\Big|\frac{s-y}{\epsilon}\Big|^2 r^2 + \Big|\frac{t-v}{\epsilon}\Big|^2 r^2 + r^p\right) + C_{K'}\Big|\frac{u-u'}{\epsilon}\Big|^2 r^2.$$

In view of (15)

$$\operatorname{Re} Z \leq -\frac{\delta_{K'} R'^2 r^2}{4\epsilon^2} + \frac{C_{K'}\Delta_{K'}^2 r^2}{\epsilon^2}.$$

By shrinking $\Delta_{K'}$ we can arrange

$$\operatorname{Re} Z \leq -\frac{\delta_{K'} R'^2 r^2}{8\epsilon^2}.$$

In either case ($|s-y| \geq d_K$ or $|s-y| \leq 2d_K$), the integrand of each term in (14) is dominated by Ce^{-r^2/ϵ^2} or $Ce^{-\delta_{K'} R'^2 r^2/(8\epsilon^2)}$ provided $u, u' \in I$ and the diameter of I is less than $\Delta_{K'}$.

Since $\int_0^\infty e^{-r^2/\epsilon^2}\, dr \to 0$ and $\int_0^\infty e^{-\delta_{K'} R'^2 r^2/(8\epsilon^2)}\, dr \to 0$ as $\epsilon \to 0$, we conclude that the sum in (14) converges to zero.

Now we examine the sum S_1 in (13), which we restate as

$$\sum_{k=1}^{N'} \int_{(\zeta,\eta)\in \widetilde{M}_{01}} \int_{\alpha \in C_{\theta_{Q,k}}} (\bar{\partial}\phi_k f)(\zeta,\eta) E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n d\zeta\, d\eta\, d\alpha.$$

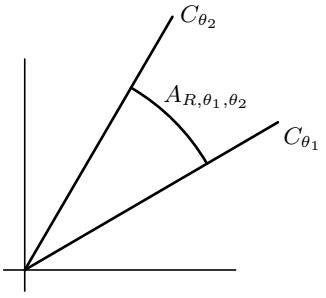
We have already determined the compact set K' (the size of which had to be chosen large enough to make the term in (14) converge to zero as $\epsilon \to 0$). Thus, all the constants in our integral kernels (i.e., $\Lambda_{K'}$ and $p = p_{K'} > 2$) are now determined.

Note that the kernel $E_\epsilon(\zeta,\eta,\alpha,z,w)$ in the sum immediately above is the same for all k. The only terms that appear to vary with k are the cutoff functions ϕ_k and the contours $C_{\theta_{Q,k}}$. The following lemma states that the integral of the kernel is independent of this contour.

Sublemma 1. *Let Q'_1 and Q'_2 be intersecting open sets from our cover. Let θ_1 and θ_2 be any two angles that satisfy the requirements of Lemma 1 relative to Q'_1 and Q'_2 (in particular, $|\theta_1|, |\theta_2| < \frac{\pi}{4} - \delta_{K'}$). Then*

$$\int_{\alpha \in C_{\theta_1}} E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n d\alpha = \int_{\alpha \in C_{\theta_2}} E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n d\alpha. \tag{16}$$

To prove this, recall that C_θ is a ray in the complex plane that makes an angle θ with the positive real axis. For $R > 0$, let A_{R,θ_1,θ_2} be the arc of a circle of radius R that lies between C_{θ_1} and C_{θ_2}:



Since $E_\epsilon(\zeta,\eta,\alpha,z,w)$ is holomorphic in α, Equation (16) will follow from Cauchy's Theorem provided we show that

$$\sup_{\alpha\in A_{R,\theta_1,\theta_2}} R^{n+1}\left|E_\epsilon(\zeta,\eta,\alpha,z,w)\right| \to 0 \quad \text{as } R\to\infty. \tag{17}$$

We restate the kernel:

$$E_\epsilon(\zeta,\eta,\alpha,z,w) = \exp\biggl(\alpha^2\Bigl(\frac{\zeta-z}{\epsilon}\Bigr)^2 + \alpha^2\Lambda_{K'}\Bigl(\frac{w-\eta}{\epsilon}\Bigr)^2 - \alpha^p\biggr).$$

Keep in mind that ϵ, ζ, η, z and w are fixed. Since $p>2$, the dominant term in the exponent is α^p as $|\alpha| = R\to\infty$. Let $\alpha = Re^{i\phi}$ be an arbitrary point on the arc A_{R,θ_1,θ_2}, so $\theta_1\le\phi\le\theta_2$. Since $|p\phi|\le p\,|\max(\theta_1,\theta_2)| < \frac{\pi}{2}$ by the choice of p, there exists $\delta>0$ such that $-\operatorname{Re}\alpha^p \le -\delta R^p$ for $\alpha\in A_{R,\theta_1,\theta_2}$. Equation (17) now follows since $R^{n+1}e^{-\delta R^p}\to 0$ as $R\to\infty$. This concludes the proof of the sublemma.

Using the fact that $\sum_j \phi_j = 1$, the sum in (13) can be rewritten as

$$\sum_{j=1}^{N'}\sum_{k=1}^{N'}\int_{(\zeta,\eta)\in\widetilde{M}_{01}}\int_{\alpha\in C_{\theta_{Q,k}}} \phi_j(\bar\partial\phi_k f)(\zeta,\eta)E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n d\zeta\,d\eta\,d\alpha. \tag{18}$$

The support of $\phi_j\bar\partial\phi_k$ is contained in $Q'_j\cap Q'_k\cap M$. Using Equation (16), the integral over $C_{\theta_{Q,k}}$ can be replaced by the integral over $C_{\theta_{Q,j}}$. Therefore (18) becomes

$$\sum_{j,k=1}^{N'}\int_{(\zeta,\eta)\in\widetilde{M}_{01}}\int_{\alpha\in C_{\theta_{Q,j}}} \phi_j(\bar\partial\phi_k f)(\zeta,\eta)E_\epsilon(\zeta,\eta,\alpha,z,w)\alpha^n d\zeta\,d\eta\,d\alpha.$$

Summing out k we see that this term is zero because $\sum_k\bar\partial\phi_k = 0$ on $M\cap K'$. Thus the term in (13) is zero. Since the terms in (14) converge to zero as $\epsilon\to 0$, the proof of Lemma 3 is now complete.

Lemmas 2 and 3 imply:

Corollary 1. *With the notation of Lemma* 3*, suppose f is* CR *on $K'\cap M$ and suppose that $Q = I\times J$, where the diameter of I is less than $\Delta_{K'}$. If u_0 belongs to I, then*

$$F_\epsilon^{u_0,Q}(f)(z,w)\to f \quad \text{as } \epsilon\to 0,$$

uniformly for $(z,w)\in Q\cap M$.

7. Globalization

Now $F_\epsilon^{u_0,Q}(f)$ is holomorphic on Q and $F_\epsilon^{u_0,Q}(f)\to f$ on $Q\cap M$. The set Q is an appropriately small open set about an arbitrary point (z,w) in K. Our next and final step is to piece together these locally defined holomorphic functions into a sequence of functions that are globally holomorphic on K

with the corresponding convergence to f on $K \cap M$. This will require solving a $\bar{\partial}$ problem on K with estimates.

To get started, we will assume K is a ball (in particular, K is a convex domain and so we can solve $\bar{\partial}$ on K). We cover K with open sets Q_j $(j = 1, \ldots, N)$ of the form $Q_j = I_j \times J_j$ where I_j is an open set in $\mathbf{R}^{n-1}$ (with coordinates u) and J_j is an open set in $\mathbf{R}^{n+1} = \mathbf{R} \times \mathbf{R} \times \mathbf{R}^{n-1}$ (with coordinates $x \in \mathbf{R}$, $y \in \mathbf{R}$, $v \in \mathbf{R}^{n-1}$). We assume that Q_j is small enough to satisfy the requirements of Lemma 3 and that the diameter of each I_j is smaller than $\Delta_{K'}$ (from Lemma 3). Choose points $u_j \in I_j$, $j = 1 \ldots N$. We let ψ_j be a partition of unity for K subordinate to the cover Q_j. We define

$$
\begin{aligned}
(19) \quad F_\epsilon(f)(z,w) &= \left(\sum_j \psi_j F_\epsilon^{u_j, Q_j}(f)(z,w) \right) - v_\epsilon(z,w) \\
&= \sum_{j,k} \psi_j(z,w) \int_{(\zeta,\eta) \in M_{u_j}} \int_{\alpha \in C_{\theta_{Q_j,k}}} (\phi_k f)(\zeta,\eta) \\
&\qquad \times E_\epsilon(\zeta,\eta,\alpha,z,w) \alpha^n d\zeta \, d\eta \, d\alpha - v_\epsilon(z,w),
\end{aligned}
$$

where v_ϵ will be chosen so that $F_\epsilon(f)$ is holomorphic in K. We will also show that $|v_\epsilon|$ converges to zero uniformly on K as $\epsilon \to 0$. Then, Corollary 1 will imply that $F_\epsilon(f) \to f$ uniformly on $M \cap K$, as desired.

In order to arrange that $F_\epsilon(f)$ is holomorphic on K, we must require

$$
\begin{aligned}
(20) \qquad \bar{\partial} v_\epsilon(z,w) &= \sum_j \bar{\partial}\{\psi_j\} F^{u_j, Q_j}(f)(z,w) \\
&= \sum_l \psi_l \sum_j \bar{\partial}\{\psi_j\} F_\epsilon^{u_j, Q_j}(f)(z,w),
\end{aligned}
$$

using $\sum_l \psi_l = 1$ on K. If $\psi_l \bar{\partial}\{\psi_j\}$ is nonzero, Q_l and Q_j must overlap. We wish to change $F_\epsilon^{u_j, Q_j}$ to $F_\epsilon^{u_l, Q_l}$ in the sum above. Changing Q_j to Q_l involves changing the angle of the contour from $C_{\theta_{Q_j,k}}$ to $C_{\theta_{Q_l,k}}$ on the second line of (19), which Sublemma 1 allows us to do. Changing u_j to u_l is allowed by Lemma 3 but with a resulting error that tends to zero with ϵ. Therefore, (20) becomes

$$
\bar{\partial} v_\epsilon(z,w) = \sum_l \psi_l \sum_j \bar{\partial}\{\psi_j\} F_\epsilon^{u_l, Q_l}(f)(z,w) + O(\epsilon),
$$

where $O(\epsilon)$ stands for terms that converge to zero uniformly on K as $\epsilon \to 0$. Summing out j and using the fact that $\sum_j \bar{\partial}\{\psi_j\} = 0$, we conclude that $\bar{\partial} v_\epsilon = O(\epsilon)$ on K. Solving this $\bar{\partial}$ equation with sup-norm estimates on K, we can find a solution with $|v_\epsilon| = O(\epsilon)$. Now returning to (19) and using Corollary 1, we conclude that $F_\epsilon(f)$ is analytic on K and converges uniformly to f on $M \cap K$. Since K is convex, it is also polynomially convex.

Therefore, there is also a sequence of polynomials that converge uniformly on $K \cap M$ to f. This concludes the proof of our main theorem.

References

[BT] M.S. Baouendi and F. Trèves, *A property of the functions and distributions annihilated by a locally integrable system of complex vector fields*, Ann. of Math. (2), **113** (1981), 387–421, MR 0607899 (82f:35057), Zbl 0491.35036.

[B] A. Boggess, *CR extension for L^p CR functions on a quadric submanifold of $\mathbf{C}^n$*, Pacific J. Math., **201**(**1**) (2001), 1–18, MR 1867889 (2002m:32051).

[Ch] E.M. Chirka, *On the uniform approximation on totally real sets in $\mathbf{C}^n$*, in 'Complex methods in approximation theory' (Almeria, 1995), 23–31, Monogr. Cienc. Tecnol., **2**, Univ. Almeria, Almeria, 1997, 23–31, MR 1634170 (99k:32021), Zbl 0942.32014.

[DG] R. Dwilewicz and P.M. Gauthier, *Global holomorphic approximations of CR functions on CR manifolds*, Complex Variables Theory Appl., **4** (1985), 377–391, MR 0858919 (88b:32041), Zbl 0548.32011.

[HW] F.R. Harvey and R.O. Wells, Jr., *Holomorphic approximation and hyperfunction theory on a C^1 totally real submanifold of a complex manifold*, Math. Ann., **197** (1972), 287–318, MR 0310278 (46 #9379), Zbl 0246.32019.

[N] J. Nunemacher, *Approximation theory on totally real submanifolds*, Math. Ann., **224** (1976), 129–141, MR 0422684 (54 #10670), Zbl 0333.32015.

Received June 16, 2003. The second author was partially supported by Polish Committee for Scientific Research Grant KBN 2 P03A 044 15.

Department of Mathematics
Texas A&M University
College Station TX 77843-3368
E-mail address: boggess@math.tamu.edu

Department of mathematics and statistics
University of Missouri
Rolla MO 65409
E-mail address: romand@umr.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

MAPS CHARACTERIZED BY ACTION ON ZERO PRODUCTS

M.A. Chebotar, Wen-Fong Ke and Pjek-Hwee Lee

For prime rings containing nontrivial idempotents, we describe the bijective additive maps which preserve zero products. Also, we describe the additive maps which behave like derivations when acting on zero products.

1. Introduction

In the last decade considerable works have been done concerning local properties of maps; see [**1**], [**7**], and [**9**]–[**32**], where other references can be found. The goal of this paper is to show that automorphisms and derivations of prime rings with nontrivial idempotents can be "almost" determined by the action on the zero-product elements.

Our first theorem generalizes a similar result of Wong [**34**, Corollary D] for simple algebras with nontrivial idempotents, as well as some other results obtained for operator algebras [**1**, **13**, **16**, **33**].

Let R be a prime ring. The definitions and some basic properties of the maximal right quotient ring $Q(R)$ and extended centroid $C(R)$ of R can be found in [**6**]. Recall that an element x in $Q(R)$ is said to be algebraic of degree $\leq n$ over $C(R)$ if there exist $c_0, c_1, c_2, \ldots, c_{n-1} \in C(R)$ such that $\sum_{i=0}^{n-1} c_i x^i + x^n = 0$. By $\deg R \geq n$ we mean that there exists an element x in R that is not algebraic of degree $\leq n-1$ over $C(R)$. The condition that $\deg R \geq n$ is equivalent to that R can not be embedded in the ring of $(n-1) \times (n-1)$ matrices over a field.

Theorem 1. *Let A and B be prime rings and $\theta : A \to B$ a bijective additive map such that $\theta(x)\theta(y) = 0$ for all $x, y \in A$ with $xy = 0$. Suppose that the maximal right quotient ring $Q(A)$ of A contains a nontrivial idempotent e such that $eA \cup Ae \subseteq A$.*

(i) *If $1 \in A$, then $\theta(xy) = \lambda\theta(x)\theta(y)$ for all $x, y \in A$, where $\lambda = 1/\theta(1)$ and $\theta(1) \in Z(B)$, the center of B. In particular, if $\theta(1) = 1$, then θ is a ring isomorphism from A onto B.*

(ii) *If $\deg B \geq 3$, then there exists $\lambda \in C(B)$, the extended centroid of B, such that $\theta(xy) = \lambda\theta(x)\theta(y)$ for all $x, y \in A$.*

It is clear that Theorem 1 can not be extended to arbitrary prime rings, since these rings may not have "enough" zero-divisors. The condition that the idempotent e be an element of $Q(A)$ (instead of A) enables us to consider matrix rings over arbitrary prime rings (not necessarily with units).

Our second result is an analog of Theorem 1 for derivations. In particular, it generalizes [**19**, Theorem 6].

Theorem 2. *Let A be a prime ring with center Z, maximal right quotient ring Q and extended centroid C. Let $\delta : A \to A$ be an additive map such that $\delta(x)y + x\delta(y) = 0$ for all $x, y \in A$ with $xy = 0$. Suppose that Q contains a nontrivial idempotent e such that $eA \cup Ae \subseteq A$.*

(i) *If $1 \in A$, then $\delta(xy) = \delta(x)y + x\delta(y) - \lambda xy$ for all $x, y \in A$, where $\lambda = \delta(1) \in Z$. In particular, if $\delta(1) = 0$, then δ is a derivation on A.*

(ii) *If $\deg A \geq 3$, there exists $\lambda \in C$ such that $\delta(xy) = \delta(x)y + x\delta(y) - \lambda xy$ for all $x, y \in A$.*

2. Isomorphisms

We start with a key result of the paper.

Theorem 3. *Let A and B be prime rings and $\theta : A \to B$ a bijective additive map such that $\theta(x)\theta(y) = 0$ for all $x, y \in A$ with $xy = 0$. Suppose that the maximal right quotient ring $Q(A)$ of A contains a nontrivial idempotent e such that $eA \cup Ae \subseteq A$. Then $\theta(x)\theta(yz) = \theta(xy)\theta(z)$ for all $x, y, z \in A$. Moreover, if A contains the unity 1, then:*

(i) *$\theta(1)$ lies in the center $Z(B)$ of B.*

(ii) *$\theta(1)\theta(xy) = \theta(x)\theta(y)$ for all $x, y \in A$. In particular, if $\theta(1) = 1$, then θ is a ring isomorphism from A onto B.*

(iii) *θ preserves commutativity, that is, $\theta(x)\theta(y) = \theta(y)\theta(x)$ for all $x, y \in A$ with $xy = yx$.*

Proof. Set $f = 1 - e$. Then f is a nontrivial idempotent in $Q(A)$ such that $e + f = 1$, $ef = fe = 0$ and $fA \cup Af \subseteq A$. Since θ is additive and $y = eye + eyf + fye + fyf$ for all $y \in A$, it suffices to show that the identity $\theta(x)\theta(yz) = \theta(xy)\theta(z)$ holds for y in eAe, eAf, fAe and fAf, respectively.

Let $x, z \in A$. Since θ preserves zero products, $(xe)(z - ez) = 0$ implies that $\theta(xe)\theta(z - ez) = 0$ and hence

$$\theta(xe)\theta(z) = \theta(xe)\theta(ez).$$

Similarly, it follows from $(x - ex)(ez) = 0$ that

$$\theta(x)\theta(ez) = \theta(xe)\theta(ez).$$

Thus we have

$$\theta(x)\theta(ez) = \theta(xe)\theta(ez) = \theta(xe)\theta(z) \quad \text{for all } x, z \in A. \tag{1}$$

By the symmetry of e and f, we also have

$$\theta(x)\theta(fz) = \theta(xf)\theta(fz) = \theta(xf)\theta(z) \quad \text{for all } x, z \in A. \tag{2}$$

Note that

$$(xe + xeyf)(eyfz - fz) = 0 \quad \text{for all } x, y, z \in A,$$

so

$$\theta(xe + xeyf)\theta(eyfz - fz) = 0 \quad \text{for all } x, y, z \in A.$$

Since $\theta(xe)\theta(fz) = 0$ and $\theta(xeyf)\theta(eyfz) = 0$, this results in

$$\theta(xe)\theta(eyfz) = \theta(xeyf)\theta(fz) \quad \text{for all } x, y, z \in A,$$

and hence

$$\theta(x)\theta(eyfz) = \theta(xeyf)\theta(z) \quad \text{for all } x, y, z \in A \tag{3}$$

in light of (1) and (2). By the symmetry of e and f, we also have

$$\theta(x)\theta(fyez) = \theta(xfye)\theta(z) \quad \text{for all } x, y, z \in A. \tag{4}$$

Thus it remains to show that

$$\theta(x)\theta(eyez) = \theta(xeye)\theta(z) \quad \text{for all } x, y, z \in A \tag{5}$$

and

$$\theta(x)\theta(fyfz) = \theta(xfyf)\theta(z) \quad \text{for all } x, y, z \in A. \tag{6}$$

Applying (1), (2), (3) and (4) we shall rewrite the product

$$\theta(xeyezeufv_1)\theta(fv_2)\theta(fv_3)\theta(ew)$$

in two ways, via the following sequences of steps (read down each column; each entry can be seen to be equal to the one immediately above):

$\theta(xeyezeufv_1)\theta(fv_2)\theta(fv_3)\theta(ew)$	$\theta(xeyezeufv_1)\theta(fv_2)\theta(fv_3)\theta(ew)$
$\theta(x(eyezeufv_1f))\theta(fv_2)\theta(fv_3)\theta(ew)$	$\theta(xeye(ezeufv_1f))\theta(fv_2)\theta(fv_3)\theta(ew)$
$\theta(x)\theta(eyezeufv_1fv_2)\theta(fv_3)\theta(ew)$	$\theta(xeye)\theta(ezeufv_1fv_2)\theta(fv_3)\theta(ew)$
$\theta(x)\theta(eyeze(eufv_1fv_2f))\theta(fv_3)\theta(ew)$	$\theta(xeye)\theta(ze(eufv_1fv_2f))\theta(fv_3)\theta(ew)$
$\theta(x)\theta(eyeze)\theta(eufv_1fv_2fv_3)\theta(ew)$	$\theta(xeye)\theta(ze)\theta(eufv_1fv_2fv_3)\theta(ew)$
$\theta(x)\theta(eyeze)\theta(u(fv_1fv_2fv_3e))\theta(ew)$	$\theta(xeye)\theta(ze)\theta(u(fv_1fv_2fv_3e))\theta(ew)$
$\theta(x)\theta(eyeze)\theta(u)\theta(fv_1fv_2fv_3ew)$	$\theta(xeye)\theta(ze)\theta(u)\theta(fv_1fv_2fv_3ew)$

Comparing the two expressions on the last line, we get

$$\big(\theta(x)\theta(eyeze) - \theta(xeye)\theta(ze)\big)\theta(u)\theta(fv_1fv_2fv_3ew) = 0$$

for all $x, y, z, u, v_i, w \in A$. Since A and B are prime and θ is bijective, we obtain

$$\theta(x)\theta(eyeze) = \theta(xeye)\theta(ze) \quad \text{for all } x, y, z \in A. \tag{7}$$

Similarly we express the product

$$\theta(xeyezfufv_1)\theta(fv_2)\theta(ev_3)\theta(ew)$$

in two other ways:

$\theta(xeyezfufv_1)\theta(fv_2)\theta(ev_3)\theta(ew)$ $\quad$ $\theta(xeyezfufv_1)\theta(fv_2)\theta(ev_3)\theta(ew)$

$\theta(x(eyezfufv_1f))\theta(fv_2)\theta(ev_3)\theta(ew)$ $\quad$ $\theta(xeye(ezfufv_1f))\theta(fv_2)\theta(ev_3)\theta(ew)$

$\theta(x)\theta(eyezfufv_1fv_2)\theta(ev_3)\theta(ew)$ $\quad$ $\theta(xeye)\theta(ezfufv_1fv_2)\theta(ev_3)\theta(ew)$

$\theta(x)\theta(eyezf(fufv_1fv_2e))\theta(ev_3)\theta(ew)$ $\quad$ $\theta(xeye)\theta(zf(fufv_1fv_2e))\theta(ev_3)\theta(ew)$

$\theta(x)\theta(eyezf)\theta(fufv_1fv_2ev_3)\theta(ew)$ $\quad$ $\theta(xeye)\theta(zf)\theta(fufv_1fv_2ev_3)\theta(ew)$

$\theta(x)\theta(eyezf)\theta(u(fv_1fv_2ev_3e))\theta(ew)$ $\quad$ $\theta(xeye)\theta(zf)\theta(u(fv_1fv_2ev_3e))\theta(ew)$

$\theta(x)\theta(eyezf)\theta(u)\theta(fv_1fv_2ev_3ew)$ $\quad$ $\theta(xeye)\theta(zf)\theta(u)\theta(fv_1fv_2ev_3ew)$

Comparing both expressions, we get

$$\big(\theta(x)\theta(eyezf) - \theta(xeye)\theta(zf)\big)\theta(u)\theta(fv_1fv_2ev_3ew) = 0$$

for all $x, y, z, u, v_i, w \in A$. Since A and B are prime and θ is bijective, we obtain

$$\theta(x)\theta(eyezf) = \theta(xeye)\theta(zf) \quad \text{for all } x, y, z \in A. \tag{8}$$

Then (5) follows immediately from the identities (7) and (8). By the symmetry of e and f, we obtain (6) too. Therefore,

$$\theta(x)\theta(yz) = \theta(xy)\theta(z) \quad \text{for all } x, y, z \in A. \tag{9}$$

Suppose that A contains the unity 1. Setting $x = z = 1$ in (9), we have

$$\theta(1)\theta(y) = \theta(y)\theta(1)$$

for all $y \in A$. Since θ is surjective, $\theta(1)$ lies in the center of B. This establishes statement (i) of Theorem 3.

Setting $z = 1$ in (9), we get

$$\theta(x)\theta(y) = \theta(xy)\theta(1) = \theta(1)\theta(xy)$$

for all $x, y \in A$. In particular, if $\theta(1) = 1$, then θ is a ring isomorphism from A onto B. This establishes (ii).

Finally, by (ii) we have

$$\theta(x)\theta(y) - \theta(y)\theta(x) = \theta(1)(\theta(xy) - \theta(yx)) = \theta(1)\theta(xy - yx)$$

for all $x, y \in A$. Then (iii) follows immediately. □

In view of the preceding theorem, we see that the zero-product preserving map θ satisfies the functional identity

$$\theta(x)\theta(yz) = \theta(xy)\theta(z) \quad \text{for all } x, y, z \in A.$$

This enables us to apply the recently developed theory of functional identities. Instead of introducing complicated definitions and notations, we shall

present some special cases of the results in [**3**, **4**]. The first one follows from [**3**, Theorem 2.4] and [**4**, Theorem 1.2].

Lemma 4. *Let R be a prime ring with maximal right quotient ring Q and extended centroid C such that $\deg R \geq 3$. Let S be a set, $\theta : S \to R$ a surjective map and $M : S \times S \to Q$ a map. Suppose that*

$$\alpha_1\theta(x)M(y,z) + \alpha_2\theta(y)M(x,z) + \alpha_3\theta(z)M(x,y) + \beta_1 M(y,z)\theta(x) + \beta_2 M(x,z)\theta(y) + \beta_3 M(x,y)\theta(z) = 0$$

for all $x, y, z \in S$, where the α_i and β_i are constants in C, not all zero. Then there exist $\lambda_1, \lambda_2 \in C$, $\mu_1, \mu_2 : S \to C$ and $\nu : S \times S \to C$ such that

$$M(x,y) = \lambda_1\theta(x)\theta(y) + \lambda_2\theta(y)\theta(x) + \mu_1(x)\theta(y) + \mu_2(y)\theta(x) + \nu(x,y)$$

for all $x, y \in S$.

The second one follows from [**3**, Theorem 2.4] and [**4**, Theorem 1.1].

Lemma 5. *Let R be a prime ring with maximal right quotient ring Q and extended centroid C, such that $\deg R \geq 3$. Let S be a set and $\theta : S \to R$ a surjective map. Suppose that*

$$\sum_{\sigma\in\mathrm{Sym}(3)} \alpha_\sigma\theta(x_{\sigma(1)})\theta(x_{\sigma(2)})\theta(x_{\sigma(3)}) + \sum_{\sigma\in\mathrm{Sym}(3)} \beta_\sigma(x_{\sigma(1)})\theta(x_{\sigma(2)})\theta(x_{\sigma(3)}) + \gamma_1(x_2,x_3)\theta(x_1) + \gamma_2(x_1,x_3)\theta(x_2) + \gamma_3(x_1,x_2)\theta(x_3) = 0$$

for all $x_1, x_2, x_3 \in S$, where $\mathrm{Sym}(3)$ is the symmetric group on 3 letters, the α_σ are constants in C, the β_σ are maps $S \to C$ and the γ_i are maps $S \times S \to C$. Then the constants α_σ and the maps β_σ and γ_i are all zero.

With these results at hand, we are ready to prove our first main theorem.

Proof of Theorem 1. Since (i) follows from Theorem 3, it remains to prove (ii).

Define a map $M : A \times A \to B$ by $M(x,y) = \theta(xy)$ for $x, y \in A$. By Theorem 3, we have

$$\theta(x)M(y,z) - M(x,y)\theta(z) = 0 \quad \text{for all } x, y, z \in A. \tag{10}$$

Then it follows from Lemma 4 that there exist $\lambda_1, \lambda_2 \in C$, $\mu_1, \mu_2 : A \to C$ and $\nu : A \times A \to C$ such that

$$M(x,y) = \lambda_1\theta(x)\theta(y) + \lambda_2\theta(y)\theta(x) + \mu_1(x)\theta(y) + \mu_2(y)\theta(x) + \nu(x,y)$$

for all $x, y \in A$. Substituting this into (10), we obtain

$$\lambda_2\theta(x)\theta(z)\theta(y) - \lambda_2\theta(y)\theta(x)\theta(z) + \mu_2(z)\theta(x)\theta(y) + \big(\mu_1(y) - \mu_2(y)\big)\theta(x)\theta(z) - \mu_1(x)\theta(y)\theta(z) + \nu(y,z)\theta(x) - \nu(x,y)\theta(z) = 0,$$

for all $x, y, z \in A$. By Lemma 5, the constant λ_2 and the maps μ_1, μ_2 and ν are all zero. In other words, $M(x, y) = \theta(xy) = \lambda_1\theta(x)\theta(y)$ for all $x, y \in A$. Thus the proof is complete. □

3. Derivations

Next we prove a result analogous to Theorem 3. The idea is essentially the same as in the proof of Theorem 3, although the computations are a bit more complicated.

Theorem 6. *Let A be a prime ring with maximal right quotient ring Q and $\delta : A \to A$ an additive map such that $\delta(x)y + x\delta(y) = 0$ for all $x, y \in A$ with $xy = 0$. Suppose that Q contains a nontrivial idempotent e such that $eA \cup Ae \subseteq A$. Then $\delta(x)yz + x\delta(yz) = \delta(xy)z + xy\delta(z)$ for all $x, y, z \in A$. Moreover, if A contains the unity 1, then*

$$\delta(xy) = \delta(x)y + x\delta(y) - \lambda xy,$$

where $\lambda = \delta(1)$ is a central element of A. In particular, if $\delta(1) = 0$, then δ is a derivation on A.

Proof. As before, we set $f = 1 - e$. Then f is a nontrivial idempotent in Q such that $e + f = 1$, $ef = fe = 0$ and $fA \cup Af \subseteq A$. Since δ is additive and $y = eye + eyf + fye + fyf$ for all $y \in A$, it suffices to show that the identity $\delta(x)yz + x\delta(yz) = \delta(xy)z + xy\delta(z)$ holds for y in eAe, eAf, fAe and fAf respectively.

Let $x, z \in A$. Since $(xe)(z-ez) = 0$, we have $\delta(xe)(z-ez)+xe\delta(z-ez) = 0$ by assumption and hence

$$\delta(xe)z + xe\delta(z) = \delta(xe)ez + xe\delta(ez).$$

Similarly, it follows from $(x - xe)(ez) = 0$ that

$$\delta(x)ez + x\delta(ez) = \delta(xe)ez + xe\delta(ez).$$

Thus

(11)
$$\delta(x)ez + x\delta(ez) = \delta(xe)z + xe\delta(z) = \delta(xe)ez + xe\delta(ez) \quad \text{for all } x, z \in A.$$

By the symmetry of e and f, we also have

(12)
$$\delta(x)fz + x\delta(fz) = \delta(xf)z + xf\delta(z) = \delta(xf)fz + xf\delta(fz) \quad \text{for all } x, z \in A.$$

Note that for $x, y, z \in A$, we have

$$(xe)(fz) = 0,$$
$$(xeyf)(eyfz) = 0,$$
$$(xe + xeyf)(eyfz - fz) = 0,$$

so

$$\delta(xe)fz + xe\delta(fz) = 0,$$
$$\delta(xeyf)eyfz + xeyf\delta(eyfz) = 0,$$
$$\delta(xe + xeyf)(eyfz - fz) + (xe + xeyf)\delta(eyfz - fz) = 0.$$

Combining these three identities, we get

$$\delta(xe)eyfz + xe\delta(eyfz) = \delta(xeyf)fz + xeyf\delta(fz),$$

and hence

$$\delta(x)eyfz + x\delta(eyfz) = \delta(xeyf)z + xeyf\delta(z) \quad \text{for all } x, y, z \in A, \tag{13}$$

in light of (11) and (12). By the symmetry of e and f, we also have

$$\delta(x)fyez + x\delta(fyez) = \delta(xfye)z + xfye\delta(z) \quad \text{for all } x, y, z \in A. \tag{14}$$

Thus it remains to show that

$$\delta(x)eyez + x\delta(eyez) = \delta(xeye)z + xeye\delta(z) \quad \text{for all } x, y, z \in A \tag{15}$$

and

$$\delta(x)fyfz + x\delta(fyfz) = \delta(xfyf)z + xfyf\delta(z) \quad \text{for all } x, y, z \in A. \tag{16}$$

Applying (11), (12), (13) and (14), we shall express the sum

$$\delta(x)eyezfufvew + x\delta(eyezfufv)ew + xeyezfufv\delta(ew)$$

in two other ways. On the one hand, we have

$$\begin{aligned}
&\delta(x)eyezfufvew + x\delta(eyezfufv)ew + xeyezfufv\delta(ew)\\
&\quad = \delta(x)eyezfufvew + x\delta(eyez(fufve))ew + xeyez(fufve)\delta(ew)\\
&\quad = \delta(x)eyezfufvew + x\delta(eyez)fufvew + xeyez\delta(fufvew).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\delta(x)eyezfufvew + x\delta(eyezfufv)ew + xeyezfufv\delta(ew)\\
&\quad = \delta(x)(eyezfuf)fvew + x\delta((eyezfuf)fv)ew + xeyezfufv\delta(ew)\\
&\quad = \delta(xeyezfuf)fvew + xeyezfuf\delta(fv)ew + xeyezfufv\delta(ew)\\
&\quad = \delta(xey(ezfuf))fvew + xey(ezfuf)\delta(fv)ew + xeyezfufv\delta(ew)\\
&\quad = \delta(xey)ezfufvew + xey\delta(ezfufv)ew + xeyezfufv\delta(ew)\\
&\quad = \delta(xey)ezfufvew + xey\delta(ez(fufve))ew + xeyez(fufve)\delta(ew)\\
&\quad = \delta(xey)ezfufvew + xey\delta(ez)fufvew + xeyez\delta(fufvew)\\
&\quad = \delta(xeye)zfufvew + xeye\delta(z)fufvew + xeyez\delta(fufvew).
\end{aligned}$$

Comparing both expressions, we get

$$\big(\delta(x)eyez + x\delta(eyez) - \delta(xeye)z - xeye\delta(z)\big)fufvew = 0$$

for all $x, y, z, u, v, w \in A$. Since A is prime, we obtain

$$\big(\delta(x)eyez + x\delta(eyez)\big)f = \big(\delta(xeye)z + xeye\delta(z)\big)f. \tag{17}$$

for all $x, y, z \in A$. Similarly we express the sum

$$\delta(x)eyezeufvfwft + x\delta(eyezeufvfw)ft + xeyezeufvfw\delta(ft)$$

in two other ways. On the one hand, we have

$$\begin{aligned}
&\delta(x)eyezeufvfwft + x\delta(eyezeufvfw)ft + xeyezeufvfw\delta(ft)\\
&= \delta(x)eyezeufvfwft + x\delta(eyez(eufvfwf))ft + xeyez(eufvfwf)\delta(ft)\\
&= \delta(x)eyezeufvfwft + x\delta(eyez)eufvfwft + xeyez\delta(eufvfwft).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\delta(x)eyezeufvfwft + x\delta(eyezeufvfw)ft + xeyezeufvfw\delta(ft)\\
&= \delta(x)(eyezeuf)vfwft + x\delta((eyezeuf)vfw)ft + xeyezeufvfw\delta(ft)\\
&= \delta(xeyezeuf)vfwft + xeyezeuf\delta(vfw)ft + xeyezeufvfw\delta(ft)\\
&= \delta(xey(ezeuf))vfwft + xey(ezeuf)\delta(vfw)ft + xeyezeufvfw\delta(ft)\\
&= \delta(xey)ezeufvfwft + xey\delta(ezeufvfw)ft + xeyezeufvfw\delta(ft)\\
&= \delta(xey)ezeufvfwft + xey\delta(ez(eufvfwf))ft + xeyez(eufvfwf)\delta(ft)\\
&= \delta(xey)ezeufvfwft + xey\delta(ez)eufvfwft + xeyez\delta(eufvfwft)\\
&= \delta(xeye)zeufvfwft + xeye\delta(z)eufvfwft + xeyez\delta(eufvfwft).
\end{aligned}$$

Comparing both expressions, we get

$$\big(\delta(x)eyez + x\delta(eyez) - \delta(xeye)z - xeye\delta(z)\big)eufvfwft = 0$$

for all $x, y, z, u, v, w, t \in A$. Since A is prime, we obtain

$$\big(\delta(x)eyez + x\delta(eyez)\big)e = \big(\delta(xeye)z + xeye\delta(z)\big)e \tag{18}$$

for all $x, y, z \in A$. Then (15) follows immediately from the identities (17) and (18). By the symmetry of e and f, we have (16) too. Therefore,

$$\delta(x)yz + x\delta(yz) = \delta(xy)z + xy\delta(z) \quad \text{for all } x, y, z \in A. \tag{19}$$

Suppose that A contains the unity 1. Setting $x = z = 1$ in (19), we get $\delta(1)y = y\delta(1)$ for all $y \in A$. That is, $\lambda = \delta(1)$ is a central element of A. Setting $z = 1$ in (19) we get

$$\delta(xy) = \delta(x)y + x\delta(y) - \lambda xy \quad \text{for all } x, y \in A.$$

Clearly, if $\delta(1) = 0$, then δ is a derivation. □

Now we need some lemmas to deal with the functional identity $\delta(x)yz + x\delta(yz) = \delta(xy)z + xy\delta(z)$. The following two results are special cases of [**5**, Corollary 2.9]. The first one also appears in [**2**, Theorem 1.2], where bi-additivity of the maps F_i and G_i is assumed.

Lemma 7. *Let R be a prime ring with maximal right quotient ring Q and extended centroid C, such that* $\deg R \geq 3$. *Let $F_i, G_i : R \times R \to Q$, $i = 1, 2, 3$, be maps. Suppose that*

$$x_1F_1(x_2, x_3) + x_2F_2(x_1, x_3) + x_3F_3(x_1, x_2) + G_1(x_2, x_3)x_1 + G_2(x_1, x_3)x_2 + G_3(x_1, x_2)x_3 = 0$$

for all $x_1, x_2, x_3 \in R$. Then there exist unique maps $p_{i,j} : R \to Q$, for $1 \leq i \neq j \leq 3$, with $p_{i,j} = p_{j,i}$, and maps $\lambda_i : R \times R \to C$, for $i = 1, 2, 3$, such that

$$F_i(x_j, x_k) = p_{i,j}(x_k)x_j + p_{i,k}(x_j)x_k + \lambda_i(x_j, x_k),$$
$$G_j(x_i, x_k) = -x_ip_{i,j}(x_k) - x_kp_{j,k}(x_i) - \lambda_j(x_i, x_k)$$

for all $x_i, x_j, x_k \in R$, where i, j, k are distinct, and $\lambda_i = 0$ if either $F_i = 0$ or $G_i = 0$.

The next result also appears in [**8**, Lemma 4.5], where additivity of the maps f_i and g_i is assumed.

Lemma 8. *Let R be a prime ring with maximal right quotient ring Q and extended centroid C. Let $f_i, g_i : R \to R$, for $i = 1, 2$, be maps. Suppose that*

$$f_1(x)y + f_2(y)x + xg_1(y) + yg_2(x) = 0$$

for all $x, y \in R$. Then there exist unique constants $c_1, c_2 \in Q$ and maps $\lambda_1, \lambda_2 : R \to C$ such that

$$f_1(x) = xc_1 + \lambda_1(x), \qquad f_2(y) = yc_2 + \lambda_2(y),$$
$$g_1(y) = -c_1y - \lambda_2(y), \qquad g_2(x) = -c_2x - \lambda_1(x)$$

for all $x, y \in R$, where $\lambda_i = 0$ if either $f_i = 0$ or $g_i = 0$.

Now we are ready to conclude the paper by proving our second main result.

Proof of Theorem 2. Since (i) follows from Theorem 6, it remains to prove (ii).

Define two maps $F, G : A \times A \to A$ by $F(x, y) = \delta(xy) - x\delta(y)$ and $G(x, y) = \delta(xy) - \delta(x)y$ for $x, y \in A$. By Theorem 6, we have

$$xF(y, z) - G(x, y)z = 0 \quad \text{for all } x, y, z \in A.$$

Then it follows from Lemma 7 that there exists a map $p : R \to Q$ such that

$$F(x, y) = \delta(xy) - x\delta(y) = p(x)y,$$
$$G(x, y) = \delta(xy) - \delta(y)x = xp(y)$$

for all $x, y \in A$. Thus,

$$\delta(xy) = x\delta(y) + p(x)y = \delta(x)y + xp(y), \tag{20}$$

and hence

$$\big(\delta(x) - p(x)\big)y - x\big(\delta(y) - p(y)\big) = 0,$$

for all $x, y \in A$. By Lemma 8, there exist constants $c_1, c_2 \in Q$ such that $\delta(x) - p(x) = xc_1$ and $\delta(y) - p(y) = c_2 y$ for all $x, y \in R$. Then $c_1 = c_2$ is an element in C. Denote this element by λ. Thus

$$p(x) = \delta(x) - \lambda x$$

for all $x \in R$. Substituting this into (20), we get

$$\delta(xy) = x\delta(y) + \delta(x)y - \lambda xy$$

for all $x, y \in R$. The proof is complete. □

References

[1] J. Araujo and K. Jarosz, *Biseparating maps between operator algebras*, J. Math. Anal. Appl., **282**(1) (2003), 48–55, MR 2000328.

[2] K.I. Beidar, *On functional identities and commuting additive mappings*, Comm. Algebra, **26** (1998), 1819–1850, MR 1621755 (99f:16023), Zbl 0901.16011.

[3] K.I. Beidar and M.A. Chebotar, *On functional identities and d-free subsets of rings* I, Comm. Algebra, **28** (2000), 3925–3951, MR 1767598 (2001j:16046), Zbl 0991.16017.

[4] K.I. Beidar and M.A. Chebotar, *On functional identities and d-free subsets of rings* II, Comm. Algebra, **28** (2000), 3953–3972, MR 1767598 (2001j:16046), Zbl 0991.16018.

[5] K.I. Beidar and W.S. Martindale, *On functional identities in prime rings with involution*, J. Algebra, **203** (1998), 491–532, MR 1622795 (99f:16024), Zbl 0904.16012.

[6] K.I. Beidar, W.S. Martindale and A.V. Mikhalev, *Rings with Generalized Identities*, Pure and Applied Mathematics **196**, Marcel Dekker, New York, 1996, MR 1368853 (97g:16035), Zbl 0847.16001.

[7] M. Brešar, *Characterizations of derivations on some normed algebras with involution*, J. Algebra, **152** (1992), 454–462, MR 1194314 (94e:46098), Zbl 0769.16015.

[8] M. Brešar, *On generalized biderivations and related maps*, J. Algebra, **172** (1995), 764–786, MR 1324181 (96c:16046), Zbl 0827.16024.

[9] M. Brešar and P. Šemrl, *Mappings which preserve idempotents, local automorphisms, and local derivations*, Canad. J. Math., **45** (1993), 483–496, MR 1222512 (94k:47054), Zbl 0796.15001.

[10] M. Brešar and P. Šemrl, *On local automorphisms and mappings that preserve idempotents*, Studia Math., **113** (1995), 101–108, MR 1318418 (96i:47058), Zbl 0835.47020.

[11] M. Brešar and P. Šemrl, *Linear preservers on $\mathcal{B}(X)$*, in *Linear Operators* (Warsaw, 1994), Banach Center Publ., **38** (1997), 49–58, Polish Acad. Sci., Warsaw, MR 1457000 (99c:47044), Zbl 0939.47031.

[12] L.J. Bunce and J.D.M. Wright, *Velocity maps in von Neumann algebras*, Pacific J. Math., **170** (1995), 421–427, MR 1363871 (97c:46087), Zbl 0851.46048.

[13] M.A. Chebotar, W.-F. Ke, P.-H. Lee and N.-C. Wong, *Mappings preserving zero products*, Studia Math., **155**(1) (2003), 77–94, MR 1961162 (2003m:47066), Zbl 1032.46063.

[14] R. Crist, *Local derivations on operator algebras*, J. Funct. Anal., **135** (1996), 76–92, MR 1367625 (96m:46128), Zbl 0902.46046.

[15] R. Crist, *Local automorphisms*, Proc. Amer. Math. Soc., **128** (2000), 1409–1414, MR 1657786 (2001a:47078), Zbl 0948.47063.

[16] J. Cui and J. Hou, *Linear maps on von Neumann algebras preserving zero products or tr-rank*, Bull. Austral. Math. Soc., **65** (2002), 79–91, MR 1889381 (2002m:46092).

[17] W. Jing, *Local derivations of reflexive algebras*, Proc. Amer. Math. Soc., **125** (1997), 869–873, MR 1814104 (2002d:47098), Zbl 0865.47031.

[18] W. Jing, *Local derivations of reflexive algebras* II, Proc. Amer. Math. Soc., **129** (2001), 1733–1737, MR 1814104 (2002d:47098), Zbl 0969.47046.

[19] W. Jing, S. Lu and P. Li, *Characterization of derivations on some operator algebras*, Bull. Austral. Math. Soc., **66** (2002), 227–232, MR 1932346 (2003f:47059), Zbl 1035.47019.

[20] R.V. Kadison, *Local derivations*, J. Algebra, **130** (1990), 494–509, MR 1051316 (91f:46092), Zbl 0751.46041.

[21] D.R. Larson, *Reflexivity, algebraic reflexivity and linear interpolation*, Amer. J. Math., **110** (1988), 283–299, MR 0935008 (89d:47096), Zbl 0654.47023.

[22] D.R. Larson and A.R. Sourour, *Local derivations and local automorphisms of* $\mathcal{B}(X)$, in *Operator Theory: Operator Algebras and Applications* (Durham, NH, 1988), Part 2, Proc. Symp. Pure Math. **51**, Amer. Math. Soc., Providence, 1990, 187–194, MR 1077437 (91k:47106), Zbl 0713.47045.

[23] L. Molnár, *Local automorphisms of some quantum mechanical structures*, Lett. Math. Phys., **58** (2001), 91–100, MR 1876245 (2002k:47077), Zbl 1002.46044.

[24] L. Molnár, *Some characterizations of the automorphisms of* $B(H)$ *and* $C(X)$, Proc. Amer. Math. Soc., **130** (2002), 111–120, MR 1855627 (2002m:47047), Zbl 0983.47024.

[25] L. Molnár, 2*-local isometries of some operator algebras*, Proc. Edinburgh Math. Soc., **45** (2002), 349–352, MR 1912644 (2003e:47067), Zbl 1027.46062.

[26] L. Molnár and M. Györy, *Reflexivity of the automorphism and isometry groups of the suspension of* $\mathcal{B}(\mathcal{H})$, J. Funct. Anal., **159** (1998), 568–586, MR 1658095 (2000h:47110), Zbl 0933.46064.

[27] L. Molnár and P. Šemrl, *Local automorphisms of the unitary group and the general linear group on a Hilbert space*, Expo. Math., **18** (2000), 231–238, MR 1763889 (2001a:47083), Zbl 0963.46044.

[28] L. Molnár and B. Zalar, *On local automorphisms of group algebras of compact groups*, Proc. Amer. Math. Soc., **128** (2000), 93–99, MR 1637412 (2000f:43002), Zbl 0930.43002.

[29] T. Petek and P. Šemrl, *Adjacency preserving maps on matrices and operators*, Proc. R. Soc. Edinb., Sect. A, Math., **132** (2002), 661–684, MR 1912421 (2003f:47060), Zbl 1006.15015.

[30] E. Scholz and W. Timmermann, *Local derivations, automorphisms and commutativity preserving maps on* $L^+(D)$, Publ. Res. Inst. Math. Sci., **29** (1993), 977–995, MR 1256440 (94k:47071), Zbl 0817.47061.

[31] P. Šemrl, *Linear mappings preserving square-zero matrices*, Bull. Austral. Math. Soc., **48** (1993), 365–370, MR 1248039 (95a:15001), Zbl 0795.15002.

[32] P. Šemrl, *Local automorphisms and derivations on $\mathcal{B}(H)$*, Proc. Amer. Math. Soc., **125** (1997), 2677–2680, MR 1415338 (98e:46082), Zbl 0887.47030.

[33] M. Wolff, *Disjointness preserving operators on C^*-algebras*, Arch. Math., **62** (1994), 248–253, MR 1259840 (94k:46122), Zbl 0803.46069.

[34] W.J. Wong, *Maps on simple algebras preserving zero products.* I: *The associative case*, Pacific J. Math., **89**(1) (1980), 229–247, MR 0596933 (82k:15002a), Zbl 0405.16006.

Received February 19, 2003 and revised June 13, 2003.

Department of Mechanics and Mathematics
Tula State University
Tula
Russia
E-mail address: mchebotar@tula.net

Department of Mathematics
National Cheng Kung University
Tainan
Taiwan
E-mail address: wfke@mail.ncku.edu.tw

Department of Mathematics
National Taiwan University
Taipei
Taiwan
E-mail address: phlee@math.ntu.edu.tw

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

WAVE EQUATIONS FOR GRAPHS AND THE EDGE-BASED LAPLACIAN

Joel Friedman and Jean-Pierre Tillich

In this paper we develop a wave equation for graphs that has many of the properties of the classical Laplacian wave equation. This wave equation is based on a type of graph Laplacian we call the "edge-based" Laplacian. We give some applications of this wave equation to eigenvalue/geometry inequalities on graphs.

1. Introduction

The main goal of this paper is to develop a "wave equation" for graphs that is very similar to the wave equation $u_{tt} = \Delta u$ in analysis. Whenever this type of wave equation is involved in a result in analysis, our graph theoretic wave equation seems likely to provide the tool to link the result in analysis to an analogous result in graph theory.

Traditional graph theory defines a Laplacian, Δ, as an operator on functions on the vertices. This gives rise to a wave equation $u_{tt} = -\Delta u$ (since graph theory Laplacians are positive semidefinite). However, this wave equation fails to have a "finite speed of wave propagation". In other words, if $u = u(x,t)$ is a solution, we may have $u(x,0) = 0$ for all vertices x within a distance $d > 0$ to a fixed vertex, x_0, without having $u(x_0, \epsilon)$ vanishing for any $\epsilon > 0$. As such, this graph theoretic wave equation cannot link most results in analysis involving the wave equation to a graph theoretic analogue.

In this paper we study what appears to be a new type of wave equation on graphs. This wave equation (1) involves a reasonable analogue of $u_{tt} = \Delta u$ in analysis, (2) has "finite speed of wave propagation" and many other basic properties shared by its analysis counterpart, and (3) seems to be a good vehicle for translating results in analysis to those in graph theory, and vice versa. This wave equation cannot be expressed in the language of traditional graph theory; it requires some of the notions of "calculus on graphs" in [**FT99**]. It does, however, have a simple physical interpretation — namely, the edges are taut strings, fused together at the vertices. And in fact, the type of Laplacian we use has appeared in the physics literature as the "limiting case" of a "quantum wire" (see [**Hur00**, **RS01**, **KZ01**] for example); but our type of development of the wave equation and its applications to graph theory seem to have escaped the interest of physicists.

A second goal of this paper is to point out that whereas in analysis there is really *one* fundamental Rayleigh quotient, heat equation, and wave equation, in graph theory there are always *two*. These two fundamental types for an equation or concept result from the fact that graph theory involves two different volume measures. So while in analysis there is usually one fundamental, top (space) dimension volume, in graph theory one has a "vertex-based" measure, $\mathcal{V}$ (a type of vertex counting measure), and an "edge-based" measure, $\mathcal{E}$ (essentially Lebesgue measure on edges viewed as real intervals) (see below and Section 2 for details); both $\mathcal{V}, \mathcal{E}$ seem to play a volume measure type of role. So we get a "vertex-based" Rayleigh quotient, heat equation, wave equation, etc. and their "edge-based" counterpart. (We can also form "mixed" equations and concepts from these two pure types, as well as vary coefficients, add lower order terms, etc.) However, sometimes it turns out that one of the two types of equation or concept is less interesting. This definitely seems to be the case in the wave equation, where the vertex-based equation does not have finite propagation speed.

A third goal of this paper is to give some examples of how to translate results in analysis to graphs and vice versa using the edge-based Laplacian (including the wave equation). As an example, we give a simple proof of a relation between distances of sets, their sizes, and the first nonzero edge-based eigenvalue; our result can be better (or worse) than that of Chung–Faber–Manteuffel; our proof also works in analysis, and rederives the results in [**FT**] with a simpler proof. As another example, we show that the optimal generalized Alon–Milman type bound is essentially achieved by a generalized Chung–Faber–Manteuffel bound derived by Chung–Grigor'yan–Yau. We briefly describe how to convert graph theoretic results into analysis results. The results in analysis turn out to be independent of any discussion of edge-based Laplacians, and hence appear as a short, separate article [**FT**] (especially for analysts who wish to learn as little graph theory as possible).

In this paper we at times restrict ourselves to finite graphs; at other times we insist that the graphs be locally finite, i.e., that each vertex meets only finitely many edges; finally, some discussion is valid for arbitrary graphs. We will indicate at the beginning of each section and/or subsection when assumptions are made on the graphs therein.

The rest of this introduction, aside from closing remarks, is devoted to giving an informal overview of the simplest form of our wave equation. If this overview seems cryptic, the reader may wish to consult [**FT99**] or Section 2 of this paper.

Let $G = (V, E)$ be a graph. Let $\mathcal{G}$ be its *geometric realization*, i.e., the metric space consisting of V with a real interval of length 1 joining u and v "glued in" for edge $\{u, v\}$. Let $\mathcal{V}$ be the vertex counting measure, and $\mathcal{E}$ be Lebesgue measure on the edges. Then the (positive semidefinite) Laplacian, as in [**FT99**], takes a function, f, and returns an "integrating

factor" (essentially a measure, see Section 2),

$$\Delta f = (\Delta_V f)\, d\mathcal{V} + (\Delta_E f)\, d\mathcal{E}$$

where Δ_E is minus the usual real Laplacian (i.e., second derivative), and Δ_V is essentially a sum of normal derivatives along edges incident with each vertex. It therefore makes no sense to write a wave equation[1]

$$u_{tt} = -\Delta u,$$

for the left-hand side should be a function, and the right-hand side an integrating factor.

The vertex-based wave equation is

$$u_{tt}\, d\mathcal{V} = -\Delta u.$$

This means that $\Delta_E u = 0$, and so u is "edgewise linear" (i.e., a linear function when restricted to any edge). For such a u, $\Delta_V u$ coincides with the traditional graph theoretic Laplacian, and we recover the wave equation based on traditional graph theory.

The edge-based wave equation is the equation

$$u_{tt}\, d\mathcal{E} = -\Delta u,$$

or $\Delta_V u = 0$ and $u_{tt} = -\Delta_E u$. This equation has wave propagation speed 1, and has many other properties befitting a wave equation.

When using an edge-based concept, one may speak of Laplacian eigenvalues. In this case one is referring to the set Λ_E of λ with $\Delta_E f = \lambda f$ and $\Delta_V f = 0$. However, traditional graph theory deals with Λ_V, defined analogously. Fortunately, it is easy to relate the two notions of eigenvalues, assuming we "normalize" the Laplacian Δ_V (see Section 3). Namely, assuming Δ_V is normalized and our graph has all edge lengths one, we have

$$\lambda \in \widetilde{\Lambda_E} \iff 1 - \cos\sqrt{\lambda} \in \Lambda_V$$

where $\widetilde{\Lambda_E}$ is Λ_E with some "less interesting" eigenvalues (certain squares of multiples of π) discarded. Λ_V is a finite set of values between 0 and 2, and Λ_E is an infinite set of nonnegative values (whose square roots are periodic, and whose values satisfy a one-dimensional Weyl's asymptotic law).

In this paper we will mildly generalize this setup, allowing for edges of variable "length" and "weight," and vertices of various "weight."

The rest of this paper is organized as follows: in Section 2 we review some notions from the calculus on graphs of [**FT99**]. In Section 3 we discuss the edge-based eigenfunction theory; it closely resembles standard eigenfunction theory. In Section 4 we discuss the wave equation and its basic properties. In Section 5 we give some applications of the edge-based Laplacian and the wave equation.

[1]Recall that the minus sign appears in the wave equation since the Laplacian is positive semidefinite.

2. Calculus on graphs

2.1. The setup. We use a similar setting as in [**Fri93**], and we recall this setting here. Let $G = (V, E)$ be a graph (undirected), such that with each edge, $e \in E$, we have associated a length, $\ell_e > 0$. We form the *geometric realization*, $\mathcal{G}$, of G, which is the metric space consisting of V and a closed interval of length ℓ_e from u to v for each edge $e = \{u, v\}$. When there is no confusion, we identify a $v \in V$ with its corresponding point in $\mathcal{G}$ and identify an $e \in E$ with its corresponding closed interval in $\mathcal{G}$. By an *edge interior* we mean the interior of an edge in $\mathcal{G}$.

Definition 2.1. The *boundary*, $\partial\mathcal{G}$, of a graph, $\mathcal{G}$, is simply a specified subset of its vertices. By the *interior of* $\mathcal{G}$, denoted $\mathring{\mathcal{G}}$, we mean $\mathcal{G} \setminus \partial\mathcal{G}$; similarly the *interior vertices*, denoted $\mathring{V}$, we mean $V \setminus \partial\mathcal{G}$. We say that $\partial\mathcal{G}$ is separated if each boundary vertex is incident upon exactly one edge.

Boundary separation is a property whose analogue for manifolds is always true. In most practical situations one can assume the boundary is separated.

Convention 2.2. Unless specified, in this article we assume all graphs have a separated boundary.

In this article we will give "boundary condition" for functions to satisfy at the boundary. Neumann or mixed boundary conditions (see the next section) behave bizarrely unless the boundary is separated.[2]

Convention 2.3. By a *traditional graph* we mean an undirected graph $G = (V, E)$. In this article we assume our graphs are always given with:

(1) lengths associated to each edge,
(2) a specified boundary (i.e., a specified subset of vertices).

Whenever an edge length is not specified, it is taken to be one. Whenever a boundary is not specified, it is taken to be empty. We refer to the geometric realization, $\mathcal{G}$, of the graph as the graph, when no confusion may arise.

An edge $e = \{u, v\}$ of length, ℓ, is a real interval of length ℓ, and as such has two *standard coordinates*, one that sets u to 0 and v to ℓ, and the other vice versa. Whenever we speak of a property such as differentiability, we always mean with respect to these standard coordinates.

Definition 2.4. By $C^k(\mathcal{G})$, the set of *k-times continuously differentiable functions on* $\mathcal{G}$, we mean the set of continuous functions on $\mathcal{G}$ whose restriction to each edge interior is k-times uniformly continuously differentiable

[2]It is not hard to see that the Neumann condition for a function, f, on a boundary vertex, v, is equivalent to $\Delta_V f = 0$ at v (see this section and the next). This is only equivalent to the normal derivative at f vanishing along all boundary edges if v is incident upon only one edge.

(as a function on that real interval). The same definition applies with $\mathcal{G}$ replaced by $C^k(\mathcal{G} \setminus V)$.

We cannot differentiate functions on $\mathcal{G}$ without orienting the edges; however, we can always take the gradient of a differentiable function as long as we know what is meant by a vector field. Recall that a vector field on an interval is a section of its tangent bundle or, what is the same, a function on the interval with an orientation of the interval, where we identify f plus an orientation with $-f$ with the opposite orientation.

Definition 2.5. By $C^k(\mathrm{T}\mathcal{G})$, the set of k-times continuously differentiable vector fields on $\mathcal{G}$, we mean those data consisting of a k-times uniformly continuously differentiable vector field on each open interval corresponding to each edge interior.

Notice that a vector field is not defined at a vertex, only on edge interiors.

Definition 2.6. For $f \in C^k(\mathcal{G})$ we may form, by differentiation, its *gradient*, $\nabla f \in C^{k-1}(\mathrm{T}\mathcal{G})$. For $X \in C^k(\mathrm{T}\mathcal{G})$ we can form, by differentiation, its *calculus divergence*, $\nabla_{\mathrm{calc}} \cdot X \in C^{k-1}(\mathrm{T}\mathcal{G} \setminus V)$.

Definition 2.7. A subset $\Omega \subset \mathcal{G}$ is of *finite type* if it lies in the union of finitely many vertices and edges. A function on $\mathcal{G}$ is of *finite type* if its support (i.e., the closure of the set where it does not vanish) is of finite type. We set $C^k_{\mathrm{fn}}(\mathcal{G})$ to be those elements of $C^k(\mathcal{G})$ of finite type; we similarly define $C^k_{\mathrm{fn}}(\mathcal{G} \setminus V)$ and $C^k_{\mathrm{fn}}(\mathrm{T}\mathcal{G})$.

Definition 2.8. An $f \in C^k_{\mathrm{fn}}(\mathcal{G})$ is said to satisfy the *Dirichlet condition* if f vanishes on $\partial\mathcal{G}$. We let $C^k_{\mathrm{Dir}}(\mathcal{G})$ denote the set of such functions.

Definition 2.9. $\mathrm{Lip}(\mathcal{G})$ denotes the class of *Lipschitz continuous* functions on $\mathcal{G}$, i.e., those $f \in C^0(\mathcal{G})$ whose restriction to each edge interior is uniformly Lipschitz continuous. We similarly define $\mathrm{Lip}_{\mathrm{fn}}(\mathcal{G})$ and $\mathrm{Lip}_{\mathrm{Dir}}(\mathcal{G})$.

2.2. Two volume measures. In analysis concepts such as Laplacians, Rayleigh quotients, and isoperimetric constants are defined using one volume measure; in calculus on graphs we use two "volume" measures.

Definition 2.10. A *vertex measure*, $\mathcal{V}$, is a measure supported on V with $\mathcal{V}(v) > 0$ for all $v \in V$. An *edge measure*, $\mathcal{E}$, is a measure with $\mathcal{E}(v) = 0$ for all $v \in V$ and whose restriction to any edge interior, $e \in E$, is Lebesgue measure (viewing the interior as an open interval) times a constant $a_e > 0$.

Traditional graph theory usually works with the traditional vertex and edge measures, $\mathcal{V}_{\mathrm{T}}$ and $\mathcal{E}_{\mathrm{T}}$, given by $\mathcal{V}_{\mathrm{T}}(v) = 1$ for all $v \in V$ and $a_e = 1$ for all $e \in E$, i.e., $\mathcal{E}_{\mathrm{T}}$ is just Lebesgue measure at each edge.

Convention 2.11. Henceforth we assume that any graph has an associated vertex measure, $\mathcal{V}$, and an edge measure, $\mathcal{E}$. When $\mathcal{V}$ is not specified we take it to be $\mathcal{V}_{\mathrm{T}}$; similarly, when unspecified we take $\mathcal{E}$ to be $\mathcal{E}_{\mathrm{T}}$.

In this article we write $\int f\,d\mathcal{E}$ and $\int f\,d\mathcal{V}$ for $\int_{\mathcal{G}} f\,d\mathcal{E}$ and $\int_{\mathcal{G}} f\,d\mathcal{V}$.

2.3. Integrating factors. In this paper a somewhat formal notion will become extremely important.

Definition 2.12. By an *integrating factor* on $\mathcal{G}$ we mean a formal expression of the form $\mu = \alpha\,d\mathcal{V} + \beta\,d\mathcal{E}$ where α is a function defined (at least) on the vertices of $\mathcal{G}$, and $\beta \in C^0(\mathcal{G} \setminus V)$.

The continuity assumption on β is not essential, but it makes things nicer for the following reason: an integrating factor as above determines a linear functional $\mathcal{L}_\mu$, on $C^0_{\mathrm{Dir}}(\mathcal{G})$ via

$$\mathcal{L}_\mu = \int f\mu = \int f\alpha\,d\mathcal{V} + \int f\beta\,d\mathcal{E}.$$

We say that two integrating factors $\mu_i = \alpha_i(x)\,d\mathcal{V} + \beta_i(x)\,d\mathcal{E}$ for $i = 1, 2$ are equal if the $\mathcal{L}_{\mu_i}$ are equal; clearly this amounts to $\alpha_1 = \alpha_2$ at the interior vertices and $\beta_1 = \beta_2$ everywhere on $\mathcal{G} \setminus V$ (since the β_i are continuous there).

We will sometimes wish to insist that $\alpha_1 = \alpha_2$ on boundary vertices as well; this corresponds to viewing integrating factors as linear functionals on $C^0(\mathcal{G})$. In this case we will speak of *boundary inclusive equality*.

In the calculus on graphs we have two measures, and thus a need for integrating factors, i.e., the need to mark functions with a $d\mathcal{V}$ or $d\mathcal{E}$ to remind us how to integrate the function against other functions. For example, we shall soon see that the divergence of a vector field or the Laplacian of a function is an integrating factor. Consequently, any wave or heat equation is most correctly regarded as an equation between integrating factors. (In traditional graph theory, all integrating factors have a vanishing $d\mathcal{E}$ component, i.e., $\beta = 0$ in the above, and they may be considered as functions on the vertices, i.e., they may be identified with α's values on the vertices.)

2.4. Regular graphs. In this article, r-regularity has a slightly more general meaning than in traditional graph theory where all vertices and edges have weight 1 (in other words $\mathcal{E} = \mathcal{E}_{\mathrm{T}}$ and $\mathcal{V} = \mathcal{V}_{\mathrm{T}}$).

Here we mean the following:

Definition 2.13. We say that a graph $\mathcal{G}$ is *r-regular* if for any $v \in \mathring{V}$ we have $\rho(v) = r$, where

$$\rho(v) = \mathcal{V}(v)^{-1} \sum_{e \ni v} \mathcal{E}(e).$$

Clearly graphs that are r-regular in the traditional sense are regular with our definition. The quantity $\rho(v)$ arises quite naturally when we consider *edgewise linear* functions, i.e., continuous functions whose restriction to each edge is a linear function. For these functions we clearly have

$$\int_e f\,d\mathcal{E} = \tfrac{1}{2}\big(f(u) + f(v)\big)\mathcal{E}(e)$$

for each edge $e = \{u, v\}$. Hence

$$\int f \, d\mathcal{E} = \int \tfrac{1}{2} f \rho \, d\mathcal{V}. \tag{2.1}$$

2.5. The divergence. The divergence of a vector field and the Laplacian of a function can be defined in terms of concepts that are already fixed, namely a graph (encompassing measures $\mathcal{E}$ and $\mathcal{V}$) and the gradient. Interestingly enough, the divergence turns out to be different from the "calculus divergence" described earlier.

Before defining the divergence we record a "divergence theorem" for the calculus divergence.

Let $X \in C^1(\mathrm{T}\mathcal{G})$. For any edge $e = \{u, v\}$ let $X|_e$ denote X restricted to the interior of e and then extended to u and v by continuity. We clearly have

$$\int_e \nabla_{\mathrm{calc}} \cdot X \, d\mathcal{E} = a_e\big(\mathbf{n}_{e,u} \cdot X|_e(u) + \mathbf{n}_{e,v} \cdot X|_e(v)\big),$$

where $\mathbf{n}_{e,u}, \mathbf{n}_{e,v}$ denote outward pointing unit (normal) vectors. Hence we obtain:

Proposition 2.14. *For all $X \in C^1_{\mathrm{fn}}(\mathcal{G})$ we have*

$$\int \nabla_{\mathrm{calc}} \cdot X \, d\mathcal{E} = \int \widetilde{\mathbf{n}} \cdot X \, d\mathcal{V}, \tag{2.2}$$

where

$$(\widetilde{\mathbf{n}} \cdot X)(v) = \mathcal{V}(v)^{-1} \sum_{e \ni v} a_e \mathbf{n}_{e,v} \cdot X|_e(v).$$

Let $C^k_{\mathrm{Dir}}(\mathcal{G})$ denote those functions in $C^k_{\mathrm{fn}}(\mathcal{G})$ that vanish on the boundary of $\mathcal{G}$.

Definition 2.15. For a vector field, X, its *divergence functional* is the linear functional $\mathcal{L}_X : C^\infty_{\mathrm{Dir}}(\mathcal{G}) \to \mathbb{R}$ given by

$$\mathcal{L}_X(g) = -\int X \cdot \nabla g \, d\mathcal{E}.$$

Proposition 2.16. *For any $X \in C^1(\mathrm{T}\mathcal{G})$ and $g \in C^\infty_{\mathrm{Dir}}(\mathcal{G})$ we have*

$$\mathcal{L}_X(g) = \int (\nabla_{\mathrm{calc}} \cdot X) g \, d\mathcal{E} - \int (\widetilde{\mathbf{n}} \cdot X) g \, d\mathcal{V},$$

i.e., the divergence functional of X is represented by $(\nabla_{\mathrm{calc}} \cdot X)\, d\mathcal{E} - (\widetilde{\mathbf{n}} \cdot X)\, d\mathcal{V}$ (viewed as a linear functional via integration).

Proof. We substitute Xg for X in Equation (2.2), and note that

$$\nabla_{\mathrm{calc}} \cdot (Xg) = g \nabla_{\mathrm{calc}} \cdot X + X \cdot \nabla g.$$

Definition 2.17. For $X \in C^1(\mathrm{T}\mathcal{G})$ we define its *divergence*, $\nabla \cdot X$, to be the integrating factor

$$(\nabla_{\mathrm{calc}} \cdot X)\, d\mathcal{E} - (\widetilde{\mathbf{n}} \cdot X)\, d\mathcal{V}.$$

If X is edgewise constant, so that $\nabla_{\mathrm{calc}} \cdot X = 0$, we will also refer to

$$-\widetilde{\mathbf{n}} \cdot X$$

(a function defined only on vertices) as its divergence, and write it $\nabla \cdot X$.

We conclude:

Proposition 2.18. *For any $g \in C^1_{\mathrm{fn}}(\mathcal{G})$ and $X \in C^1_{\mathrm{fn}}(\mathrm{T}\mathcal{G})$ we have*

$$\int_{\mathcal{G}} (\nabla \cdot X) g + \int X \cdot \nabla g \; d\mathcal{E} = 0.$$

To make this look more like analysis we can write this as:

$$\int_{\mathcal{G} \setminus \partial\mathcal{G}} (\nabla \cdot X) g + \int X \cdot \nabla g \, d\mathcal{E} = \int_{\partial\mathcal{G}} (\widetilde{\mathbf{n}} \cdot X) g \, d\mathcal{V}.$$

2.6. The Laplacian. In graph theory we usually define positive semidefinite Laplacians. So we define

$$\Delta f = -\nabla \cdot (\nabla f).$$

As integrating factors we have

$$\Delta f = (\Delta_E f)\, d\mathcal{E} + (\Delta_V f)\, d\mathcal{V},$$

with $\Delta_E f = -\nabla_{\mathrm{calc}} \cdot \nabla f$ and $\Delta_V f = \widetilde{\mathbf{n}} \cdot \nabla f$. It is easy to see that:

Proposition 2.19. *For all $f \in C^2_{\mathrm{fn}}(\mathcal{G})$ and $g \in C^1_{\mathrm{fn}}(\mathcal{G})$ we have*

$$\int (\Delta f) g = \int \nabla f \cdot \nabla g \, d\mathcal{E}. \tag{2.3}$$

If also $g \in C^2_{\mathrm{fn}}(\mathcal{G})$ we have

$$\int (\Delta f) g = \int (\Delta g) f. \tag{2.4}$$

The link with the traditional graph theoretic Laplacian is as follows:

Proposition 2.20. *For $f \in C^2_{\mathrm{fn}}(\mathcal{G})$ edgewise linear we have*

$$\nabla_{\mathrm{calc}} \cdot \nabla f = 0, \quad \textit{hence} \quad \Delta f = \widetilde{\mathbf{n}} \cdot \nabla f \, d\mathcal{V}.$$

Viewing Δf as a function on vertices we therefore have:

$$(\Delta f)(v) = \mathcal{V}(v)^{-1} \sum_{e \sim \{u,v\}} a_e \frac{f(v) - f(u)}{\ell_e}. \tag{2.5}$$

When restricting to edgewise linear functions, it is common (in graph theory) to write Δ as $D - A$, where D is the diagonal matrix or operator (classically the "degree" matrix) whose v, v entry is

$$L(v) = \mathcal{V}(v)^{-1} \sum_{e \sim \{u,v\}} a_e/\ell_e,$$

where we omit e's that are self-loops from the summation, and where A is the "adjacency" matrix or operator given by

$$(Af)(v) = \mathcal{V}(v)^{-1} \sum_{e \sim \{u,v\}} (a_e/\ell_e) f(u),$$

again omitting self-loops, e.

2.7. Variable integrals. We shall wish to consider the derivative at $t = t_0$ of the function

$$\mathcal{I}(t) = \int_{S(t)} f(x,t)\, d\mathcal{E}(x),$$

where t is a real parameter, $S(t)$ is a decreasing family of open subsets of $\mathcal{G}$, and $f(x,t)$ is continuous in x (taking values in $\mathcal{G}$) and differentiable in t (taking values near t_0).

For this paper we only need consider $S(t)$ given by the set of points within a distance $\tau - ct$ from a fixed set $A \subset \mathcal{G}$, with τ fixed and $0 \le t_0 < \tau$. We will assume $S(t)$ is of finite type for t near t_0, and that $\partial S(t_0)$ is finite and contains no vertices.

With the above assumptions, the formula for $\mathcal{I}'(t_0)$ is very easy. We will discuss more general variants of these formulas later. These more general variants are not needed in this paper, but are interesting to consider and compare with the co-area formula and its problems at the vertices as described in [**FT99**].

Calculus says that if a, b, c are constants with $c > 0$, and $f = f(x,t) : \mathbb{R}^2 \to \mathbb{R}$ is continuous in x and differentiable in t, then

$$\frac{d}{dt} \int_{a+ct}^{b-ct} f(x,t)\, dx = -c\big(f(a+ct,t) + f(b-ct,t)\big) + \int_{a+ct}^{b-ct} f_t(x,t)\, dx,$$

where $f_t = \partial f/\partial t$. If we replace $a + ct$ by a in the above integral, the $f(a+ct,t)$ drops out, and similarly for $b - ct$ replaced by b.

Summing the above calculus equality over all edges yields the following easy proposition:

Proposition 2.21. *Let $S(t), f(x,t), t_0, \mathcal{I}(t)$ be as above. Then*

$$\mathcal{I}'(t_0) = - \sum_{\substack{(x,e) \text{ s.t. } e \ni x, \\ x \in \partial S(t_0)}} f(x,t_0) c a_e + \int_{S(t_0)} f_t(x,t_0)\, d\mathcal{E}(x). \tag{2.6}$$

In other words, the above sum is over all boundary points, x, of $S(t_0)$, and involves the edge, e, on which x lies.

Finally, we remark that if $S(t)$ does not decrease "linearly" with speed c everywhere, then we simply replace c by the speed at with $\partial S(t)$ moves at x in the summation of any of the above formulas.

We finish this subsection by describing what happens when $\partial S(t_0)$ contains vertices. If so, then the left and right derivatives of $\mathcal{I}(t)$ at t_0 will exist but won't usually be equal. Equation 2.6 will essentially hold, but different (x, e) pairs appear in the sum. So consider pairs (x, e) with $x \in e \cup \partial S(t_0)$ (remember we identify an edge with its *closed* interval in $\mathcal{G}$, so x may be a vertex); we call a pair *future active* if e's interior intersected with $S(t_0)$ contains an open interval ending at x. In other words, the picture of $S(t)$ near x and along e is changing for $t > t_0$ near t_0 (since $S(t)$ is decreasing in t). We say that a pair (x, e) is *past active* if either x is not a vertex, or x is a vertex and (x, e) is not future active, see Figure 1. (Geometrically, since $S(t)$ is decreasing in t, we are saying that for t near t_0 and at a boundary point, x, the picture of $S(t)$ always changes when x is not a vertex, and when x is a vertex it changes along some edges in the future ($t > t_0$) and other edges in the past ($t < t_0$).)

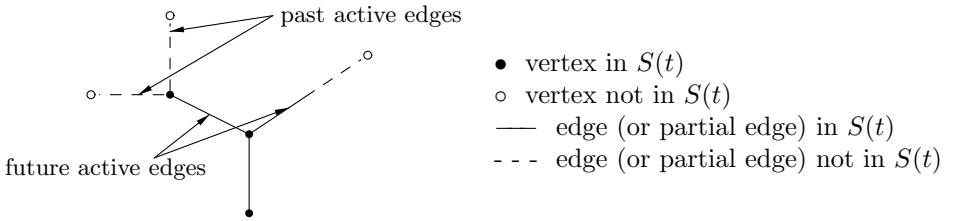


Figure 1.

Summing the calculus formula shows that the right derivative of $\mathcal{I}$ at t_0 exists and equals

$$\mathcal{I}'(t+0) = -\sum_{(x,e) \text{ future active}} f(x,t_0) c a_e + \int_{S(t_0)} f_t(x,t_0)\, d\mathcal{E}(x). \tag{2.7}$$

Similarly the left derivative is the same, with future active replaced by past active pairs (x, e).

Notice that the notion of "future active" essentially arose in the definition of the area of the boundary of a subset, in [**FT99**], Section 2, in connection with the co-area formula.

3. The edge-based eigenvalues and eigenfunctions

In this section we discuss some facts about the "edge-based" Laplacian and its eigenpairs, i.e., pairs (f, λ) with $\Delta_E f = \lambda f$ and $\Delta_V f = 0$. Such pairs

are crucial to understanding the solutions to the wave equation and invariants associated with it. We concern ourselves with the basic cases at first, later illustrating fancier boundary conditions and mixed edge and vertex Laplacians.

It is worth mentioning that the equations $\Delta_E f = \lambda f$ and $\Delta_V f = 0$ describe the modes associated with a physical object with a metal string for each edge, with strings being fused together at the vertices. For example, if we "pluck" such an object, it would produce tones with the frequencies of $\sqrt{\lambda}$ with λ ranging over the edge-based eigenvalues (this is seen from considering the wave equation of the next section).

3.1. Basic existence theory.

Definition 3.1. We say that (f, λ) is an *eigenpair for the "edge-based" Laplacian* if $f \in C^\infty(\mathcal{G})$ and satisfies $\Delta_E f = \lambda f$ and $\Delta_V f = 0$. We say that f satisfies the *Dirichlet condition* if f vanishes at all boundary points; the similarly for the *Neumann condition*, where f's normal derivatives along its edges evaluated at any boundary point vanish.

Notice that the condition $\Delta_E f = \lambda f$ implies that f's restriction to each edge is given by $A\cos(\omega x + B)$, where A, B are constants, $\omega = \lambda^{1/2}$, and x represents one of the two standard coordinates on the edge.

The existence of a complete set of eigenpairs for the Laplacian is well understood in analysis for compact domains, and the same techniques carry over to our setting, for finite graphs, with almost no modifications. We only summarize the theory, and refer the reader to [**GT83**] or [**Fri69**].

Proposition 3.2. *Let $\mathcal{G}$ be a finite graph. There exists eigenpairs (f_i, λ_i) for the edge based Laplacian, such that:*

(1) $0 \le \lambda_1 \le \lambda_2 \le \cdots$,
(2) *the f_i satisfy the Dirichlet condition,*
(3) *the f_i form a complete orthonormal basis for $L^2_{\mathrm{Dir}}(\mathcal{G}, \mathcal{E})$, and*
(4) $\lambda_i \to \infty$.

The same statement holds with Dirichlet replaced by Neumann and $L^2_{\mathrm{Dir}}(\mathcal{G}, \mathcal{E})$ replaced by $L^2(\mathcal{G}, \mathcal{E})$.

Proof. Consider the Rayleigh quotient

$$\mathcal{R}(f) = \frac{\int |\nabla f|^2 \, d\mathcal{E}}{\int |f|^2 \, d\mathcal{E}},$$

which is certainly defined for $f \in C^1(\mathcal{G})$. Let $u_1, u_2, \ldots$ be a minimizing sequence for $\mathcal{R}$ in $C^1_{\mathrm{Dir}}(\mathcal{G})$, i.e., $u_i \in C^1_{\mathrm{Dir}}(\mathcal{G})$ with

$$\mathcal{R}(u_i) \to \inf_{f \in C^1_{\mathrm{Dir}}(\mathcal{G})} \mathcal{R}(f).$$

We may assume $\int |u_i|^2 \, d\mathcal{E} = 1$, and thus that $\int |\nabla u_i|^2 \, d\mathcal{E}$ are bounded. Let $H^1_{\mathrm{Dir}}(\mathcal{G})$ be the closure of $C^1_{\mathrm{Dir}}(\mathcal{G})$ under the norm

$$\|f\|^2_{H^1} = \int [|\nabla f|^2 + |f|^2] \, d\mathcal{E};$$

$H^1_{\mathrm{Dir}}(\mathcal{G})$ can also be described with Fourier transforms, or as the set of L^2 functions with a weak derivative in L^2; it is well-known to be a separable Hilbert space, therefore having a weakly compact unit ball. Hence by passing to a subsequence we may assume that the u_i converge weakly in H^1 to a $u \in H^1_{\mathrm{Dir}}(\mathcal{G})$.

The u_i are uniformly Hölder continuous of exponent $1/2$. To see this let $x, y \in \mathcal{G}$ be of distance ρ, and fix a path γ of length ρ from x to y. We have

$$\big|u_i(x) - u_i(y)\big| \le \int_\gamma |\nabla u_i| \, d\mathcal{E} \le \left(\int_\gamma d\mathcal{E}\right)^{\frac{1}{2}} \left(\int_\gamma |\nabla u_i|^2 \, d\mathcal{E}\right)^{\frac{1}{2}} \le C\rho^{1/2} \|u_i\|_{H^1},$$

where C is the maximum over edges e of $a_e^{-1/2}$. Hence our claim holds, and by Ascoli's lemma we can pass to a further subsequence and assume that the u_i converge uniformly to a u that is Hölder continuous of exponent $1/2$; since the u_i vanish on the boundary, so does u.

We have (by the uniform convergence and the weak H^1 convergence)

$$\mathcal{R}(u) \le \liminf \mathcal{R}(u_i),$$

and so equality must hold and u minimizes $\mathcal{R}$ over all of $H^1_{\mathrm{Dir}}(\mathcal{G})$.

Now we claim that u is our desired eigenfunction, and $\lambda = \mathcal{R}(u)$ its eigenvalue. This is seen by setting $\lambda = \mathcal{R}(u)$ and considering $\mathcal{R}(u + \epsilon w)$ for various $w \in H^1_{\mathrm{Dir}}(\mathcal{G})$ and taking $\epsilon \to 0$. We conclude that

$$\int \nabla u \nabla w \, d\mathcal{E} = \lambda \int uw \, d\mathcal{E} \tag{3.1}$$

for all $w \in H^1_{\mathrm{Dir}}(\mathcal{G})$. Now standard estimates for elliptic equations (e.g., Lemma 15.4 of [**Fri69**]) show that in fact u is C^∞ at all edge interiors, and satisfies

$$\Delta_E u = \lambda u. \tag{3.2}$$

Hence u's restriction to any edge is given as $A\cos(\omega x + B)$ for $\omega = \lambda^{1/2}$, and A, B are constants depending on e (and which of the two standard coordinates we place on e). Since u is Hölder continuous on $\mathcal{G}$, it is certainly continuous everywhere, including all its vertices. Hence $u \in C^\infty_{\mathrm{Dir}}(\mathcal{G})$. Finally, given a vertex, v, let us show that $(\Delta_V u)(v) = 0$. From Proposition 2.19 we know that

$$\int \nabla u \cdot \nabla w \, d\mathcal{E} = \int (\Delta_E u) w \, d\mathcal{E} + \int (\Delta_V u) w \, d\mathcal{V} = \lambda \int uw \, d\mathcal{E} + \int (\Delta_V u) w \, d\mathcal{V}.$$

From (3.1) and (3.2) we obtain

$$\int (\Delta_V u) w \, d\mathcal{V} = 0. \tag{3.3}$$

We choose $w \in C^1_{\text{fn}}(\mathcal{G})$ in Equation (3.1), such that $w(v) = 1$ and $w(v') = 0$ on other vertices and conclude

$$(\Delta_V u)(v) = 0.$$

Letting $h \to 0$ we conclude that u is an "edge-based" Laplacian eigenfunction satisfying the Dirichlet condition, with eigenvalue λ.

Set $f_1 = u$ and $\lambda_1 = \lambda$. Now repeat the same argument, except minimizing $\mathcal{R}$ over those functions that are orthogonal to f_1. With the same argument, we find an eigenpair (f_2, λ_2) with $\lambda_1 \leq \lambda_2$ and f_1 orthogonal to f_2. Now repeat again, minimizing over functions orthogonal to f_1 and f_2. In this way we get a sequence of orthogonal eigenpairs (f_i, λ_i) with λ_i nondecreasing in i.

Next we show that $\lambda_i \to \infty$. Otherwise the H^1 norms of the f_i are uniformly bounded, and so the f_i are uniformly Hölder continuous, and so a subsequence of the f_i converges uniformly to a g. But since the f_i are orthogonal, g would be orthogonal to all f_i and therefore to itself, so $g = 0$. This contradicts the uniform convergence of the subsequence of f_i to g.

It remains to show that the f_i are complete. If the f_i are not complete, there is a nonzero $g \in L^2_{\text{Dir}}(\mathcal{G})$ orthogonal to all the f_i. By convolving g with smooth approximations to Dirac's delta function, and modifying it at the vertices, we can find a $g_\epsilon \in C^\infty_{\text{Dir}}(\mathcal{G})$ for any $\epsilon > 0$ with $\|g - g_\epsilon\|_{L^2} \leq \epsilon$. Hence for small ϵ the function, h, which is the projection of g_ϵ onto the complement of the f_i's, is: (1) nonzero, (2) orthogonal to all f_i, and (3) lies in $H^1_{\text{Dir}}(\mathcal{G})$. Now $\mathcal{R}(h)$ must upper bound the λ_i, by their minimizing property. Hence λ_i are bounded, which we know is impossible.

We finish by remarking that the same proof holds for the Neumann condition, except that we begin by working with $C^1(\mathcal{G})$ instead of $C^1_{\text{Dir}}(\mathcal{G})$. □

The same theorem holds for more general "mixed" boundary conditions. Namely, we consider the condition that a function, f, satisfy the mixed condition

$$f = 0 \quad \text{on } K_1, \tag{3.4}$$

$$\widetilde{\mathbf{n}} \cdot \nabla f + \sigma f = 0 \quad \text{on } K_2, \tag{3.5}$$

where K_1, K_2 are a partition of $\partial\mathcal{G}$ and where σ is nonnegative. Then the same theorem and proof hold, provided that we replace $C^1_{\text{Dir}}(\mathcal{G})$ by

$$\big\{ f \in C^1(\mathcal{G}) \mid f = 0 \text{ on } K_1 \big\},$$

and replace the Rayleigh quotient, $\mathcal{R}$, by

$$\widetilde{\mathcal{R}}(f) = \frac{\int |\nabla f|^2 \, d\mathcal{E} + \int_{K_2} \sigma f^2 \, d\mathcal{V}}{\int |f|^2 \, d\mathcal{E}}. \tag{3.6}$$

We mention that $\mathcal{V}$ does not enter in any essential way into the edge-based eigenvalues. Only $\mathcal{E}$ should affect those eigenvalues.

We can get mixed edge-vertex Laplacians. So consider the Rayleigh quotient:

$$\mathcal{R}(f) = \frac{\int \gamma |\nabla f|^2 \, d\mathcal{E}}{\int \alpha f^2 \, d\mathcal{V} + \int \beta f^2 \, d\mathcal{E}},$$

with α, β, γ continuous, nonnegative and γ differentiable and never vanishing. Its successive minimizers satisfy

$$-\nabla \cdot (\gamma \nabla f) = \lambda f(\alpha \, d\mathcal{V} + \beta \, d\mathcal{E}). \tag{3.7}$$

The theorem above gives us a complete eigenbasis for $L^2(\mathcal{G}, \mu)$ with $\mu = \alpha\mathcal{V} + \beta\mathcal{E}$. This basis is infinite provided that β is not identically zero.

Finally, we mention that it is easy to modify the above to work for mixed boundary conditions with mixed edge-vertex Laplacians.

3.2. Weyl's law. One fundamental result about edge-based eigenvalues that is true for any finite graph is that their growth rate is determined, to first-order, by the sum of the lengths of their edges. In analysis the analogous quantity is the volume of the subdomain or manifold, and Weyl's proof of this fact (see [**Wey12**]) in analysis immediately carries over here.

For a finite graph, $\mathcal{G}$, let $N_{\mathrm{Dir}}(\lambda, \mathcal{G})$ be the number of Dirichlet edge-based eigenvalues $\leq \lambda$ for $\mathcal{G}$, and similarly for $N_{\mathrm{Neu}}(\lambda, \mathcal{G})$.

Proposition 3.3 (Weyl's Law)**.** *Fix a finite graph, $\mathcal{G}$. Let $N(\lambda)$ be either $N_{\mathrm{Dir}}(\lambda, \mathcal{G})$ or $N_{\mathrm{Neu}}(\lambda, \mathcal{G})$. There is a constant, C, such that*

$$|N(\lambda) - L\lambda^{1/2}/\pi| \leq C,$$

where L is the sum of all the lengths of the edges in the graph.

Proof. Consider the graph, $\mathcal{G}_1$, where every vertex is a boundary point. Then the edge-based eigenvalues of $\mathcal{G}_1$ are found by minimizing the same Rayleigh quotient over a more restrictive class of functions. Hence, by the min-max principle,

$$N_{\mathrm{Dir}}(\lambda, \mathcal{G}_1) \leq N_{\mathrm{Dir}}(\lambda, \mathcal{G}).$$

Similarly,

$$N_{\mathrm{Neu}}(\lambda, \mathcal{G}) \leq N_{\mathrm{Neu}}(\lambda, \mathcal{G}_1),$$

for $N_{\mathrm{Neu}}(\lambda, \mathcal{G}_1)$ corresponds to a Rayleigh quotient over the space of functions that needn't be continuous at any vertex. For similar reasons we have

$$N_{\mathrm{Dir}}(\lambda, \mathcal{G}) \leq N_{\mathrm{Neu}}(\lambda, \mathcal{G}),$$

the former N corresponding to the same class of functions as the latter except the latter need not vanish at boundary vertices. To summarize, we have shown

$$N_{\mathrm{Dir}}(\lambda, \mathcal{G}_1) \leq N_{\mathrm{Dir}}(\lambda, \mathcal{G}) \leq N_{\mathrm{Neu}}(\lambda, \mathcal{G}) \leq N_{\mathrm{Neu}}(\lambda, \mathcal{G}_1).$$

Hence it suffices to prove the proposition for $\mathcal{G}_1$.

But the values of the $\mathcal{G}_1$ eigenfunctions don't interact across vertices, so

$$N_{\mathrm{Dir}}(\lambda, \mathcal{G}_1) = \sum_{e \in E} N_{\mathrm{Dir}}(\lambda, e),$$

where edges, e, are also viewed as graphs. If e is an edge of length ℓ, its Dirichlet eigenfunctions are $f_n(x) = \sin(nx\pi/\ell)$ for $n = 1, 2, \ldots$, with eigenvalues $\lambda_n = (n\pi/\ell)^2$. Hence we have

$$N_{\mathrm{Dir}}(\lambda, e) = \lfloor \lambda^{1/2} \ell/\pi \rfloor.$$

It follows that for some constant $C > 0$,

$$N_{\mathrm{Dir}}(\lambda, \mathcal{G}_1) \geq -C + (\lambda^{1/2}/\pi) \sum_{e \in E} \ell_e = -C + L\lambda^{1/2}/\pi.$$

For similar reasons the result also holds for $N_{\mathrm{Neu}}(\lambda, \mathcal{G}_1)$, where the eigenfunctions on an edge are $f_n(x) = \cos\big((n-1)x\pi/\ell\big)$ for $n = 1, 2, \ldots$, with eigenvalues $\lambda_n = \big((n-1)\pi/\ell\big)^2$. □

Again, a similar result holds for mixed boundary conditions. To see this, we shall show that for any fixed mixed boundary condition (as in the end of the previous subsection)

$$N_{\mathrm{Dir}}(\lambda, \mathcal{G}) \leq N_{\mathrm{mixed}}(\lambda, \mathcal{G}) \leq N_{\mathrm{Neu}}(\lambda, \mathcal{G}),$$

where N_{mixed} counts eigenvalues with a mixed condition. Indeed, N_{Dir} can be viewed as having the same Rayleigh quotient, as in Equation (3.6), as N_{mixed}, except over a smaller space (i.e., the space of functions vanishing over all of $\partial\mathcal{G}$, not just K_1). Furthermore, the Rayleigh quotient for N_{mixed} is no less than that for N_{Neu}, and the space for the former is more restrictive. Hence the claim that $N_{\mathrm{Dir}} \leq N_{\mathrm{mixed}} \leq N_{\mathrm{Neu}}$, and hence the asymptotic law.

To get an asymptotic law for mixed edge-vertex Laplacians (as in Equation (3.7)), the above arguments show it suffices to consider Dirichlet and Neumann eigenvalues for an edge, e. Partition e into k intervals, $I_1, \ldots I_k$, and on a fixed interval, I_j, set $\gamma_{\max}$ to be the maximum value of γ there, and $\beta_{\min}$ similarly. The Rayleigh quotient with $\gamma_{\max}$ replacing γ and $\beta_{\min}$ replacing β is never smaller, and so the Dirichlet eigenvalue counting function on I_j for the Rayleigh quotient with β and γ is at least

$$\left\lfloor \frac{\lambda^{1/2} |I_j| \, \beta_{\min}}{\gamma_{\max} \pi} \right\rfloor.$$

We conclude that for any $\epsilon > 0$ we have that the Dirichlet eigenvalue counting function on e is at least

$$\lambda^{1/2}(1/\pi)\left(-\epsilon + \int_e \beta(x)/\gamma(x)\,dx\right),$$

where x is a standard coordinate on e. We conclude a similar upper bound for the Neumann eigenvalues, and conclude that $N(\lambda) \sim \lambda^{1/2}C/\pi$, where

$$C = \sum_e \int_e \frac{\beta(x)}{\gamma(x)}\,dx.$$

3.3. A condition on edge-based eigenfunctions.

Proposition 3.4. *Let $\mathcal{G}$ be a locally finite graph. Let (f,ω) be an edge-based eigenpair for the Laplacian, and let $\omega = \lambda^{1/2}$. Let v be an interior vertex such that $\omega\ell_e$ is not a multiple of π for any e incident upon v. Then*

$$\sum_{e\sim\{v,u\}} a_e \frac{f(u) - \cos(\omega\ell_e)f(v)}{\sin(\omega\ell_e)} = 0.$$

If the degree of v is infinite, the theorem still holds provided the above sum converges absolutely (and Δ_V is understood in the natural way).

Proof. This is a simple consequence of the fact that if f is an eigenfunction then $\Delta_V f(v) = 0$ at any interior vertex. Fix an edge $e \sim \{v,u\}$, and let x be the standard coordinate on e with $x(v) = 0$ and $x(u) = \ell_e$. We have f's restriction to e with coordinate x, $f_e = f_e(x) = A\cos(\omega x + B)$ for some A, B. Hence

$$f(v) = f_e(0) = A\cos B,$$

and

$$f(u) = f_e(\ell_e) = A\cos B\cos(\omega\ell_e) - A\sin B\sin(\omega\ell_e),$$

and the outward normal derivative of f at v is

$$f'(0) = -A\omega\sin B = -\omega\frac{f(u) - \cos(\omega\ell_e)f(v)}{\sin(\omega\ell_e)}.$$

Since $\Delta_V f = \widetilde{\mathbf{n}}\cdot\nabla f$ at v is $\mathcal{V}(v)^{-1}$ times the sum of the above times a_e, we conclude the proposition. □

Using this proposition one can rather easily determine all the edge-based eigenpairs in terms of the eigenpairs of the "normalized" adjacency matrix of the graph, provided that all edge lengths are equal; the ω's will turn out to be periodic of period 2π. However, we do not know of such a determination when edge lengths vary; we shall show that a graph with two vertices joined by three edges of varying edge lengths will not have periodic ω's.

3.4. The equilength case. In this subsection we consider a finite graph, $\mathcal{G}$, all of whose edges have length 1 (however the a_e can vary). We can similarly deal with any graph whose edge lengths are equal.

For any $v \in \mathring{V}$, let a_v be the sum of the a_e over all edges, e, incident with v (these edges, e, may be incident with boundary vertices). We will assume that $\mathcal{G}$ is connected with at least one edge, so that $a_v > 0$ for all v. Let $\widetilde{A}$ be the "normalized" adjacency matrix, which is just the adjacency matrix of Section 2, normalized by dividing each row by its corresponding a_v. $\widetilde{A}$ represents a Markov chain if and only if $\mathcal{G}$ has no boundary vertices (and it always represents a Markov chain if we add in the boundary vertices and make them "absorbing" states).

Our main result describes the edge-based eigenvalues in terms of $\widetilde{A}$ and the number of edges and interior vertices. We state this first, and then prove it in a series of propositions. First we set

$$\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \dots\},$$

so that for any $\tau \in \mathbb{R}$ we can write

$$\tau + 2\pi\mathbb{Z}_{\geq 0} = \{\tau, \tau + 2\pi, \tau + 4\pi, \dots\}.$$

Theorem 3.5. *Let $\mathcal{G}$ be a connected graph with at least one edge, and let $\widetilde{A}$ be its normalized adjacency matrix, as above. The edge-based eigenvalues is the multiset sum*[3] *of the following sets: for each eigenvalue, λ, of $\widetilde{A}$, there is a unique $\cos^{-1}(\lambda) \in [0, \pi]$; corresponding to this λ we have eigenvalues*

$$\cos^{-1}(\lambda) + 2\pi\mathbb{Z}_{\geq 0}, \qquad \text{and} \qquad 2\pi - \cos^{-1}(\lambda) + 2\pi\mathbb{Z}_{\geq 0}. \tag{3.8}$$

Additionally, the sets

$$\pi + 2\pi\mathbb{Z}_{\geq 0}, \qquad \text{and} \qquad 2\pi + 2\pi\mathbb{Z}_{\geq 0} \tag{3.9}$$

occur with multiplicity $|E| - |\mathring{V}|$. This means that if $|E| - |\mathring{V}| = -1$, i.e., $\mathcal{G}$ is a tree without boundary, then we subtract the list in Equation (3.9) *once from the union over Equation* (3.8)*; i.e., $n\pi$ for nonnegative integer n occurs with multiplicity one.*

Let us mention that if $\mathcal{G}$ has separated boundary, then the Neumann condition at a boundary vertex is equivalent to considering that vertex to be an interior vertex. Hence our theorem really also handles the case where we impose a Dirichlet condition on some vertices, and a Neumann condition on the rest, assuming the rest are separated.

Proof. The proof of this theorem occupies the rest of this section. We prove it in a sequence of propositions. Notice that our proof provides a method for finding a basis for the eigenspaces.

[3] i.e., if λ occurs five times in the lists below, then its multiplicity is five.

Proposition 3.6. *Let $\omega \in \mathbb{R} \setminus (\pi\mathbb{Z})$. Then ω^2 is an edge-based Dirichlet eigenvalue if and only if $\cos\omega$ is an eigenvalue of $\widetilde{A}$. If so, ω^2's multiplicity is that of $\cos\omega$ in $\widetilde{A}$, and for each corresponding eigenfunction, f, of $\widetilde{A}$ (therefore defined on the vertices), we may extend f along each edge (as $A\cos(\omega x + B)$ for some A, B and standard edge coordinate x) to an edge-based eigenfunction.*

Proof. If $\omega \in \mathbb{R} \setminus (\pi\mathbb{Z})$ has ω^2 an edge-based eigenvalue, then by Proposition 3.4 we have for each v

$$\sum_{e\sim\{v,u\}} a_e f(u) = \cos(\omega) f(v) \sum_{e\sim\{v,u\}} a_e.$$

In other words, since f is Dirichlet, $\widetilde{A}f = \cos(\omega)f$; that is, $\cos\omega$ is an eigenvalue of $\widetilde{A}$.

Conversely, let $\cos\omega \neq \pm 1$ be an eigenvalue of $\widetilde{A}$ with eigenfunction g. We claim that for any fixed $e = \{u, v\}$, there is a unique way to extend g to a function along e of the form $A\cos(B + \omega x)$, where x is the standard coordinate on e with $x(v) = 0$. Indeed, consider the equations in A, B for fixed $g(u), g(v)$:

$$g(v) = A\cos B, \qquad g(u) = A\cos(B + \omega).$$

Since

$$A\sin B = \frac{g(u) - g(v)\cos\omega}{-\sin\omega},$$

$A\cos B$ and $A\sin B$ are uniquely determined by $g(u), g(v)$. From what we know of polar coordinates, this means either $A = 0$ and B is arbitrary, in which case g's extension along e is by 0 (and $g(u) = g(v) = 0$), or there is a unique positive $A = A_0$ and unique $B = B_0$ modulo 2π satisfying the equations, with the only other solution being $A = -A_0$ and, modulo 2π, $B = B_0 + \pi$. Since the function $A\cos(B + \omega x)$ is the same for these solutions, and does have the right value at $x = x(v) = 0$ and $x = x(u) = 1$, g can be extended to satisfy $\Delta_E g = \omega^2 g$. Also clearly g satisfies the condition in Proposition 3.4, which is equivalent to $\Delta_V g = 0$.

To sum up, we know that ω^2 is an edge-based eigenvalue for $\omega \in \mathbb{R} \setminus (\mathbb{Z}\pi)$ if and only if $\cos\omega$ is an eigenvalue of $\widetilde{A}$. We know that the restriction of any Δ_E eigenfunction gives an $\widetilde{A}$ eigenfunction, and we know that (given ω) a $\widetilde{A}$ eigenfunction g has a unique extension to a Δ_E eigenfunction. Hence the multiplicities of ω^2 in Δ_E and $\cos\omega$ in $\widetilde{A}$ are equal. □

The story when $\omega \in \pi\mathbb{Z}$ is less elegant. Indeed, there are many edge-based eigenfunctions, f, whose restriction to the vertices vanishes.

Let Y_ω be the eigenspace corresponding to the Dirichlet edge-based eigenfunctions with eigenvalue ω^2,

$$Y_\omega = Y_\omega(\mathcal{G}) = \{f \in C^\infty_{\mathrm{Dir}}(\mathcal{G}) \mid \Delta_E f = \omega^2 f, \Delta_V f = 0\},$$

and let Z_ω be those elements of Y_ω vanishing on all vertices,

$$Z_\omega = Z_\omega(\mathcal{G}) = \{f \in Y_\omega \mid f|_V = 0\}.$$

We reduce the study of Y_ω for $\omega \in \pi\mathbb{Z}$ to that of Z_ω by the following proposition:

Proposition 3.7. *Y_ω/Z_ω for $\omega \in 2\pi\mathbb{Z}$ is one-dimensional if $\partial\mathcal{G} = \emptyset$, and otherwise zero. Similarly for $\omega \in \pi + 2\pi\mathbb{Z}$, except that we require $\partial\mathcal{G} = \emptyset$ and that $\mathcal{G}$ is bipartite.*

Proof. If $f \in Y_{2\pi n}$ with $n \in \mathbb{Z}$, and $e = \{u, v\}$ is an edge, then $f(u) = f(v)$, since f's restriction to e is of the form $A\cos(B + 2\pi n x)$. Hence f must be constant on V. This implies that Y_ω/Z_ω is at most one-dimensional, and must be zero if $\partial\mathcal{G} \neq \emptyset$. On the other hand, if $\partial\mathcal{G} = \emptyset$ then the function whose restriction to each edge is $\cos(\omega x)$ gives a nonzero element of Y_ω/Z_ω. The case $\omega = (2n+1)\pi$ is handled similarly. □

The next two propositions essentially finish our work in this section.

Proposition 3.8. *For $\omega > 0$ there is a natural isomorphism of Y_ω with $Y_{\omega+2\pi}$ that restricts to an isomorphism of Z_ω with $Z_{\omega+2\pi}$. The same is true if $\omega + 2\pi$ is replaced by $\omega + \pi$, provided that $\mathcal{G}$ is bipartite.*

Proof. For an arbitrary $f \in Y_\omega$, let ιf be the function whose restriction to e is $A\cos\big(B + (2\pi + \omega)x\big)$, where A, B are given by f's restriction to e being $A\cos(B + \omega x)$. Then $\Delta_V f = \Delta_V(\iota f)\omega/(\omega + 2\pi)$ so $\Delta_V(\iota f) = 0$. From here it is clear that ι is the desired isomorphism. □

Proposition 3.9. *Let $\mathcal{G}' = \mathcal{G} \cup \{e\}$ be the graph formed by adding an edge, e, to $\mathcal{G}$. Then $Z_\omega(\mathcal{G})$ naturally injects into $Z_\omega(\mathcal{G}')$, and the quotient is of dimension 1 or zero.*

Proof. If $f \in Z_\omega(\mathcal{G})$, we simply extend it by zero on e to get a member of $Z_\omega(\mathcal{G}')$. Member of $Z_\omega(\mathcal{G}')$ restricts to $A\sin(\omega x)$ along e for some A, and hence any two of them are scalar multiples of each other modulo $Z_\omega(\mathcal{G})$. □

First a corollary of these two propositions:

Corollary 3.10. *Let b be the number of ± 1's that appear among $\widetilde{A}$'s eigenvalues, i.e.,*

$$b = b(\mathcal{G}) = \begin{cases} 0 & \text{if } \partial\mathcal{G} \neq \emptyset, \\ 1 & \text{if } \partial\mathcal{G} = \emptyset \text{ and } \mathcal{G} \text{ is not bipartite, and} \\ 2 & \text{if } \partial\mathcal{G} = \emptyset \text{ and } \mathcal{G} \text{ is bipartite.} \end{cases}$$

Then

$$\dim Y_\pi + \dim Y_{2\pi} + 2(|\mathring{V}| - b) = 2|E|.$$

Hence, if $\mathcal{G}$ is bipartite, we further have

$$\dim Y_\pi = \dim Y_{2\pi} = |E| - |\mathring{V}| + b.$$

Proof. The first part follows from the above and Weyl's law. The second part follows from Proposition 3.8. □

To prove Theorem 3.5, first note that since we are working with the Dirichlet condition we can assume the boundary is separated — if not, we just give each boundary edge its own boundary vertex.

The corollary above shows that the theorem is true for a tree. Any connected graph is the union of a tree and a number of edges. Hence it suffices to show that if the theorem is true for $\mathcal{G}$ then it is true for $\mathcal{G}$ with an edge thrown in, $\mathcal{G}' = \mathcal{G} \cup \{e\}$.

So assume the theorem is true for $\mathcal{G}$. If $b(\mathcal{G}) = b(\mathcal{G}')$ then $\dim Y_\pi + \dim Y_{2\pi}$ increases by 2. But each dimension can increase by at most 1, so they increase precisely by that much, and the theorem holds for $\mathcal{G}'$.

The only change in b that can happen by adding an edge is that $b(\mathcal{G}) = 2$ but $b(\mathcal{G}') = 1$, i.e., $\mathcal{G}$ is bipartite but $\mathcal{G}'$ isn't. In this case $\dim Y_\pi + \dim Y_{2\pi}$ remains the same. While Z_π cannot decrease in going from $\mathcal{G}$ to $\mathcal{G}'$, Y_π/Z_π goes from one to zero-dimensional from $\mathcal{G}$ to $\mathcal{G}'$. Since $Y_{2\pi}/Z_{2\pi}$ remains one-dimensional, $Z_{2\pi}$ increases by one. Once again we see how the Z_ω's and Y_ω/Z_ω's change for $\mathcal{G}'$, and it is easy to see the theorem holds there. □

4. The wave equation

In this section we usually only assume that the graphs are locally finite. This is because the wave equation has finite propagation speed, and "cannot tell" whether or not a graph is finite (in any fixed interval of time).

Definition 4.1. Given a graph, fix nonnegative $\alpha, \beta \in C^0(\mathcal{G})$, nonnegative $\gamma \in C^1(\mathcal{G})$, and an interval $I \subset \mathbb{R}$. A function $u = u(x,t) : \mathcal{G} \times I \to \mathbb{R}$ is said to satisfy the wave equation with coefficients α, β, γ if:

(1) u is continuous on $\mathcal{G} \times I$ and $u(\,\cdot\,, t) \in C^2(\mathcal{G})$ for all $t \in I$;
(2) for all $x \in \mathring{\mathcal{G}}$ and $t \in \mathring{I}$ the derivative u_{tt} exists at (x,t) and is continuous in x; and
(3) for fixed t we have
$$(\alpha\, d\mathcal{V} + \beta\, d\mathcal{E})u_{tt} = \nabla \cdot (\gamma \nabla u)$$
as integrating factors.

If u vanishes on $\partial\mathcal{G} \times I$ we say that u satisfies the *Dirichlet condition*; similarly for the *Neumann condition* if ∇u vanishes along all edges at all boundary vertices.

Having the equality above as integrating factors means that
$$\alpha u_{tt} = -\gamma \Delta_V u$$
at all (x,t) with $x \in \mathring{V}$ and $t \in \mathring{I}$, and that
$$\beta u_{tt} = -\gamma \Delta_E u + \nabla\gamma \cdot \nabla u$$

at all (x,t) with $x \in \mathring{\mathcal{G}} \setminus V$ and $t \in \mathring{I}$.

In the above definition we may also allow mixed boundary conditions as in Equations (3.4) and (3.5).

The the vertex-based wave equation, we mean the wave equation with coefficients $\alpha = \gamma = 1$ and $\beta = 0$. In this case $\Delta_E u = 0$ and so u must be edgewise linear. As remarked in Section 2, Δ_V on a finite graph can be viewed as a bounded operator on $L^2_{\mathrm{Dir}}(\mathcal{G},\mathcal{V})$. As such, for any edgewise linear $f \in L^2_{\mathrm{Dir}}(\mathcal{G},\mathcal{V})$ we can form

$$u(x,t) = \cos\bigl(t\sqrt{\Delta_V}\,\bigr)f = f - t^2\Delta_V f/2 + t^4\Delta_V^2 f/4! - \cdots$$

and we easily see it satisfies the wave equation with $u(x,0) = f(x)$. The boundedness of Δ_V implies that if $f = \chi_v$ is the edgewise linear characteristic function[4] at v, then for fixed x,

$$u(x,t) = \cos\bigl(t\sqrt{\Delta_V}\,\bigr)\chi_v = (-1)^d t^{2d} N_{x,v}/(2d)! + O(t^{2d+2})$$

where d is x's distance to v and $N_{x,v}$ is the number of paths from x to v of length d. It follows that $u(x,t) > 0$ for small t, and so the vertex-based wave equation does not have a finite wave propagation speed.

Since traditional graph theory is based on Δ_V restricted to edgewise linear functions, the above explains why approaching the wave equation with traditional graph theory leads to unsatisfactory results.

A much better model of the wave equation appearing in analysis is the edge-based wave equation, where $\beta = \gamma = 1$ and $\alpha = 0$. We will show in the next subsection that these waves propagate "along the edges" and have finite wave speed equal to 1.

4.1. The energy inequality. Many basic properties of the wave equation follow from well-known energy inequality, which we state and apply in this subsection.

If $A \subset \mathcal{G}$, then the energy of $u = u(x,t)$ over A at time t is defined to be

$$\mathrm{Energy}(A;t) = \int_A \bigl(\gamma(\nabla u)^2\,d\mathcal{E} + u_t^2(\alpha\,d\mathcal{V} + \beta\,d\mathcal{E})\bigr).$$

For real $h > 0$ let

$$A^h = \bigl\{x \in \mathcal{G} \mid \mathrm{dist}(x,A) < h\bigr\}.$$

Now fix coefficients α,β,γ for a wave equation and let c be the smallest constant such that $\gamma \le c^2\beta$ throughout $\mathcal{G}$ (we assume this c exists).

Theorem 4.2. *Let A be an open set with A^{ct_0} of finite type for some $t_0 > 0$. Then if u is a solution to the Dirichlet or Neumann wave equation we have*

$$\mathrm{Energy}(A^{ct_0};0) \ge \mathrm{Energy}(A;t_0).$$

[4] i.e., this function is 1 on v, 0 on other vertices, and edgewise linear.

The same holds of the mixed boundary condition, as in Equations (3.4) *and* (3.5)*, provided we add*

$$\int_{K_2 \cap A} \gamma \sigma u^2 \, d\mathcal{V}$$

to the energy.

This theorem will be proven in the next subsection; its proof is virtually identical to its well-known proof in analysis.

We remark that the wave equation has a time symmetry, in that if u satisfies the wave equation then so does $w(x,t) = u(x,-t)$. We may therefore conclude the symmetric fact that

$$\text{Energy}(A;0) \le \text{Energy}(A^{ct_0};t_0).$$

From the energy inequality we easily conclude the following proposition:

Proposition 4.3. *Assume $\mathcal{G}$ is locally finite, that β is strictly positive on $\mathcal{G}$, and that c exists as before (i.e., $\gamma \le c^2\beta$ everywhere). Let u be a solution to the wave equation on $\mathcal{G} \times [0,T]$ with $u(x,0) = 0$ and $u_t(x,0) = 0$ for all x within a distance ct to a fixed $y \in \mathcal{G}$. Then $u(y,t) = 0$.*

Proof. For any $\epsilon > 0$ we have

$$\text{Energy}(\{y\}^{c(t-\epsilon)};0) = 0.$$

We conclude that $u_t(y,s)$ vanishes from $s = 0$ to $s = t - 2\epsilon$, and hence $u(y,t-2\epsilon) = 0$. Now we let $\epsilon \to 0$ and use the continuity of u. □

Some immediate corollaries of this are:

Corollary 4.4. *If u, w are two solutions to the wave equation such that u and w agree at time $t = 0$ on all points within distance ct to y, and the same for u_t and w_t, then $u(y,t) = w(y,t)$. In other words, the value of $u(y,t)$ depends only on the value of u at a fixed time in the "space-time cone of speed c at y". In other words, this wave equation has finite speed of wave propagation bounded by c.*

Corollary 4.5. *Fixing $u(\,\cdot\,,0)$ and $u_t(\,\cdot\,,0)$, there is at most one solution, $u(x,t)$ for $t > 0$, to the wave equation.*

4.2. A proof of the energy inequality. Let $I \subset \mathbb{R}$ be an interval. By a (*graph-time*) *vector field* on $\mathcal{G} \times I$ we mean a pair (G,F) where $F = F(t)$ is an integrating factor on $\mathcal{G}$ and $G = G(t)$ is vector field on $\mathcal{G}$ both depending on $t \in I$. By its *divergence* we mean

$$\nabla_{\text{gt}} \cdot (G,F) = \nabla \cdot G + F_t,$$

which is an integrating factor that depends on time, t, where by F_t we mean the partial derivative of F with respect to t, i.e., we differentiate F's $d\mathcal{V}$ component and its $d\mathcal{E}$ component with respect to time.

Proposition 4.6. *Consider a divergence free graph-time vector field, (G, F), i.e., $\nabla \cdot G + F_t = 0$ as a boundary inclusive equality. Write $F = F_V\, d\mathcal{V} + F_E\, d\mathcal{E}$. Assume that $F_V \geq 0$ at all vertices and that $cF_E \geq |G|$ on $(\mathcal{G} \setminus V) \times I$. If A is any open set with A^{ct_0} of finite type, and if $0, t_0 \in I$, then*

$$\mathcal{I}(t) = \int_{A^{c(t_0-t)}} F(t)$$

is a nonincreasing function in $t \in [0, t_0]$.

The energy inequality in Section 4.1 follows almost at once by taking

$$F = \gamma\mathcal{E}(\nabla u)^2 + (\alpha\mathcal{V} + \beta\mathcal{E})u_t^2 \quad \text{and} \quad G = -2\gamma u_t \nabla u,$$

adding $\gamma\sigma u^2\mathcal{V}|_{K_2}$ to F for the mixed boundary condition.

Proof. Let $S(t) = A^{c(t_0-t)}$. If $\partial S(t)$ contains a vertex, then $\mathcal{I}(t)$ is right continuous at t, and has a jump (if any) from the left of

$$\mathcal{I}(t-0) - \mathcal{I}(t) = \sum_{x \in \mathring{V} \cap \partial S(t)} F_V(x)\mathcal{V}(x).$$

Next partition $\partial S(t)$ into interior and boundary points,

$$\mathring{B} = \partial S(t) \cap \mathring{\mathcal{G}} \qquad \text{and} \qquad B^{\partial} = \partial S(t) \cap \partial\mathcal{G}.$$

$\mathring{B}$ will contain no vertices for all but finitely many t. If $\mathring{B}$ contains no vertex, then using Proposition 2.21 we see that

$$\mathcal{I}'(t) = -\sum_{x \in \mathring{B}, e \ni x} F_E(x,t) c a_e + \int_{S(t) \cup B^{\partial}} F_t(t)$$

(where F_t contains both a $d\mathcal{V}$ term and a $d\mathcal{E}$ term). By taking $S(t)$ and adding to it $\partial S(t)$ as new vertices each of weight 1, we get a graph, $\mathcal{G}_t$; taking $F_V = 0$ at all new vertices, we may write:

$$\begin{aligned}\int_{S(t) \cup B^{\partial}} F_t(t) &= \int_{\mathcal{G}_t \setminus \mathring{B}} F_t(t) = -\int_{\mathcal{G}_t \setminus \mathring{B}} \nabla \cdot G(t) \\ &= \int_{\mathring{B}} \widetilde{\mathbf{n}} \cdot G(t) = \sum_{x \in \mathring{B}, e \ni x} \mathbf{n}_{e,x} \cdot G(x,t) a_e,\end{aligned}$$

using the divergence theorem.

Recalling that $\mathbf{n}_{e,x}$ is a unit vector, we have

$$-F_E(x,t)c + \mathbf{n}_{e,x} \cdot G(x,t) \leq -F_E(x,t)c + |G(x,t)| \leq 0$$

for all x, t. Hence

$$\mathcal{I}'(t) = \sum_{x \in \mathring{B}, e \ni x} a_e\big(-F_E(x,t)c + \mathbf{n}_{e,x} \cdot G(x,t)\big) \leq 0.$$

We have shown that $\mathcal{I}(t)$ is nonincreasing at all but finitely many points (i.e., the t's with $\mathring{B}$ containing a vertex), and at these finitely many points we know that $\mathcal{I}(t)$ is continuous or has a decreasing jump. Hence $\mathcal{I}(t)$ is nonincreasing. □

The proposition is usually proven (in analysis) by invoking a space-time divergence theorem on a truncated cone in space-time (or graph-time here), such as those (x,t) with $\mathrm{dist}(x,A) < c(t_0 - t)$ (see, for example, [**Smo83**] Chapter 4 or [**CH89**]). While this approach also works, setting up a graph-time divergence theorem seems like more trouble than it's worth for our purposes at this point.

4.3. More general wave equations. One can generalize the uniqueness results for the wave equation, i.e., Corollaries 4.4 and 4.5, to the same results for a wave equation of the form

$$(\alpha\, d\mathcal{V} + \beta\, d\mathcal{E})u_{tt} = \nabla\cdot(\gamma\nabla u) + \delta\cdot\nabla u + \epsilon u,$$

where δ is a C^0 vector field and ϵ is a C^0 function, and we assume β never vanishes. The proof is a simple adaptation of the analysis proof given in, for example, Smoller's book [**Smo83**]. We define the energy exactly as before (ignoring δ and ϵ), but now prove

$$e^{Kt_0}\mathrm{Energy}(A^{ct_0};0) \geq \mathrm{Energy}(A;t_0) \tag{4.1}$$

for some constant K depending on δ,ϵ, provided that u, u_t vanish on A^{ct_0} at $t=0$. Uniqueness with "propagation speed" at most c follows as before.

We shall outline the proof of Equation (4.1) when $\delta = 0$; the general case is a bit messier but similar (see [**Smo83**] for details). First we notice that if all is the same as in Proposition 4.6 except that $\nabla\cdot G + F_t$ does not necessarily vanish (where by F and G we mean $F = \gamma\mathcal{E}(\nabla u)^2 + (\alpha\mathcal{V}+\beta\mathcal{E})u_t^2$ and $G = -2\gamma u_t\nabla u$), then we have

$$\mathcal{I}(t) \leq \mathcal{I}(0) + \int_0^t \int_{A^{c(t_0-s)}} (\nabla\cdot G + F_t)\, ds.$$

Hence setting

$$E(t) = \mathrm{Energy}(A^{c(t_0-t)};t)$$

for $0 \leq t \leq t_0$, we have for such t

$$E(t) \leq E(0) - 2\int_0^t \int_{A^{c(t_0-s)}} \epsilon u(x,s)u_t(x,s)\, d\mathcal{E}(x)\, ds. \tag{4.2}$$

The integral on the right-hand side can be bounded by

$$\left|\int_0^t \int_{A^{c(t_0-s)}} uu_t\right| \leq \int\int u^2/2 \;+\; \int\int u_t^2/2.$$

The $u_t^2/2$ term integrated over space is bounded by a constant times $E(t)$ (since β vanishes nowhere), and the $u^2/2$ can be bounded by a constant

times t^2 times a u_t^2 integral by Poincaré's inequality (see [**Smo83**]). It easily follows that the double integral in Equation (4.2) is bounded by a constant times $\int_0^t E(s)\,ds$, and hence

$$E(t) \leq E(0) + K \int_0^t E(s)\,ds$$

for a constant K. Gronwall's inequality implies $E(t) \leq e^{Kt}E(0)$, which is just Equation (4.1).

4.4. Differentiability and the wave equation. In this subsection we show how to prove the existence of a solution to the wave equation for sufficiently "differentiable" initial conditions. On the circle, i.e., $\mathbb{R}/(2\pi\mathbb{Z})$, there is a rough correspondence between differentiability and having Fourier coefficients decaying. We shall use the same for graphs, to give a nice description of when we are sure that the wave equation has a solution.

Let $f_1, f_2, \ldots$ be the eigenfunctions of Δ_E with eigenvalues $\lambda_1, \lambda_2, \ldots$ with some boundary conditions of any type specified before. Let D^k (which depends on the boundary conditions) be those formal sums $\sum_i a_i f_i$ with $\sum_i i^k |a_i| < \infty$ (here k is any real, although typically a nonnegative integer). Let B^k be the same, except that the a_i's are subject to the weaker condition that $i^k|a_i|$ is bounded in i. Let Diff^k be those functions $f \in C^k$ such that for $j = 0, \ldots, k-1$ we have:

(1) $f^{(j)}(v, e)$ depends only on v if j is even, where $f^{(j)}(v, e)$ is the j-th derivative of f at v along e.
(2) The sum of $f^{(j)}(v, e)$ over all e for fixed v is zero for all v, if j is odd.

It is not hard to prove the following facts:

Proposition 4.7. *We have natural inclusions $B^{k+1+\epsilon} \subset D^k \subset B^k$ for any k and $\epsilon > 0$, and inclusions $D^k \subset \mathrm{Diff}^k \subset B^k$ for any nonnegative integer k.*

We remark that Diff^k's compatibility with D^k, B^k makes it, in some sense, a better notion of differentiability than C^k defined in Section 2.

Proof. The first inclusions are straightforward. The inclusion $D^k \subset \mathrm{Diff}^k$ follows by viewing the formal sum as an absolutely convergent sum of functions whose derivatives up to k-th order also form an absolutely convergent sum (here we use the periodic nature of the eigenpairs). Finally the inclusion $\mathrm{Diff}^k \subset B^k$ follows by integration by parts of the inner product (f, f_i) for an $f \in \mathrm{Diff}^k$ along each edge: $\int_0^1 f(x) A_e \cos(\omega x + B_e)$ is proportional to the integral of $\omega^{-k} f^{(k)}(x)$ times sine or cosine of plus or minus $\omega x + B_e$ plus boundary terms. The boundary terms at $v \in \mathring{V}$ are proportional to sums over e of $f^{(j)}(v, e) A_e$ times sine or cosine $\omega x + B_e$. Knowing that $A_e \cos(B_e)$ is independent of e (for a given v, with $x = 0$ corresponding to v), and knowing that the sum of $A_e \sin(B_e)$ vanishes (since $\Delta_V f_i = 0$ for all i), we

see that the boundary terms at $\mathring{V}$ disappear; similarly we see that boundary terms at boundary vertices vanish. Now we use Weyl's law and the fact that $f \in C^k$ to see that $i^k(f, f_i)$ is bounded. □

We now state an existence theorem in terms of D^2; it follows from the above proposition that our theorem also applies to the class Diff^4, which is easier to understand, in a sense, than D^2.

Proposition 4.8. *Let $g, h \in D^2$. If $g = \sum a_i f_i$ and $h = \sum b_i f_i$, then*

$$u(x,t) = \sum_i f_i\Big(a_i \cos(\sqrt{\lambda_i}\, t) + \frac{b_i}{\sqrt{\lambda_i}} \sin(\sqrt{\lambda_i}\, t)\Big) \tag{4.3}$$

is a solution to the wave equation with $u(\,\cdot\,, 0) = g$ and $u_t(\,\cdot\,, 0) = h$.

Proof. We need to know that u_{tt} exists, and if $g, h \in D^2$ then the sum of twice differentiated terms is absolutely convergent and this u_{tt} exists. The rest is an easy verification. □

Any g, h in L^2, say, will have eigenfunction expansions. It makes sense to define $u(x,t)$ by the formal sum above (in Equation (4.3)), which for fixed t will always lie in L^2 (although u_{tt} need not exist).

4.5. Chebyshev polynomials and the wave operator. We again assume that $\mathcal{G}$ is finite in this subsection. Let f_i, λ_i be as in the Section 4.4, and set Summ to be D^0 in the notation of the previous section, i.e.,

$$\mathrm{Summ} = \Big\{ \sum_i a_i f_i \;\Big|\; \sum |a_i| < \infty \Big\}.$$

Elements of Summ may be viewed as formal sums, or we may identify them with the bounded function on $\mathcal{G}$ to which they converge (since the f_i are uniformly bounded). Summ is the set of functions with a summable eigenfunction coefficient series; it is easy to see that it contains, for example, $H^1(\mathcal{G})$. The map sending f_i to its restriction on vertices extends to a continuous map from Summ to $L^\infty(\mathcal{G}, \mathcal{V})$. Since $\mathcal{G}$ is finite this gives rise to a continuous map $M_{E\to V} : \mathrm{Summ} \to L^2(\mathcal{G}, \mathcal{V})$.

If $g \in \mathrm{Summ}$ with $g = \sum a_i f_i$, then

$$\big(\cos\sqrt{\Delta_E}\,\big) g = \sum a_i \big(\cos\sqrt{\lambda_i}\,\big) f_i$$

lies in Summ, and we know

$$M_{E\to V}\big(\cos\sqrt{\Delta_E}\,\big) g = \sum a_i \big(\cos\sqrt{\lambda_i}\,\big) M_{E\to V} f_i.$$

Since $1 - \cos\sqrt{\lambda_i}$ is the Δ_V eigenvalue corresponding to $M_{E\to V} f_i$ and $\widetilde{A} = I - \Delta_V$, we have

$$\widetilde{A} M_{E\to V} = M_{E\to V} \cos\sqrt{\Delta_E}$$

as maps from Summ to $L^2(\mathcal{G},\mathcal{V})$, assuming all edge lengths are 1. It follows that for any polynomial, P, we have

$$P\big(\widetilde{A}\big)M_{E\to V} = M_{E\to V}P\big(\cos\sqrt{\Delta_E}\,\big).$$

If T_k is the k-th Chebyshev polynomial, given by

$$T_k(\cos x) = \cos(kx),$$

we have

$$T_k\big(\widetilde{A}\big)M_{E\to V} = M_{E\to V}\cos\big(k\sqrt{\Delta_E}\,\big).$$

We conclude the following theorem:

Theorem 4.9. *Let all edge lengths be* 1. *Let* $g \in$ Summ*, and let* u *be the formal solution to the wave equation as in Equation* (4.3) *with* $h = 0$. *Then for any integer* k *and interior vertex* v *we have*

$$u(v,k) = T_k\big(\widetilde{A}\big)M_{E\to V}g.$$

So if we wish to know this wave equation solution at vertices and at only integral times, we need only know the initial condition at the vertices. The values of the solution there are given in term of Chebyshev polynomials of the normalized adjacency matrix. Notice that knowing the values of u at nonintegral times requires knowing f along the edges.

Notice that Theorem 4.9 is valid for infinite (locally finite) graphs, since the wave equation solution is determined locally for any finite time, and the equality in Theorem 4.9 is a local statement as well.

We finish this subsection by remarking that if we ignore the map $M_{E\to V}$, we can say that $\widetilde{A}$ "acts like" $\cos\sqrt{\Delta_E}$. Noting that $\widetilde{A}$ is $I - \Delta_V$ for an appropriate graph theoretic Laplacian, Δ_V, we can see that $I - \Delta_V$ "acts like" $\cos\sqrt{\Delta_E}$.

4.6. Wave propagation through vertices. Using Theorem 4.9 it is not hard to see what happens when a wave is sent through a vertex. More precisely, let G be the infinite d-regular star,[5] i.e., G's vertices are v_0 union $v_{i,j}$ with $i = 1,\dots,d$ and j a positive integer, and G's edges are $\{v_0, v_{i,1}\}$ for all i, and $\{v_{i,j}, v_{i,j+1}\}$ for all i and for all positive j (see Figure 2).

Taking g to be a function which is zero on all vertices except $v_{1,j}$ for some $j \geq 2$, we apply Theorem 4.9 to see that of the "wave" traveling towards v_0, we have $(2/d) - 2$ of the wave comes back along the $i = 1$ edge, and $2/d$ of it travels down each $i > 1$ edge.

This motivates the following theorem. We state this theorem in terms of the length 1 d-regular star, by which we mean the graph, G, as above,

[5]Note that the wave equation effectively ignores vertices of degree 2, i.e., one gets the same equation if one treats the two edges incident with the vertex as one longer edge. So the infinite d-regular star is equivalent to a graph with one degree d vertex and d edges of infinite length.

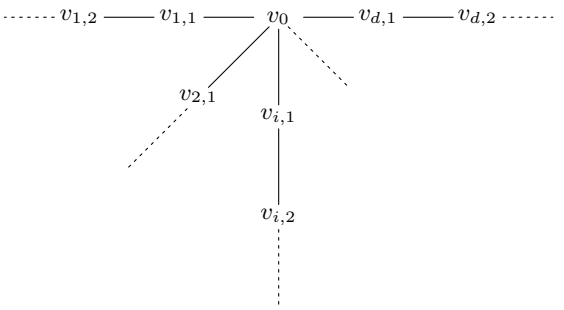


Figure 2.

except we restrict j to take on only the value 1. We endow the edges of G with standard coordinates $x_1, \ldots, x_d$ where $x_i(0) = v_0$ and $x_i(1) = v_{i,1}$.

Theorem 4.10. *Let $u(x,t)$ be the solution to the wave equation for $0 \le t \le 1/4$ given by*

$$u(x_i,t) = \begin{cases} f(x_1+t) & \text{for } i=1, \\ 0 & \text{otherwise,} \end{cases}$$

where f is any twice differentiable function supported on $(1/4, 3/4)$. Then the solution for $0 \le t \le 5/4$ to the wave equation exists and is given by

$$\widetilde{u}(x_i,t) = \begin{cases} f(x_1+t) + \big((2/d)-1\big) f(t-x_1) & \text{for } i=1, \\ (2/d) f(t-x_i) & \text{otherwise.} \end{cases}$$

Notice that this theorem tells us how waves propagate through vertices. Also, notice that we know that $\widetilde{u}$ is unique.

Proof. Clearly $\widetilde{u}$ satisfies the wave equation on edge interiors. At v_0, the only vertex of interest for $t \le 5/4$, we have $\widetilde{u}$ is continuous (taking the limiting value $2f(t)/d$ along each edge at v_0) and satisfies $\Delta_V \widetilde{u} = 0$ there. Hence $\widetilde{u}$ satisfies the wave equation. □

4.7. Finite propagation speed of wave operators. There are a number of operators on $L^2(\mathcal{G}, \mathcal{E})$ that arise from the wave operator, which have a "finite speed of propagation". We mention one classical one, and a generalization of it. First we need some definitions.

By the support of a function, f, in $L^2(\mathcal{G}, \mathcal{E})$ we mean the complement of the union of those open sets, U, for which $f = 0$ almost everywhere in U.

Definition 4.11. Let A_t be a family of bounded (everywhere defined) operators on $L^2(\mathcal{G}, \mathcal{E})$ indexed on $t \ge 0$. We say that A_t have *speed of propagation at most c* if $(A_t f, g) = 0$ for any $f, g \in L^2(\mathcal{G}, \mathcal{E})$ and t with the supports of f and g a distance at least ct apart.

Definition 4.12. A subset, D, of $L^2(\mathcal{G},\mathcal{E})$ is called *supportingly dense* if for any $f \in L^2(\mathcal{G},\mathcal{E})$ and $\epsilon > 0$ there is $\widetilde{f} \in D$ such that $\|f - \widetilde{f}\|_2 < \epsilon$ and (each point of) the support of $\widetilde{f}$ is within distance ϵ of the support of f.

In our propagation speed definition, rather than requiring $f, g \in L^2(\mathcal{G},\mathcal{E})$, it suffices to take $f, g \in D$ where D is any supportingly dense subset of $L^2(\mathcal{G},\mathcal{E})$. We also remark that standard mollification arguments show that for any k, Diff^k is supportingly dense; it follows that B^k and D^k are, as well.

Consider the operator

$$W_t = \cos\bigl(t\sqrt{\Delta}\,\bigr).$$

By the spectral theorem this can be viewed as an operator on $L^2(\mathcal{G},\mathcal{E})$ whose norm is bounded by 1. We know by Proposition 4.8 that W_t restricted to $D^2 \cap L^2(\mathcal{G},\mathcal{E})$ has propagation speed at most 1. Since D^2 is supportingly dense in $L^2(\mathcal{G},\mathcal{E})$ we see that W_t has speed of propagation ≤ 1.

Next for any $a \in \mathbb{R}$ consider the operator:

$$W_{t,a} = h\bigl(t^2(\Delta - a)\bigr),$$

where

$$h(x) = 1 - \frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \cdots = \begin{cases} \cos\sqrt{x} & \text{if } x \geq 0, \\ \cosh\sqrt{-x} & \text{if } x < 0. \end{cases}$$

So $W_{t,0}$ is just W_t as above. Since Δ is positive semidefinite, $\|W_{t,a}\| \leq \cosh\bigl(t\sqrt{a}\,\bigr)$. The analogue of Proposition 4.8 for the wave equation $u_{tt} = -\Delta u + au$ is easily verified, and we conclude (using Subsection 4.3) as we did for W_t that:

Proposition 4.13. *For fixed a, the propagation speed of $W_{t,a}$ is ≤ 1.*

5. Applications

In this section we give examples of how to apply our edge-based approach to get graph theoretic results from analysis results. We give a new bound on eigenvalues based on set distances in graph theory; the bound it gives on diameters can be better or worse than the well-known bound of Chung, Faber, and Manteuffel (in [**CFM94**]); our technique is very simple and works in analysis (to give the result of Friedman and Tillich in [**FT**]). We also show, for example, that the graph diameter inequalities of Chung, Grigor'yan, and Yau in [**CGY96**, **CGY97**], which mildly generalize that of Chung, Faber, and Manteuffel (in [**CFM94**]), is optimal to first-order, in a certain sense, for small Laplacian eigenvalue. We give other results that illustrate our ability to translate from analysis to graph theory, although these other results do not improve the best known graph theory results.

5.1. Distances and diameter: using old results. A number of articles have inequalities relating distances to sets or diameters of a space to Laplacian eigenvalues, both for manifolds and graphs (see [**AM85**, **Moh91**, **LPS88**, **CFM94**, **CGY96**, **BL97**, **CGY97**]). In [**FT**] the philosophy that $I - \Delta_G$ "acts like" $\cos\sqrt{\Delta_E}$ of the preceding section is used to apply graph theoretic techniques (namely those of [**CGY97**]) to analysis and get an improved eigenvalue bound based. Here we go the other way, taking analysis results or more general results, and apply them to get graph theoretic results. The result we obtained is not the best known, but it is better than some results, especially one that can be derived from the same technique using the vertex-based Laplacian. However, in the next subsection we give a distances/eigenvalues technique that yields new results in graph theory.

Consider the result of Bobkov and Ledoux, which states that for any metric probability space, (M, ρ, μ), and disjoint Borel sets X, Y we have

$$\sqrt{\lambda}\,\rho(X,Y) \le -\log(\mu X\ \mu Y) \tag{5.1}$$

where $\rho(X,Y)$ is the distance from X to Y and λ is the optimal constant in the Poincaré inequality (in other words, λ is the first nonzero Neumann eigenvalue in the case of a compact manifold or finite graph). Let us apply this to bounding the diameter of a graph.

One way is to directly apply Equation (5.1) to the vertex-based situations, with $\lambda = \lambda_V$, the vertex-based Laplacian eigenvalue. We get $\mu X = |X|/n$, and we take X, Y to consist of single points of distance D where D is the diameter. We conclude that $\sqrt{\lambda_V}\ D \le \log(n^2) = 2\log n$, or

$$D \le \frac{2\log n}{\sqrt{\lambda_V}}.$$

Here λ_V is the first nonzero eigenvalue of the Laplacian $I - \widetilde{A}$ where $\widetilde{A}$ is the normalized adjacency matrix. Notice that if our graph is d-regular, then $\lambda_V = \lambda_T/d$ where λ_T is the first nonzero eigenvalue of the traditional graph theoretic Laplacian. We conclude:

$$D \le 2\log n\sqrt{d/\lambda_T}.$$

Alternatively we may apply Equation (5.1) to the edge-based Laplacian. Notice that the similar and slightly weaker inequalities of [**CGY96**, **CGY97**] require the Laplacian, Δ, to have $\cos\bigl(t\sqrt{-\Delta}\,\bigr)$ extend supports of functions by a distance at most t. So our edge-based Laplacian could be used with these earlier results, whereas the vertex-based Laplacian could not. We take X, Y to be balls of size $1/2$ about two points of distance D, the diameter. We conclude:

$$\sqrt{\lambda_E}\ (D-1) \le \log(n^2) = 2\log n.$$

Seeing as $\sqrt{\lambda_E} = \cos^{-1}(1 - \lambda_V)$, we have

$$D - 1 \leq \frac{2 \log n}{\cos^{-1}(1 - \lambda_V)}.$$

But it is easy to see that $\cos^{-1}(1 - a) \geq \sqrt{2a}$ for any a. Hence we have

$$D - 1 \leq \sqrt{2/\lambda_V} \ \log n.$$

In the d-regular case this gives

$$D - 1 \leq \sqrt{(2d)/\lambda_T} \ \log n. \tag{5.2}$$

We conclude:

(1) With the same technology, i.e., using Equation (5.1), the edge-based technique does better than the vertex-based technique by roughly a factor of $\sqrt{2}$.
(2) In previous Laplacian bounds (e.g., [**CGY96**, **CGY97**]) the edge-based technique can be applied whereas the vertex-based technique cannot (by support extending restrictions that are equivalent to the wave equation having propagation speed 1).

It is interesting to note that the edge-based diameter bound we derived (in Equation (5.2)) improves the original Alon–Milman bound[6] of

$$D \leq 2\lfloor \sqrt{(2d/\lambda_T)} \ \log_2 n \rfloor$$

(see [**AM85**]), in the case where d/λ_T and n are large. Similarly we have improved Mohar's improvement (see [**Moh91**]) of the Alon–Milman result, taking $\lambda_\infty \leq 2d$ (the highest Laplacian eigenvalue)[7]). But the result of Chung, Faber, and Manteuffel (see [**CFM94**]) improves on our edge-based result by a factor of 2. In fact, it is precisely this factor of 2 that we gain in applying the result in Chung, Faber, and Manteuffel to analysis, in [**FT**], using the relation in this paper between graph Laplacians and analysis-like (e.g., edge-based) Laplacians.

5.2. Distances and diameters: new results. In this subsection we give a new way of proving distance/eigenvalues or diameter/eigenvalue results. Graph theoretically this yields a new result, which is an edge-based analogue of the Chung, Faber, and Manteuffel (see [**CFM94**]) result; our new results cannot be compared to that of Chung, Faber, and Manteuffel — it can yield better or worse results. This proof also carries over to analysis, where it gives a different (and in some sense shorter) proof of the Friedman–Tillich result in [**FT**]. But however different our proofs or results are, they can be

[6]It is interesting that this bound is often misquoted, as the authors use $[a]$ to denote $\lfloor a \rfloor$ (the greatest integer $\leq a$) without ever explicitly saying so in the paper. Many authors incorrectly guess the meaning of $[a]$.

[7]Since we are thinking of d/λ_T as being large, it seems reasonable to expect that λ_∞ should not be far from $2d$.

viewed as variants of older results stemming from the same basic technique used in [**CGY97**] and [**FT**].

Let X_0 denote the constants in $L^2(\mathcal{G},\mathcal{E})$, and X_1 the orthogonal complement of X_0; for $i=0,1$ let π_i denote the projection onto X_i (so $\pi_0+\pi_1$ is the identity). Let $W_{t,a}$ be the operators defined at the end of Section 4, and fix $a=\lambda_E$ to be the first nonzero eigenvalue of Δ.

Proposition 5.1. *If f,g are two functions whose supports are at a distance of at least d, we have*

$$\cosh\bigl(d\sqrt{\lambda_E}\,\bigr) \geq \frac{\|\pi_1 f\|\|\pi_1 g\|}{\|\pi_0 f\|\|\pi_0 g\|}.$$

Proof. By Proposition 4.13 we have

$$0=(W_{d,a}f,g)=(W_{d,a}\pi_0 f,\pi_0 g)+(W_{d,a}\pi_1 f,\pi_1 g).$$

Now

$$(W_{d,a}\pi_0 f,\pi_0 g)=\pm\cosh\bigl(d\sqrt{\lambda_E}\,\bigr)\|\pi_0 f\|\|\pi_0 g\|$$

since $\pi_0 f,\pi_0 g$ are both constants. Since $W_{d,a}$ restricted to X_1 has norm at most one, we have

$$\bigl|(W_{d,a}\pi_1 f,\pi_1 g)\bigr| \leq \|\pi_1 f\|\|\pi_1 g\|,$$

and the proposition follows. □

Corollary 5.2. *Let X,Y be disjoint measurable subsets of distance $\geq d$. Then*

$$d\sqrt{\lambda_E} \leq \cosh^{-1}\left(\sqrt{\frac{\mathcal{E}(X^{\mathrm{c}})\mathcal{E}(Y^{\mathrm{c}})}{\mathcal{E}(X)\mathcal{E}(Y)}}\right).$$

Proof. Take f,g to be the respective characteristic functions of X,Y. □

We finish this subsection with a discussion of the above proposition and its corollary.

First of all, these proofs carry right over to analysis, where they yield the results of Friedman and Tillich (in [**FT**]) with a considerably simpler proof; however, the basic idea that $(W_{d,a}f,g)=0$ based on the supports of f,g the the speed of propagation of W appears before.

Second, the above proposition and corollary can be generalized to k functions or k sets for any $k\geq 2$, with the same technique that appears in [**CGY96**], also used in [**CGY97**, **FT**]. In the analysis case we recover the results for k functions or sets that appear in [**FT**]. For graphs, our corollary would read that if $X_1,\dots,X_k$ were disjoint measurable subsets any two of which had supports of distance $\geq d$, then

$$d\sqrt{\lambda_E} \leq \min_{i\neq j}\cosh^{-1}\left(\sqrt{\frac{\mathcal{E}(X_i^{\mathrm{c}})\mathcal{E}(X_j^{\mathrm{c}})}{\mathcal{E}(X_i)\mathcal{E}(X_j)}}\right).$$

Our proposition would read similarly.

Finally, we remark that the above corollary yields a diameter result that can be better (or worse) than that of Chung, Faber, and Manteuffel (see [**CFM94**]). Similarly our corollary can be better or worse than the comparable theorem in [**CGY97**]. For example, the Chung, Faber, and Manteuffel result states that in a regular graph

$$(D-1)\cosh^{-1}\left(1+\frac{2\lambda_V}{\lambda_n-\lambda_V}\right)\le \cosh^{-1}(n-1),$$

where D is the diameter, $n=|V|$, and λ_V,λ_n are respectively the smallest and largest positive Δ_V eigenvalues. Taking balls of radius $\delta/2$ about two points of distance D and applying Corollary 5.2 we get

$$(D-\delta)\sqrt{\lambda_E}\le \cosh^{-1}(2n/\delta-1)$$

for any $\delta\le 2$. It follows that the result obtained here, taking $\delta=2$, is better than the result in [**CGY97**], provided that λ_V and $2-\lambda_n$ are both $\le c/\log n$ for a constant, c (using $\cos\sqrt{\lambda_E}=1-\lambda_V$).

5.3. Invariants. A typical spectral invariant studied in the analysis literature is the wave invariant

$$W(t)=\operatorname{Trace}\big(\cos t\sqrt{\Delta_E}\,\big)=\sum_j\big(\cos t\sqrt{\lambda_j}\,\big),$$

with λ_j running through all Laplacian eigenvalues. This sum can be understood in several ways; here we think of

$$\widetilde{W}(t)=\sum_j\big(e^{it\sqrt{\lambda_j}}\,\big)$$

as a complex analytic function defined on the subset of complex numbers with positive imaginary part, and then we extend $\widetilde{W}$ analytically to the whole complex plane; W is just the real part of $\widetilde{W}$.

It is well-known that in analysis the real singularities of W are at $t=0$ and t being (plus or minus) the length of a closed geodesic. If all edges of a graph have length one, then we know we have $\alpha_1,\dots,\alpha_{2|E|}$ such that the edge-based eigenvalues (with their respective multiplicities) are precisely those squares of $\alpha_j+2\pi\mathbb{Z}_{\ge 0}$. We have

$$\widetilde{W}(t)=\frac{1}{1-e^{2\pi it}}\sum_{j=1}^{2|E|}e^{it\alpha_j}.$$

This has a pole at $t\mathbb{R}$ precisely when $t=0$ or t is an integer with $\sum e^{it\alpha_j}\neq 0$.

To understand the vanishing or not of $\sum e^{it\alpha_j}$, consider that the α_j's are of two types; one type is a $(\pi,2\pi)$ pair coming from an edge and interior vertex count, and such α_j's cancel in the sum $\sum\cos(it\alpha_j)$ for t an odd

integer and contribute 2 when t is even; the other type is a $\alpha, 2\pi - \alpha$ pair with

$$e^{it\alpha} + e^{it(2\pi-\alpha)} = 2\cos(t\alpha)$$

for t an integer. Hence this sum is essentially the trace of the t-th Chebyshev polynomial in $\widetilde{A}$.

It follows that the first odd $t > 0$ for which $\widetilde{W}$ has a pole is the length of the smallest odd cycle. However it doesn't seem like such a simple statement holds for higher odd values of t or even values, and so the invariant $\widetilde{W}$ is not entirely analogous to its analysis counterpart.

It is natural to ask if any spectral, edge-based invariants (such as those whose analysis analogues are interesting, for example) yield new and interesting graph invariants. Such invariants would, in particular, include traces of Chebyshev polynomials of $\widetilde{A}$ when the edge lengths are one.

5.4. Cheeger's inequality. In this section we mention that Cheeger's inequality holds for the edge-based Laplacian as well as the vertex-based, but the second one, at least for r-regular graphs yields the first one. We will require some notions from [**FT99**].

Consider an open subset of $A \subset \mathcal{G}$ whose boundary contains no vertices, and let $\mathcal{A}(\partial A)$ be the "area" of A's boundary (see [**FT99**]); this is just the sum of a_e for each boundary point of A lying on e. Also $\mathcal{E}(A)$ is just the total $\mathcal{E}$ measure of A, and we set

$$h_E = \min_{\mathcal{E}(A) \le \mathcal{E}(\mathcal{G})/2} \frac{\mathcal{A}(\partial A)}{\mathcal{E}(A)}.$$

The co-area formula of [**FT99**] and the arguments to prove Cheeger's inequality immediately carry over here to yield:

$$\lambda_E \ge h_E^2/4. \tag{5.3}$$

We can compare this to Cheeger's inequality for the vertex-based case (i.e., Dodziuk's inequality, see [**Dod84**, **FT99**]),

$$\lambda_V \ge h_V^2/2 \tag{5.4}$$

for the 1-regular graphs (see Section 2), where

$$h_V = \min_{B \subset V,\, \mathcal{V}(B) \le\, \mathcal{V}(V)/2} \frac{\mathcal{E}(E(B, B^{\mathrm{c}}))}{\mathcal{V}(B)},$$

where $E(B, B^{\mathrm{c}})$ denotes the set of edges with one endpoint in B and one in B^{c}.

Let us compare these two Cheeger's inequalities, in case the graph is d-regular in the traditional sense (each vertex is the endpoint of d edges) with unit edge lengths and weights. We consider now the graph $\mathcal{G}$ derived from it, where $\mathcal{V}$ at any vertex is taken to be d so as to make the graph 1-regular.

We shall need a simple lemma:

Lemma 5.3. *For a* 1*-regular graph* $h_V \geq h_E/2$.

Proof. Let B be the subset of vertices of size smaller than $\mathcal{V}(V)/2$ for which

$$h_V = \frac{\mathcal{E}(E(B, B^{\mathrm{c}}))}{\mathcal{V}(B)}.$$

For a subset X of vertices we denote by X_t the set of points lying on edges that have both endpoints on B or are within a distance t from X. Note that

$$\mathcal{E}(B_{1/2}) + \mathcal{E}({B^{\mathrm{c}}}_{1/2}) = \mathcal{E}(\mathcal{G}).$$

It follows that the measure of one of these two sets is smaller than or equal to $\mathcal{V}(\mathcal{G})/2$.

Case 1: $\mathcal{E}(B_{1/2}) \leq \mathcal{V}(\mathcal{G})/2$. Here $\frac{\mathcal{A}(\partial B_{1/2})}{\mathcal{E}(B_{1/2})} \geq h_E$. Since $\mathcal{E}(E(B, B^{\mathrm{c}})) = \mathcal{A}(\partial B_{1/2})$ and $\mathcal{V}(B) \leq 2\mathcal{E}(B_{1/2})$, we obtain

$$h_V = \frac{\mathcal{E}(E(B, B^{\mathrm{c}}))}{\mathcal{V}(B)} \geq \frac{\mathcal{A}(\partial B_{1/2})}{2\mathcal{E}(B_{1/2})} \geq h_E/2.$$

Case 2: $\mathcal{V}({B^{\mathrm{c}}}_{1/2}) \leq \mathcal{E}(\mathcal{G})/2$. This implies that

$$\mathcal{V}(V)/4 = \mathcal{E}(\mathcal{G})/2 \geq \mathcal{V}({B^{\mathrm{c}}}_{1/2}) \geq \mathcal{V}(B^{\mathrm{c}})/2,$$

so B^{c} is also of measure smaller than or equal to $\mathcal{V}(V)/2$, and therefore $\mathcal{V}(B) = \mathcal{V}(B^{\mathrm{c}}) = \mathcal{V}(V)/2$. This means that

$$h_V = \frac{\mathcal{E}(E(B, B^{\mathrm{c}}))}{\mathcal{V}(B^{\mathrm{c}})}.$$

By using the same arguments as in the previous case (with B^{c} replacing B) we also get $h_V \geq h_E/2$. □

Now notice that

$$\lambda_E = (\cos^{-1}(1 - \lambda_V))^2 \geq 2\lambda_V.$$

Hence

$$\lambda_E \geq 2\lambda_V \geq h_V^2 \geq h_E^2/4.$$

In other words: for 1-regular graphs Cheeger's inequality for vertices, (5.4), implies the Cheeger inequality for edges, (5.3).

5.5. Optimal distance bounds. In this section we prove the following theorem:

Theorem 5.4. *Let* C, C_1, C_2 *be constants such that the following holds: for any graph with edge lengths one, whose diameter* D *is realized by two vertices,* u, v*, we have*

$$D - 1 \leq \frac{C}{\sqrt{2\lambda_V} + C_1\lambda_V} \log\left(\frac{C_2\mathcal{V}(V)^2}{\mathcal{V}(u)\mathcal{V}(v)}\right). \tag{5.5}$$

Then $C \geq 1/2$*. The same is true if we insist that the graph is* 1*-regular.*

The bound of Chung, Faber, and Manteuffel (in [**CFM94**]) implies Equation (5.5) with $C = 1/2$ assuming $\mathcal{V}$ is constant; the generalization to general $\mathcal{V}$ is implied by the results in [**CGY96**, **CGY97**] (with different constants C_2 in the two articles). Actually, there they assume the edge weights are integral; but this assumption clearly implies the same for rational edge weights, and therefore arbitrary real edge weights. So regarding λ_V as small, these inequalities are optimal to first-order.

Our proof is based on a standard metric probability space, the "exponential distribution on the nonnegative reals," or on a standard Riemannian manifold, the surface of revolution of $y = e^{-x}$ with x nonnegative. We model this on a graph by taking a sequence of edges with edge weight exponentially decreasing. These analysis examples (the exponential distribution and the surface of revolution of $y = e^{-x}$) prove a bound in analysis that is analogous to the bound $C \geq 1/2$ in the theorem above (see [**FT**]).

Proof. Fix a small $\eta \in (0,1)$, and let $r = 1 - \eta$. Fix an integer $D > 0$. Consider the graph, $\mathcal{G}$, whose vertices are the integers, $V = \{0, 1, \ldots, D\}$, with an edge from i to $i+1$ of weight r^i for all nonnegative integers $i < D$.

Making $\mathcal{G}$ into a 1-regular graph is done by taking $\mathcal{V}(i) = r^{i-1} + r^i$ for $i \neq 0, D$, dropping one of these summands when $i = 0$ or $i = D$. Here the vertices furthest from each other are 0 and D, and therefore

$$\log\left(\frac{\mathcal{V}(V)^2}{\mathcal{V}(u)\mathcal{V}(v)}\right) = \log\left(\frac{4\left(\frac{1-r^{D+1}}{1-r}\right)^2}{1 \cdot r^D}\right) \leq K(r) - D\log r.$$

We obtain a lower bound on λ_V by using Cheeger's inequality (5.4). Obviously to find Cheeger's constant h_V we just need to consider vertex sets that are connected, i.e., of the form $B = \{a, a+1, \ldots, b\}$.

We have to distinguish between two cases: $a > 0$ and $a = 0$. In the first case

$$\frac{\mathcal{E}(E(B, B^{\mathrm{c}}))}{\mathcal{V}(B)} = \frac{r^{a-1} + \mathbf{1}_{\{b \neq D\}} r^b}{r^{a-1} + 2\sum_{i=a}^{b-1} r^i + \mathbf{1}_{\{b \neq D\}} r^b} \geq \frac{1}{2(1 + r + r^2 + \cdots)} = \frac{1-r}{2}.$$

In the second case we get the same lower bound with a slightly more tedious calculation. First we note that in this case $b \neq D$ (since $\mathcal{V}(B) \leq \mathcal{V}(V)/2$) and therefore

$$\mathcal{V}(B) = r^b + 2\sum_{i=0}^{b-1} r^i \leq \mathcal{V}(V)/2 = \frac{1 - r^{D+1}}{1-r}. \tag{5.6}$$

This implies that

$$r^b \geq \frac{1 + r^{D+1}}{1+r}. \tag{5.7}$$

We use the previous upper bound on $\mathcal{V}(B)$, and this lower bound on r^b to obtain

$$\frac{\mathcal{E}(E(B,B^c))}{\mathcal{V}(B)} = \frac{r^b}{\mathcal{V}(B)} \geq \frac{(1+r^{D+1})(1-r)}{(1+r)(1-r^{D+1})} \geq \frac{1-r}{2}.$$

By using (5.4) it follows that

$$\lambda_V \geq (1-r)^2/8 = \eta^2/8,$$

and therefore

$$D-1 \leq \frac{C}{\eta/2 + C_1\eta^2/8}\left(K(r) - D\log r\right).$$

Taking $D \to \infty$ we conclude

$$1 \leq -\frac{C}{\eta/2 + C_1\eta^2/8}\ \log(1-\eta).$$

Taking $\eta \to 0^+$ yields $1 \leq 2C$. □

References

[AM85] N. Alon and V.D. Milman, λ_1, *isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B, **38**(**1**) (1985), 73–88, MR 0782626 (87b:05092), Zbl 0549.05051.

[BL97] S. Bobkov and M. Ledoux, *Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution*, Probab. Theory Related Fields, **107**(**3**) (1997), 383–400, MR 1440138 (98e:60026), Zbl 0878.60014.

[CFM94] F.R.K. Chung, V. Faber and T.A. Manteuffel, *An upper bound on the diameter of a graph from eigenvalues associated with its Laplacian*, SIAM J. Discrete Math., **7**(**3**) (1994), 443–457, MR 1285582 (95j:05115), Zbl 0808.05072.

[CGY96] F.R.K. Chung, A. Grigor'yan, and S.-T. Yau, *Upper bounds for eigenvalues of the discrete and continuous Laplace operators*, Adv. Math., **117**(**2**) (1996), 165–178, MR 1371647 (96m:58254), Zbl 0844.53029.

[CGY97] F.R.K. Chung, A. Grigor'yan, and S.-T. Yau. *Eigenvalues and diameters for manifolds and graphs*, in 'Tsing Hua lectures on geometry & analysis' (Hsinchu, 1990–1991), 79–105, Internat. Press, Cambridge, MA, 1997, MR 1482032 (98h:58206), Zbl 0890.58093.

[CH89] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Partial differential equations, Wiley, New York, 1962, MR 1013360 (90k:35001), Zbl 0729.35001.

[Dod84] J. Dodziuk, *Difference equations, isoperimetric inequality and transience of certain random walks*, Trans. Amer. Math. Soc., **284** (1984), 787–794, MR 0743744 (85m:58185), Zbl 0512.39001.

[Fri69] A. Friedman, *Partial Differential Equations*. Holt, Rinehart and Winston, 1969, MR 0445088 (56 #3433), Zbl 0224.35002.

[Fri93] J. Friedman, *Some geometric aspects of graphs and their eigenfunctions*, Duke Math. J., **69**(**3**) (1993), 487–525, MR 1208809 (94b:05134), Zbl 0785.05066.

[FT] J. Friedman and J.-P. Tillich, *Laplacian eigenvalues and distances between subsets of a manifold*, J. Differential Geom., **56**(**2**) (2000), 285–299, MR 1863018 (2002g:58050), Zbl 1032.58015.

[FT99] J. Friedman and J.-P. Tillich. *Calculus on graphs*, preprint, October 1999, arXiv: cs.DM/0408028.

[GT83] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren der Math. Wissenschaften, **224**, Springer-Verlag, 2nd edition, 1983, MR 0737190 (86c:35035), Zbl 0562.35001.

[Hur00] N.E. Hurt, *Mathematical Physics of Quantum Wires and Devices: From Spectral Resonances to Anderson Localization*, Math. and its Appl., **506**, Kluwer, Dordrecht, 2000, MR 1862155 (2002h:81004).

[KZ01] P. Kuchment and H. Zeng, *Convergence of spectra of mesoscopic systems collapsing onto a graph*, J. Math. Anal. Appl., **258**(**2**) (2001), 671–700, MR 1835566 (2002d:35158), Zbl 0982.35076.

[LPS88] A. Lubotzky, R. Phillips and P. Sarnak, *Ramanujan graphs*, Combinatorica, **8**(**3**) (1988), 261–277, MR 0963118 (89m:05099), Zbl 0661.05035.

[Moh91] B. Mohar, *Eigenvalues, diameter, and mean distance in graphs*, Graphs Combin., **7**(**1**) (1991), 53–64, MR 1105467 (92d:05105), Zbl 0771.05063.

[RS01] J. Rubinstein and M. Schatzman, *Variational problems on multiply connected thin strips.* I. *Basic estimates and convergence of the Laplacian spectrum*, Arch. Ration. Mech. Anal., **160**(**4**) (2001), 271–308, MR 1869667 (2003b:35007), Zbl 0997.49003.

[Smo83] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Grundlehren der Mathematischen Wissenschaften, **258**, Springer-Verlag, New York, 1983, MR 0688146 (84d:35002), Zbl 0508.35002.

[Wey12] H. Weyl, *Der asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichunger*, Math. Ann., **71** (1912), 441–469, JFM 43.0436.01.

Received May 24, 2002. The first author's research was supported in part by an NSERC grant. Research of the second author was supported in part by the INRIA and the CCETT.

Departments of Computer Science and Mathematics
University of British Columbia
Vancouver BC V6T 1Z4
Canada
E-mail address: jf@cs.ubc.ca

Inria Projet Codes
Domaine de Voluceau BP.105 78153
France
E-mail address: jean-pierre.tillich@inria.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

GEOMETRY OF THE UNIT BALL AND REPRESENTATION THEORY FOR OPERATOR ALGEBRAS

Elias G. Katsoulis

We investigate the relationship between the facial structure of the unit ball of an operator algebra $\mathcal{A}$ and its algebraic structure, including the hereditary subalgebras and the socle of $\mathcal{A}$. Many questions about the facial structure of $\mathcal{A}$ are studied with the aid of representation theory. For that purpose we establish the existence of reduced atomic type representations for certain nonselfadjoint operator algebras. Our results are applicable to C*-algebras, strongly maximal TAF algebras, free semigroup algebras and various semicrossed products.

The study of geometric problems in operator algebra theory goes back to the beginnings of the subject. The theory of Gelfand–Naimark and Segal identified the extreme points of the state space of a C*-algebra as functionals (pure states) that produce irreducible representations under the GNS machinery. Kadison's characterization of the isometric linear maps between C*-algebras [**29**] depended heavily on the identification of the extreme points for the unit ball. Crucial information about the algebraic structure of a C*-algebra is encoded in the geometry of its unit ball. The ideal structure of the algebra coincides with the M-structure [**4**] and the density of the invertibles is reflected in the richness of the convex hull of the unitary operators [**45**].

A subset $\mathcal{F}$ of a convex set $\mathcal{K}$ is said to be a *face* of $\mathcal{K}$ if it is convex and has the property that, if an interior point of a line segment in $\mathcal{K}$ belongs to $\mathcal{F}$, the entire line segment belongs to $\mathcal{F}$. The extreme points of a convex set, together with the empty set, form the *trivial faces* of $\mathcal{K}$. A face $\mathcal{F}$ is said to be *finite-dimensional* if the (real) linear space generated by $\mathcal{F}$ is finite-dimensional. If $\mathcal{K}$ is contained in a normed linear space, then $\mathcal{F}$ is said to be *compact* if its norm closure is a norm compact set. A comprehensive study of the facial structure for the unit ball of a C*-algebra was conducted by Akemann and Pedersen [**2**], following related work by Edwards and Ruttimann [**22**, **23**]. Beyond selfadjoint operator algebras, there has not been a systematic work addressing the nontrivial faces of the unit ball.

In this paper we begin a study of the nontrivial compact faces of the unit ball of an arbitrary operator algebra. (All operator algebras are assumed to be norm closed and contain the identity operator.) The existence of such

faces has a significant impact on the structure of the algebra. In Theorem 1.6 we show that if the unit ball of an operator algebra $\mathcal{A}$ has a nontrivial compact face $\mathcal{F}$, then $\mathrm{S}(\mathcal{F})$, and therefore $\mathcal{A}$, contains a nonscalar operator A whose spectrum has at most one limit point. (Here $\mathrm{S}(\mathcal{F})$ denotes the unique real subspace of $\mathcal{A}$ that is a translate of the affine hull of $\mathcal{F}$.) For a finite-dimensional face $\mathcal{F}$ we can offer a more definitive result. Theorem 1.8 shows that $\mathrm{S}(\mathcal{F})$ is a (real) finite-dimensional hereditary subalgebra of $\mathcal{A}$ consisting of multiples of elements with finite geometric rank. Moreover, if $\mathcal{A}$ happens to be semisimple then $\mathrm{S}(\mathcal{F})$ is contained in the socle of $\mathcal{A}$; this makes an important connection between the facial structure of the unit ball and the general theory of Banach Algebras. As a consequence, if the unit ball of a semisimple operator algebra $\mathcal{A}$ contains a nontrivial finite-dimensional face then $\mathcal{A}$ contains a minimal idempotent (Corollary 1.10). We also relate the existence of nontrivial compact faces with the concept of geometric compactness, which was first introduced by Anoussis and the author in [**5**, **6**]. Theorem 1.3 shows that the existence of nontrivial compact faces implies the existence of geometrically compact elements. Actually, we observe that just the presence of nonzero geometrically compact elements suffices for the existence of nonscalars with discrete spectrum.

The general results of the first section are complemented with several applications. In the second section of the paper we investigate the facial structure of various operator algebras with the aid of representation theory. Motivated by our earlier work in [**5**], we introduce the class of operator semisimple algebras; these are algebras that can be isometrically represented as a strongly dense subalgebra of the diagonal algebra $\bigoplus_{a\in\mathbb{A}} \mathcal{B}(\mathcal{H}_a)$. We prove that for an operator semisimple algebra $\mathcal{A}$ the existence of nontrivial compact faces, the existence of nonzero geometrically compact elements and the existence of atoms are all equivalent conditions. Moreover, we relate the concept of geometric compactness to the representation theory of $\mathcal{A}$. In Theorem 2.4 we show that an element $A \in \mathcal{A}_1$ is geometrically compact if and only if there exists an isometric representation φ of $\mathcal{A}$ such that $\varphi(A)$ is a compact operator. This generalizes our earlier selfadjoint work [**5**] to the nonselfadjoint setting.

The rest of the second section is occupied with identifying various classes of operator semisimple algebras. Clearly, any operator algebra containing the compacts acts irreducibly on the Hilbert space and hence is operator semisimple. By an old result of Gardner, all C*-algebras are also operator semisimple. Therefore, the unit ball of a unital C*-algebra $\mathcal{A}$ has nontrivial compact faces if and only if it has finite-dimensional faces if and only if $\mathcal{A}$ has an atom. This result was implicit in [**5**].

It turns out that the concept of operator semisimplicity is also applicable to TAF algebras, a class of nonselfadjoint algebras that has received a great deal of attention in recent years; cf. Power's monograph [**44**]. Here

we use the representation theory of Davidson and the author [**13**]. In [**13**] we characterized the operator primitive TAF algebras as the semisimple ones whose enveloping C*-algebra is primitive. Here we add to this result and in Theorem 2.6 we show that all semisimple TAF algebras are operator semisimple. As an immediate corollary of our theory, the unit balls of the familiar standard, alternation and $A(\mathbb{Q}, \nu)$ algebras do not contain any nontrivial compact faces.

The list of operator semisimple algebras also includes various function algebras. Indeed, in Theorem 2.9 we show that if the unitary functions in a uniform algebra $\mathcal{A}$ separate points, $\mathcal{A}$ is operator semisimple. In particular, H^∞ and the disc algebra $A(\mathbb{D})$ are operator semisimple. The methods of the second section are also applicable to semicrossed products of the form $C(\mathbb{T}) \times_\alpha \mathbb{Z}^+$, where α is an irrational rotation of the circle $\mathbb{T}$. Indeed, in [**14**] it is shown that such algebras are operator primitive. Therefore, the unit ball of such an algebra does not contain any nontrivial compact faces.

In the third section, we study the presence of compact faces in the unit ball of a free semigroup algebra. In [**15**, Theorem 4.5] it is shown that every operator in the open unit ball of a free semigroup algebra $\mathcal{A}$ is a mean of isometries from $\mathcal{A}$. This generalizes a classical result of Marshall and shows that there is an abundance of extreme points in the unit ball of these algebras. Theorem 3.1 shows now that, once again, the unit ball of a free semigroup algebra $\mathcal{A}$ contains nontrivial compact faces if and only if $\mathcal{A}$ has atoms. In particular, the unit ball of the "noncommutative Toeplitz algebra" $\mathcal{L}_n$ has no nontrivial compact faces, a result that seems to be new even for H^∞. We note that the representation techniques of the second section are not applicable here since a free semigroup algebra may not be semisimple. Instead we use the spectral properties of $\mathcal{L}_n$ together with the structure theorem of Davidson, Katsoulis and Pitts [**15**]. Similar spectral considerations also show that the unit ball of $A(\mathbb{D}) \times_\alpha \mathbb{Z}^+$ and $H^\infty \times_\alpha \mathbb{Z}^+$ do not contain any nontrivial finite-dimensional faces. These algebras were studied in [**27**, **28**].

The last section of the paper contains several remarks and observations, including a generalization of Kadison's characterization for the extreme points of the unit ball of a C*-algebra.

1. Structure for the faces of the unit ball

If $\mathcal{K} \subseteq \mathcal{X}$ is a convex subset of a complex normed space $\mathcal{X}$, then $[\mathcal{K}]_\mathbb{R}$ denotes the real subspace of $\mathcal{X}$ generated by $\mathcal{K}$:

$$[\mathcal{K}]_\mathbb{R} \equiv \left\{ \sum_{i=1}^{n} \lambda_i x_i \;\middle|\; \lambda_i \in \mathbb{R},\, x_i \in \mathcal{K},\, 1 \leq i \leq n,\, n \in \mathbb{N} \right\}.$$

(The complex subspace generated by $\mathcal{K}$ will be denoted as $[\mathcal{K}]$.) For any $x \in \mathcal{K}$, the subspace $[x - \mathcal{K}]_{\mathbb{R}}$ does not depend on the choice of $x \in \mathcal{K}$ and is denoted as $\mathrm{S}(\mathcal{K})$. The translation of $\mathrm{S}(\mathcal{K})$ by any element of $\mathcal{K}$ equals the affine hull of $\mathcal{K}$.

If x and y belong to $\mathcal{K}$, the line segment joining x and y is denoted by $[x, y]$. Thus

$$[x, y] = \big\{\lambda x + (1 - \lambda)y \mid \lambda \in [0, 1]\big\}.$$

If x and y are in $\mathcal{K}$ and $x \neq y$, then an element $v \in [x, y]$ is said to be an internal point of $[x, y]$ if $v \neq x$ and $v \neq y$. Given $v \in \mathcal{K}$, we write $\mathcal{F}(\mathcal{K}, v)$ for the union of all line segments in $\mathcal{K}$ that contain v as an internal point, provided that v is not an extreme point of $\mathcal{K}$. Otherwise, $\mathcal{F}(\mathcal{K}, v) = \{v\}$. If $\mathcal{K}$ is the unit ball of $\mathcal{X}$, then $\mathcal{F}(\mathcal{K}, v)$ is simply denoted as $\mathcal{F}(v)$. It is an important fact in elementary convexity theory that for each $v \in \mathcal{K}$, the set $\mathcal{F}(\mathcal{K}, v)$ is a face of $\mathcal{K}$ that is minimal with the property of containing v. (The proof of this is an entertaining exercise in plane geometry; or see [**1**, Theorem 1.2] for a detailed proof.)

If $\mathcal{S}$ is a nonempty subset of the unit ball of $\mathcal{X}$, then the *contractive perturbations* of $\mathcal{S}$ are defined as

$$\mathrm{cp}(\mathcal{S}) = \big\{x \in \mathcal{X} \;\big|\; \|x \pm s\| \leq 1 \text{ for all } s \in \mathcal{S}\big\}.$$

It is clear that $\mathcal{S}_1 \subseteq \mathcal{S}_2$ implies $\mathrm{cp}(\mathcal{S}_1) \supseteq \mathrm{cp}(\mathcal{S}_2)$. Also, an element of the unit ball of $\mathcal{X}$ is an extreme point if and only if $\mathrm{cp}(\{x\}) = \{0\}$.

The following result relates contractive perturbations with the facial structure of the unit ball:

Lemma 1.1. *Let $\mathcal{F}$ be a face of the unit ball of a normed space $\mathcal{X}$. If $x \in \mathcal{F}$, then,*

$$\mathrm{cp}(\{x\}) \subseteq \tfrac{1}{2}\left(\mathcal{F} - \mathcal{F}\right).$$

Proof. Let $m \in \mathrm{cp}(\{x\})$. Then $x = \frac{1}{2}\big((x + m) + (x - m)\big)$, and therefore $x \pm m \in \mathcal{F}$. Hence $m = \frac{1}{2}\big(x + m - (x - m)\big)$ belongs to $\frac{1}{2}(\mathcal{F} - \mathcal{F})$ and this proves the lemma. □

We now compute $\mathrm{S}\left(\mathcal{F}(x)\right)$ in terms of contractive perturbations for x.

Lemma 1.2. *Let $\mathcal{X}$ be a normed space and let $x \in \mathcal{X}_1$. Then*

$$[x - \mathcal{F}(x)]_{\mathbb{R}} = [\mathrm{cp}(\{x\})]_{\mathbb{R}}.$$

Proof. Assume $x + m \in \mathcal{F}(x)$. There exists $\lambda > 0$ such that $x \pm \lambda m \in \mathcal{F}(x)$, so $\lambda m \in \mathrm{cp}(\{x\})$. Conversely, if $m \in \mathrm{cp}(\{x\})$, then $[x - m, x + m] \subseteq \mathcal{X}_1$ and so by the definition of $\mathcal{F}(x)$ we have that $x \pm m \in \mathcal{F}(x)$. □

One may define contractive perturbations of higher-order by using the recursive formula $\mathrm{cp}^{(n+1)}(\mathcal{S}) = \mathrm{cp}\left(\mathrm{cp}^{(n)}(\mathcal{S})\right)$, $n \in \mathbb{N}$. These higher-order contractive perturbations satisfy the Galois duality $\mathrm{cp}^{(n+2)}(\mathcal{S}) = \mathrm{cp}^{(n)}(\mathcal{S})$,

$n \in \mathbb{N}$. The second contractive perturbations were introduced by Anoussis and the author in [**5**, **6**]. In [**5**] we defined a contraction x in a normed space space $\mathcal{X}$ to be *geometrically compact* if $\mathrm{cp}^{(2)}(\{x\})$ is norm compact. If $\mathrm{cp}^{(2)}(\{x\})$ happens to span a finite-dimensional subspace of $\mathcal{X}$, then x is said to have finite geometric rank. In [**5**, Theorem 2.2] we proved that a nonzero element A of a C*-algebra $\mathcal{A}$ is geometrically compact (resp. has finite geometric rank) if and only if there exists a faithful representation φ of $\mathcal{A}$ such that $\varphi(A)$ is a compact operator (resp. $\varphi(A)$ is a finite rank operator).

Theorem 1.3. *Let $\mathcal{X}$ be a normed space and assume that the unit ball of $\mathcal{X}$ has a nontrivial compact face $\mathcal{F}$. Then $\mathrm{S}(\mathcal{F})$ contains a nonzero geometrically compact element.*

Proof. Let x_1, x_2 be distinct elements of $\mathcal{F}$ and let $x = \frac{1}{2}(x_1 + x_2)$. Since x is not an extreme point, $\mathrm{cp}(\{x\})$ contains a nonzero element, say m. Then $\{m\} \subseteq \mathrm{cp}(\{x\})$ and so

$$\mathrm{cp}^{(2)}(\{m\}) \subseteq \mathrm{cp}^{(3)}(\{x\}) = \mathrm{cp}(\{x\}).$$

By Lemma 1.1, $\mathrm{cp}^{(2)}(m)$ is contained in $\frac{1}{2}(\mathcal{F} - \mathcal{F})$, which is a norm compact set; thus m is geometrically compact. Hence $\mathrm{cp}(\{x\})$ contains nonzero geometrically compact elements and by Lemma 1.2 the same is true for $[x - \mathcal{F}(x)]_{\mathbb{R}} \subseteq [x - \mathcal{F}]_{\mathbb{R}}$. □

It is instructive to observe that some geometrically compact elements may be located outside translates of affine hulls for compact faces. Actually, there exists a Banach space $\mathcal{X}$ containing elements with finite geometric rank but its unit ball has no compact faces. Indeed, by [**5**, Theorem 2.2], c_0 contains an abundance of elements with finite geometric rank. However, given any element x in the unit ball of c_0, it is easy to see that $\mathcal{F}(x)$ does not have a compact closure and so the unit ball of c_0 has no compact faces.

The following mild generalization of [**5**, Proposition 1.2] is necessary for deriving Theorem 1.8. Compare with [**34**, Theorem 3] and [**46**, Theorem 2], where the calculations below originate.

Proposition 1.4. *Let $\mathcal{A}$ be an operator algebra, let $\emptyset \neq \mathcal{S} \subseteq \mathcal{A}_1$ and assume that $S_1, S_2 \in \mathcal{S}$. If $X \in \mathcal{A}$ satisfies $\|X\| \leq 1/2$, then $S_1 X S_2 \in \mathrm{cp}^{(2)}(\mathcal{S})$.*

Proof. Let $B \in \mathrm{cp}(\mathcal{S})$. Since $\|S_i \pm B\| \leq 1$ we have

$$S_i^* S_i + B^* B - S_i^* B - B^* S_i \leq I,$$
$$S_i^* S_i + B^* B + S_i^* B + B^* S_i \leq I,$$

so $S_i^* S_i \leq I - B^* B$. Douglas' majorization theorem implies the existence of a contraction Q_i such that $S_i = Q_i (I - B^* B)^{1/2}$. A similar argument shows

that $S_i = (I - BB^*)^{1/2} P_i$ for some contraction P_i. Hence,

$$S_1 X S_2 = (I - BB^*)^{1/2} P_1 X Q_2 (I - B^* B)^{1/2}, \tag{1}$$
$$= (I - |B^*|)^{1/2} Y (I - |B|)^{1/2},$$

where $Y = (I + |B^*|)^{1/2} P_1 X Q_2 (I + |B|)^{1/2}$; thus $\|Y\| \leq 1$. The Heinz–Kato inequality [**32**] now asserts that for any vectors $e, f \in \mathcal{H}$ we have

$$\big|\langle Be, f\rangle\big| \leq \big\||B|^{1/2} e\big\| \, \big\||B^*|^{1/2} f\big\|. \tag{2}$$

Combining (1) and (2), we obtain $\|B \pm S_1 X S_2\| \leq 1$ and so $S_1 X S_2 \in \mathrm{cp}^{(2)}(\mathcal{S})$, as desired. □

Corollary 1.5. *Let $\mathcal{A}$ be an operator algebra, $A \in \mathcal{A}_1$ and assume that $C_1, C_2 \in \mathrm{cp}(\{A\})$. If $X \in \mathcal{A}$ satisfies $\|X\| \leq 1/2$, then $C_1 X C_2 \in \mathrm{cp}(\{A\})$.*

Proof. Apply Proposition 1.4 with $\mathcal{S} = \mathrm{cp}(\{A\})$. Then

$$C_1 X C_2 \in \mathrm{cp}^{(3)}(\{A\}) = \mathrm{cp}(\{A\}).$$ □

If a is an element of a Banach algebra $\mathfrak{A}$, let $\sigma_{\mathfrak{A}}(a)$ denote the spectrum of a as an element of $\mathfrak{A}$. The *left multiplier* L_a is defined as $L_a b = ab$, $b \in \mathfrak{A}$. The collection $\mathrm{M}_l(\mathfrak{A})$ of all left multipliers on $\mathfrak{A}$ is isometrically isomorphic as an algebra to $\mathfrak{A}$. Therefore, the map $a \longrightarrow L_a$ is spectrum preserving, i.e., $\sigma_{\mathfrak{A}}(a) = \sigma_{\mathrm{M}_l(\mathfrak{A})}(L_a)$.

Theorem 1.6. *Let $\mathcal{A}$ be an operator algebra, let $\mathcal{F}$ be a nontrivial face of its unit ball and let $\mathrm{S}(\mathcal{F})$ be the unique real subspace of $\mathcal{A}$ that is a translate of the affine hull of $\mathcal{F}$. Then the following three conditions are successively weaker:*

(i) *$\mathcal{F}$ is a compact face.*
(ii) *$\mathrm{S}(\mathcal{F})$ contains a nonzero geometrically compact element.*
(iii) *$\mathrm{S}(\mathcal{F})$ contains a nonscalar operator A whose spectrum has at most one limit point.*

If $\mathcal{A}$ is commutative and semisimple, then condition (ii) *also implies:*

(iii′) *$\mathcal{A}$ contains a minimal idempotent, i.e., a nonzero idempotent Q such that $Q\mathcal{A}Q = \mathbb{C}Q$.*

Proof. (i) ⇒ (ii): Theorem 1.3 shows that this is valid for any normed space.

(ii) ⇒ (iii): Assume that $\mathrm{S}(\mathcal{F})$ contains such an element A, so the norm closure of $\mathrm{cp}^{(2)}(\{A\})$ is a nonzero compact set. Proposition 1.4 shows that

$$A\mathcal{A}_{1/2} A \subseteq \mathrm{cp}^{(2)}(\{A\}),$$

so the norm closure of $A\mathcal{A}_1 A$ is norm compact.

Consider the elementary operator L_{A^2} acting on the operator algebra $\mathfrak{A}$ generated by the polynomials of A. Since the closure of $A\mathcal{A}_1 A$ is norm compact, the norm closure of $A^2\mathfrak{A}_1$ is also compact, so the operator L_{A^2}

is a compact operator on $\mathfrak{A}$. According to the Riesz theory for compact operators, $\sigma_{\mathcal{B}(\mathfrak{A})}(L_{A^2})$ is a countable set with 0 as its only limit point (see Theorem VII.7.1 in [**11**]). Clearly the same is true for $\sigma_{\mathcal{B}(\mathfrak{A})}(L_A)$. Since $\mathrm{M}_l(\mathfrak{A}) \subseteq \mathcal{B}(\mathfrak{A})$, Theorem VII.5.4 in [**11**] shows that $\sigma_{\mathcal{B}(\mathfrak{A})}(L_A)$ and $\sigma_{\mathrm{M}_l(\mathfrak{A})}(L_A)$ differ only by holes, and so are equal. Hence, $\sigma_{\mathrm{M}_l(\mathfrak{A})}(L_A)$ is countable with one limit point. Our remarks above show that the same is true for $\sigma_{\mathfrak{A}}(A)$. Another application of [**11**, Theorem VII.5.4] for the Banach algebras $\mathfrak{A} \subseteq \mathcal{B}(\mathcal{H})$ shows that $\sigma_{\mathcal{B}(\mathcal{H})}(A)$ has at most one limit point, as desired.

(ii) $\Rightarrow$ (iii′): Arguing as above, let $A \in \mathcal{A}$, so that the closure of $A\mathcal{A}_1 A$ is norm compact and so, by commutativity, the norm closure of $A^2\mathcal{A}_1$ is a compact set. Therefore, the left multiplier L_{A^2} is a compact operator on $\mathcal{A}$. Since $\mathcal{A}$ is semisimple, the left multiplier algebra $\mathrm{M}_l(\mathcal{A})$ is also semisimple. Hence L_{A^2} is a nonquasinilpotent compact operator. Let λ be a nonzero eigenvalue of L_A and let $E(\lambda)$ be the corresponding Riesz idempotent (see VII.6.9 in [**11**]). Since $L_A \in \mathrm{M}_l(\mathcal{A})$, we have $E(\lambda) \in \mathrm{M}_l(\mathcal{A})$, so there exists an idempotent $Q \in \mathcal{A}$ such that $E(\lambda) = L_Q$. By [**11**, Corollary VII.7.8] the idempotent $E(\lambda) = L_Q$ has finite-dimensional range, i.e., $Q\mathcal{A}$ is finite-dimensional. By [**38**, Proposition 4.3.12], $Q\mathcal{A}$ is semisimple, so by the Wedderburn-Artin Theorem, $Q\mathcal{A}$ is isomorphic to a direct sum of full matrix algebras. The existence of the minimal idempotent in $\mathcal{A}$ now follows. □

Corollary 1.7. *The unit ball of* H^∞ *has no compact faces apart from singletons.*

A (real or complex) subalgebra $\mathcal{B}$ of an operator algebra $\mathcal{A}$ is said to be *hereditary* if $B_1, B_2 \in \mathcal{B}$ implies $B_1\mathcal{A}B_2 \subseteq \mathcal{B}$. For complex selfadjoint subalgebras of C*-algebras this definition coincides with the familiar definition of a hereditary subalgebra, as it appears in [**36**, **40**]. It is easy to see that if $\mathcal{B}$ is a real hereditary subalgebra of $\mathcal{A}$, then $[B]$ is a complex hereditary subalgebra of $\mathcal{A}$.

Theorem 1.8. *Let $\mathcal{A}$ be an operator algebra, let $\mathcal{F}$ be a finite-dimensional face of the unit ball of $\mathcal{A}$ and let $\mathrm{S}(\mathcal{F})$ be the unique real subspace of $\mathcal{A}$ that is a translate of the affine hull of $\mathcal{F}$. Then $\mathrm{S}(\mathcal{F})$ is a finite-dimensional hereditary subalgebra of $\mathcal{A}$ consisting of multiples of elements with finite geometric rank. Moreover,*

$$\mathcal{F} = (A + \mathrm{S}(\mathcal{F})) \cap \mathcal{A}_1 \tag{3}$$

for any $A \in \mathcal{F}$.

Proof. If $A, B \in \mathcal{F}$, then $[A - \mathcal{F}]_{\mathbb{R}} = [B - \mathcal{F}]_{\mathbb{R}}$ and so

$$A + [A - \mathcal{F}]_{\mathbb{R}} = B + (A - B) + [B - \mathcal{F}]_{\mathbb{R}} = B + [B - \mathcal{F}]_{\mathbb{R}}.$$

Therefore it suffices to prove (3) for a specific $A \in \mathcal{F}$. Since $\mathcal{F}$ is a finite-dimensional convex set, it has nonempty relative interior. Therefore, there exists $A \in \mathcal{F}$ and $\epsilon > 0$ such that $B \in \mathcal{F}$ implies

$$[A - \epsilon(B - A), B] \subseteq \mathcal{F}. \tag{4}$$

We claim that $\mathcal{F} = \mathcal{F}(A)$. Indeed, $\mathcal{F}(A) \subseteq \mathcal{F}$. Conversely, let $B \in \mathcal{F}$. The definition of $\mathcal{F}(A)$ and (4) imply that

$$[A - \epsilon(B - A), B] \subseteq \mathcal{F}(A),$$

and so $B \in \mathcal{F}(A)$, which proves the claim.

Since $\mathcal{F} = \mathcal{F}(A)$, Lemma 1.2 shows that

$$\mathrm{S}(\mathcal{F}) = [A - \mathcal{F}]_{\mathbb{R}} = [\mathrm{cp}(A)]_{\mathbb{R}}.$$

By Corollary 1.5, $\mathrm{S}(\mathcal{F})$ is a finite-dimensional hereditary subalgebra of $\mathcal{A}$.

Since $\mathrm{cp}(\{A\})$ is convex and $\mathrm{cp}(\{A\}) = -\mathrm{cp}(\{A\})$, a moment's reflection shows that $[\mathrm{cp}(\{A\})]_{\mathbb{R}}$ consists of multiples of $\mathrm{cp}(\{A\})$. Hence $\mathrm{S}(\mathcal{F})$ consists of multiples of elements in $\mathrm{cp}(\{A\})$. However, if $X \in \mathrm{cp}(\{A\})$, then

$$\mathrm{cp}^{(2)}(\{X\}) \subseteq \mathrm{cp}^{(3)}(\{A\}) = \mathrm{cp}(\{A\}).$$

Therefore X has finite geometric rank and so $\mathrm{S}(\mathcal{F})$ consists of multiples of elements with finite geometric rank.

It remains to verify (3). Let $B \in \mathrm{S}(\mathcal{F})$ be such that $\|A + B\| = 1$. Then there exists $\lambda > 0$ such that $\lambda B \in \mathrm{cp}(A)$, and so $\|A - \lambda B\| \leq 1$. Hence $[A - \lambda B, A + B] \subseteq \mathcal{A}_1$. Since A is contained in the interior of the line segment, we conclude that $A + B \in \mathcal{F}$, as desired. □

Recall that the *socle* of a semisimple Banach algebra $\mathcal{A}$ is defined as the sum of all minimal left ideals of $\mathcal{A}$. It coincides with the sum of all minimal right ideals [**38**, Proposition 8.2.8] and therefore it is a (not necessarily closed) two-sided ideal of $\mathcal{A}$. The study of the socle has been a central theme in the theory of Banach algebras. Our next result shows that the socle is also important for the geometry of the unit ball.

Corollary 1.9. *Let $\mathcal{A}$ be an operator algebra and let $\mathcal{F}$ be a finite-dimensional face of the unit ball of $\mathcal{A}$. If $\mathcal{A}$ is semisimple then $\mathrm{S}(\mathcal{F})$ is contained in the socle of $\mathcal{A}$.*

Proof. Let $A \in \mathrm{S}(\mathcal{F})$. By Theorem 1.8, $\mathrm{S}(\mathcal{F})$ is hereditary, so the operator $X \longrightarrow AXA$, $X \in \mathcal{A}$, has finite-dimensional range. By [**3**, Theorem 7.2], A belongs to the socle of $\mathcal{A}$. □

Corollary 1.10. *Let $\mathcal{A}$ be a semisimple operator algebra. If the unit ball of $\mathcal{A}$ has a nontrivial finite-dimensional face then $\mathcal{A}$ contains a minimal idempotent.*

Proof. The socle of a semisimple Banach algebra $\mathcal{A}$ is generated by the set of minimal idempotents of $\mathcal{A}$ [**38**, Proposition 8.2.8]. □

In particular, the unit ball of a simple operator algebra has no finite-dimensional faces apart from singletons.

2. Representation theorems for operator algebras

Recall that Theorem 1.8 asserts that if $\mathcal{F}$ is a finite-dimensional face of the unit ball of an operator algebra then $\mathrm{S}(\mathcal{F})$ consists of scalar multiples of elements with finite geometric rank. So far the elements with finite geometric rank have been characterized for two classes of operator algebras: nest algebras [**6**] and C*-algebras [**5**]. Using representation theory, we now characterize the elements with finite geometric rank and the geometrically compact elements for a variety of nonselfadjoint algebras.

Our selfadjoint work in [**5**] suggests the following definition. (Also compare with [**47**].)

Definition 2.1. An operator algebra $\mathcal{A}$ is said to be *operator semisimple* if there exists a family of Hilbert space representations $(\tau_a, \mathcal{H}_a)$ of $\mathcal{A}$, $a \in \mathbb{A}$, such that their direct sum $\tau = \bigoplus_{a\in\mathbb{A}} \tau_a$ is an isometric isomorphism of $\mathcal{A}$ that maps the $(1+\epsilon)$-ball of $\mathcal{A}$ onto a strongly dense subset of $\bigoplus_{a\in\mathbb{A}} \mathcal{B}(\mathcal{H}_a)_1$, for some $\epsilon > 0$. The family $(\tau_a, \mathcal{H}_a), a \in \mathbb{A}$ is said to implement the operator semisimplicity.

By Lemma 2.1 in [**25**], each of the representations τ_a, $a \in \mathbb{A}$, is algebraically irreducible. Since $\tau = \bigoplus_{a\in\mathbb{A}} \tau_a$ is faithful for $\mathcal{A}$ we conclude that the intersection of all kernels of algebraically irreducible representations for $\mathcal{A}$ equals zero, i.e., an operator semisimple algebra is indeed semisimple.

The next result provides additional information for the operator A appearing in Theorem 1.6 (iii).

Lemma 2.2. *Let $\mathcal{A}$ be an operator semisimple algebra and let $(\tau_a, \mathcal{H}_a)$, $a \in \mathbb{A}$, be the family of representations of $\mathcal{A}$ implementing the operator semisimplicity. If A is a geometrically compact element of $\mathcal{A}$ then $\tau(A)$ is a compact operator.*

Proof. Proposition 1.4 shows that the norm closure of $A\mathcal{A}_{1/2}A$ is contained in $\mathrm{cp}^{(2)}(\{A\})$, which is a norm compact set. Therefore, the norm closure of $\tau(A)\tau(\mathcal{A}_{1+\epsilon})\tau(A)$ is also compact. However, the weak closure of $\tau(\mathcal{A}_{1+\epsilon})$ contains $\mathcal{B}(\mathcal{H})_1$ and so the norm closure of

$$\tau(A)\mathcal{B}(\mathcal{H})_1\tau(A)$$

is norm compact.

We now prove that $\tau_a(A)$ is a compact operator, for any $a \in \mathbb{A}$. Let $e \in \mathcal{H}_a$ be such that $\tau_a(A)^*e \neq 0$ and let $\{f_k\}_{k=1}^{\infty}$ be an arbitrary sequence of unit vectors from $\mathcal{H}_a$. By what we saw the previous paragraph, the sequence $\left\{\tau_a(A)(e \otimes f_k)\tau_a(A)\right\}_{k=1}^{\infty}$ has a norm convergent subsequence. However,

$$\tau_a(A)\,(e \otimes f_k)\,\tau_a(A) = (\tau_a(A)^*e) \otimes (\tau_a(A)f_k)\,,$$

so the sequence $\{\tau_a(A)f_k\}_{k=1}^{\infty}$ has a norm convergent subsequence. This proves that $\tau_a(A)$ is a compact operator.

It remains to show that $\tau(A)$ is a compact operator. Let $\mathbb{B}$ be the collection of all finite subsets of $\mathbb{A}$. For each $b \in \mathbb{B}$, let T_b be the diagonal operator satisfying $T_b|_{\mathcal{H}_a} = \tau_a(A)$, for all $a \in b$, and $T_b|_{\mathcal{H}_a} = 0$ otherwise. The previous paragraph shows that T_b is a compact operator, for any $b \in \mathbb{B}$. It suffices to show that the net $\{T_b\}_{b\in\mathbb{B}}$ converges in norm to $\tau(A)$.

By way of contradiction assume that the net $\{T_b\}_{b\in\mathbb{B}}$ does not converge in norm to $\tau(A)$. This is easily seen to imply the existence of an $\epsilon > 0$ and a sequence $\{a_n\}_{n\in\mathbb{N}} \subseteq \mathbb{A}$ such that $\|\tau_{a_n}(A)\| \geq \epsilon$, for all $n \in \mathbb{N}$. Therefore there exist unit vectors $f_n \in \mathcal{H}_{a_n}$ such that $\|\tau_{a_n}(A)f_n\| \geq \epsilon$, for all $n \in \mathbb{N}$. However, the sequence $\{f_n\}_{n\in\mathbb{N}}$ converges weakly to zero and so an argument similar to that of the second paragraph of the proof shows that the sequence $\|\tau_{a_n}(A)f_n\|$ converges to zero, a contradiction. □

Lemma 2.3. *Let $\mathcal{A}$ be an operator semisimple algebra, let $(\tau_a, \mathcal{H}_a)$, $a \in \mathbb{A}$, be the family of representations of $\mathcal{A}$ implementing the operator semisimplicity and let $\tau = \bigoplus_{a\in\mathbb{A}} \tau_a$. Then the set of compact operators in $\tau(\mathcal{A})$ forms a* C*-algebra.*

Proof. Let $T = \bigoplus_{a\in\mathbb{A}} T_a$ be a compact operator in $\tau(\mathcal{A})$. Fix an $a_0 \in \mathbb{A}$ and let $\{A_i\}_{i\in I}$ be a bounded net in $\tau(\mathcal{A})$ converging strongly to the operator that equals $T_{a_0}^{\star}$ on $\mathcal{H}_{a_0}$ and 0 everywhere else. Then $\{A_iT\}_{i\in I}$ converges in norm to a positive compact operator supported on $\mathcal{H}_{a_0}$. An application of the Spectral Theorem shows now that $\tau(\mathcal{A})$ contains a finite rank projection P supported on $\mathcal{H}_{a_0}$.

We claim that $\tau(\mathcal{A})$ contains all rank one operators supported on $\mathcal{H}_{a_0}$. (This will imply that $\tau(\mathcal{A})$ contains all compact operators supported on $\mathcal{H}_{a_0}$ and in particular $T_{a_0}^*$.) Indeed, fix a unit vector $g \in P(\mathcal{H})$ and let $e \otimes f$ be any rank one operator supported on $\mathcal{H}_{a_0}$. Since τ implements an operator semisimplicity, there exists a bounded net $\{B_i\}_{i\in I}$ in $\tau(\mathcal{A})$ converging strongly to $g \otimes f$. Hence, the net $\{B_iP\}_{i\in I}$ converges in norm to $(g \otimes f)P = g \otimes f$ and so $g \otimes f \in \tau(\mathcal{A})$. Similarly, there exists a bounded net $\{C_i\}_{i\in I}$ in $\tau(\mathcal{A})$ converging strongly to $e \otimes g$ and so $\{C_i^*\}_{i\in I}$ converges weakly to $g \otimes e$. By [**11**, Corollary IX.5.2], there exists a net $\{D_j\}_{j\in J} \subseteq \tau(\mathcal{A})$ consisting of convex combinations from $\{C_i\}_{i\in I}$ and such that $\{D_j^*\}_{j\in J}$ converges strongly to $g \otimes e$. Hence, the net $\{D_j^*P\}_{j\in J}$ converges in norm to $(g \otimes e)P = g \otimes e$ and so $g \otimes e \in (\tau(\mathcal{A}))^*$. Hence, $e \otimes g \in \tau(\mathcal{A})$ and so

$$e \otimes f = (g \otimes f)(e \otimes g) \in \tau(\mathcal{A}),$$

as desired.

Finally, an approximation argument similar to that of the last paragraphs of the proof of Lemma 2.1, combined with the above claim, shows that $T^* \in \tau(\mathcal{A})$. □

The next result clarifies the nature of the geometrically compact elements in operator semisimple algebras and generalizes the main result in [**5**].

Theorem 2.4. *Let $\mathcal{A}$ be an operator semisimple algebra and let $A \in \mathcal{A}$. Then A is geometrically compact if and only if there exists an isometric representation φ of $\mathcal{A}$ such that $\varphi(A)$ is a compact operator.*

Proof. If $\mathcal{A}$ is geometrically compact then Lemma 2.2 shows that $\tau(A)$ is a compact operator.

Conversely, assume that there exists an isometric representation φ of $\mathcal{A}$ such that $\varphi(A)$ is a compact operator. Then $\varphi(A)\varphi(\mathcal{A}_1)\varphi(A)$ is a compact set, so $\tau(A)\tau(\mathcal{A}_1)\tau(A)$ is a compact set contained in a C*-algebra consisting of compact operators (Lemma 2.3). The rest of the proof now follows from our selfadjoint arguments in [**5**, Theorem 2.2]. □

Corollary 2.5. *The geometrically compact elements of an operator semisimple algebra $\mathcal{A}$ form a C*-algebra.*

Proof. In light of Lemma 2.3 and Theorem 2.4, it suffices to show that τ^{-1} is a $*$-homomorphism when restricted to the set of compact operators in $\tau(\mathcal{A})$. However, τ^{-1} is an isometry and therefore it preserves selfadjoint projections. By the spectral theorem it preserves all selfadjoint compact operators and the conclusion follows. □

A minor modification of Lemma 2.2 shows that the socle of an operator semisimple algebra coincides with the set of all operators that can be isometrically represented as finite rank operators. Therefore Theorem 2.4 identifies the socle of such an algebra as the set of all elements with finite geometric rank.

We are now able to give a criterion for when the unit ball of an operator semisimple algebra contains nonzero geometrically compact elements.

Theorem 2.6. *If $\mathcal{A}$ is an operator semisimple algebra, then the following statements are equivalent:*

(i) *The unit ball of $\mathcal{A}$ has nontrivial compact faces.*
(ii) *The unit ball of $\mathcal{A}$ has nontrivial finite-dimensional faces.*
(iii) *$\mathcal{A}$ contains nonzero geometrically compact elements.*
(iv) *$\mathcal{A}$ contains a nonzero atom P, i.e., a nonzero selfadjoint projection $P \in \mathcal{A}$ such that* $\dim P\mathcal{A}P < \infty$.

Proof. (ii) ⇒ (i): trivial.

(i) ⇒ (iii): this follows from Theorem 1.3.

(iii) ⇒ (iv): let $\mathcal{A}$ contain a nonzero geometrically compact element A. By Corollary 2.5 the geometrically compact elements form a C*-subalgebra $\mathcal{J} \subseteq \mathcal{A}$ consisting of compact operators. Hence $\mathcal{J}$ contains an atom; since $\mathcal{J} \subseteq \mathcal{A}$ is an ideal, so does $\mathcal{A}$.

(iv) $\Rightarrow$ (ii): we claim that

$$\mathcal{F}(I-P) = \big\{(I-P) + X \mid X = PXP \in \mathcal{A}_1\big\}$$

is a face. By a standard argument, it suffices to show that if $\mathcal{F}(I-P)$ contains the midpoint of a line segment in the unit ball of $\mathcal{A}$, then the endpoints of the line segment also lie in $\mathcal{F}(I-P)$. Take $A, B \in \mathcal{A}$ with $A = (I-P) + X$ and $X = PXP$, and assume that $\|A \pm B\| \leq 1$. We need to show that $A \pm B \in \mathcal{F}(I-P)$.

Indeed, arguing as in the proof of Proposition 1.4, we produce bounded operators S and T such that

$$B = S(I - A^*A)^{\frac{1}{2}} = (I - AA^*)^{\frac{1}{2}}T.$$

However, $A^*A = (I-P) + PX^*XP$ and so

$$I - A^*A = P - PX^*XP = (I - A^*A)P.$$

Therefore, $(I - A^*A)^{\frac{1}{2}} = (I - A^*A)^{\frac{1}{2}}P$ and so

$$BP = S(I - A^*A)^{\frac{1}{2}}P = B.$$

A similar argument shows that $(I - AA^*)^{\frac{1}{2}} = P(I - AA^*)^{\frac{1}{2}}$ and so $B = PB$. Hence $B = PBP$, as desired. □

Every operator algebra that contains the compact operators acts irreducibly and is therefore operator semisimple. The existence of finite-dimensional faces is not an issue here since any such algebra contains atoms. A later result, Theorem 4.3, is relevant.

By the *reduced atomic representation* of a C*-algebra $\mathcal{A}$ we mean a representation $\tau = \bigoplus_{a\in\mathbb{A}} \tau_a$, where $\{\tau_a\}_{a\in\mathbb{A}}$ is a maximal family of pairwise inequivalent irreducible representations of $\mathcal{A}$. It is an old result in C*-algebra theory (see Proposition 13.10.13 in [**30**]) that the reduced atomic representation satisfies the properties of Definition 2.1. Therefore all C*-algebras are operator semisimple and so Theorem 2.6 applies here. (This was implicit in our earlier work in [**5**].)

A norm closed subalgebra $\mathcal{A}$ of an AF C*-algebra is a (strongly maximal) TAF algebra if and only if it is the limit $\varinjlim(\mathcal{A}_i, \varphi_i)$ of a directed system

$$\mathcal{A}_1 \xrightarrow{\varphi_1} \mathcal{A}_2 \xrightarrow{\varphi_2} \mathcal{A}_3 \xrightarrow{\varphi_3} \mathcal{A}_4 \xrightarrow{\varphi_4} \cdots \tag{5}$$

where for each $i \geq 1$, the subalgebra $\mathcal{A}_i$ satisfies:

(i) $\mathcal{A}_i$ is a direct sum of upper triangular matrix algebras.
(ii) φ_i extends to a $*$-monomorphism from $\mathrm{C}^*(\mathcal{A}_i) \equiv \overline{\mathcal{A}_i + \mathcal{A}_i^*}$ to $\mathrm{C}^*(\mathcal{A}_{i+1})$.
(iii) The extension of φ_i maps matrix units to sums of matrix units.

We call (5) a *presentation* for the algebra; clearly it is not unique. TAF algebras have received much attention in recent years. A good reference for these and more general limit algebras is Power's monograph [**44**].

Two well-known examples of TAF algebras are the standard and refinement algebras. Define the *standard embedding* σ_k by

$$\sigma_k(A) = A \oplus A \oplus \cdots \oplus A \quad (k \text{ factors})$$

and the *refinement embedding* ρ_k by

$$\rho_k([a_{s,t}]) = [a_{s,t} I_k],$$

where I_k is the $k \times k$ identity matrix. If all the embeddings φ_i in the direct limit $\mathcal{A} = \varinjlim(\mathcal{A}_i, \varphi_i)$ are standard embeddings, $\mathcal{A}$ is said to be a standard algebra. If all the φ_i are refinement embeddings, $\mathcal{A}$ is said to be a refinement algebra. An *alternation* algebra is a TAF algebra $\mathcal{A} = \varinjlim(\mathcal{A}_i, \varphi_i)$, where the φ_i alternate between the standard and the refinement embeddings.

The embeddings φ_i are said to be mixing if for every matrix unit $e^{(i)}_{s,t}$ we have $\varphi_i(e^{(i)}_{s,t})\mathcal{A}_{i+1}\varphi_i(e^{(i)}_{s,t}) \neq 0$. The property of an embedding being mixing was introduced by Donsig in his study of semisimplicity [**19**] and was further exploited in [**13**]. It turns out that a TAF algebra $\mathcal{A}$ is semisimple if it admits a presentation $\mathcal{A} = \varinjlim(\mathcal{A}_i, \varphi_i)$, where all the embeddings φ_i are mixing. All standard embeddings are mixing, so the standard and alternation algebras are semisimple.

Let $\mathfrak{A} = \varinjlim(\mathfrak{A}_i, \varphi_i)$ be an AF C*-algebra and assume each $\mathfrak{A}_i$ decomposes as a direct sum $\mathfrak{A}_i = \bigoplus_j \mathfrak{A}_{ij}$ of finite-dimensional full matrix algebras $\mathfrak{A}_{ij}$. A *path* Γ for $\mathfrak{A} = \varinjlim(\mathfrak{A}_i, \varphi_i)$ is a sequence $\{\mathfrak{A}_{ij_i}\}_{i=1}^{\infty}$ such that for each pair of nodes $\big((i, j_i), (i+1, j_{i+1})\big)$ there exists an arrow in the Bratteli diagram for $\mathfrak{A} = \varinjlim(\mathfrak{A}_i, \varphi_i)$ joining them.

For each path Γ consider a subsystem of the directed limit $\varinjlim(\mathfrak{A}_i, \varphi_i)$ consisting of all the summands of $\mathfrak{A}$ that are never mapped into some $\mathfrak{A}_{ij_i} \in \Gamma$. Evidently this system is hereditary and directed upwards; therefore it determines an ideal $\mathcal{J}_\Gamma$ of $\mathfrak{A}$. The quotient $\mathfrak{A}/\mathcal{J}_\Gamma$ is the AF algebra corresponding to the remaining summands and the remaining embeddings. The summands that eventually get mapped into some $\mathfrak{A}_{ij_i} \in \Gamma$ are denoted by $\operatorname{summ}(\Gamma)$. Two paths $\Gamma = \{\mathfrak{A}_{ij_i}\}_{i=1}^{\infty}$ and $\Gamma' = \{\mathfrak{A}_{ij_i'}\}_{i=1}^{\infty}$ are said to be *disjoint* if for all but finitely many $i \in \mathbb{N}$, the nodes $(i\, j_i)$ and $(i\, j_i')$ are distinct.

Lemma 2.7. *Let $\Gamma = \{\mathfrak{A}_{ij_i}\}_{i=1}^{\infty}$ and $\Gamma' = \{\mathfrak{A}_{ij_i'}\}_{i=1}^{\infty}$ be two disjoint paths for an* AF C*-*algebra* $\mathfrak{A} = \varinjlim(\mathfrak{A}_i, \varphi_i)$ *and let $\omega = \varinjlim \omega_i$ and $\omega' = \varinjlim \omega_i'$ be pure states such that the ω_i and ω_i' are supported on $\mathfrak{A}_{ij_i}$ and $\mathfrak{A}_{ij_i'}$ respectively. Then the states ω and ω' induce inequivalent* GNS *representations.*

Proof. Assume that the states ω and ω' produce equivalent irreducible representations π and π'. Then by [**30**, Theorem 10.2.6], there exists a unitary $u \in \mathfrak{A}$ such that $\omega = \omega' \operatorname{ad}_u$, where $\operatorname{ad}_u(a) = uau^*$. Every unitary in an AF algebra is a limit of unitaries in its finite-dimensional subalgebras. Hence there is a unitary v in some $\mathfrak{B}_i$ such that $\|\omega - \omega' \operatorname{ad}_v\| < 1$. However the

disjointness of the paths Γ and Γ' shows that this cannot occur and the conclusion follows. □

In [**13**] it was shown that the quotient $\mathcal{A}/\mathcal{J}$ of a TAF algebra $\mathcal{A}$ by a prime ideal $\mathcal{J}$ is operator primitive, i.e., it admits an isometric representation on a Hilbert space $\mathcal{H}$ such that the 2-ball of $\mathcal{A}/\mathcal{J}$ is weakly dense in the unit ball of $\mathcal{B}(\mathcal{H})$. In particular, a semisimple TAF subalgebra $\mathcal{A}$ of a *primitive* C*-algebra is operator primitive. We now generalize this to an arbitrary semisimple TAF algebra.

Theorem 2.8. *A semisimple* TAF *algebra* $\mathcal{A} = \varinjlim(\mathcal{A}_i, \varphi_i)$ *is operator semisimple.*

Proof. Before embarking on the proof we establish some terminology. If e_1 and e_2 are matrix units in $\mathcal{A}_{i_1}$ and $\mathcal{A}_{i_2}$, $i_1 < i_2$, then we say that e_2 is a *subordinate* of e_1 if there exists a diagonal matrix unit $p \in \mathcal{A}_{i_2}$ such that $e_2 = pe_1$. If e_2 and f_2 are subordinates of e_1 and f_1, then we say that e_2 *majorizes* f_2 if there exist matrix units $s, t \in \mathcal{A}_{i_2}$ such that $e_2 = sf_2t$.

For the proof, start with any direct summand $\mathfrak{A}^{(1)}_{1,j_1}$ of $\mathfrak{A}_1$. Let $e^{(1)}_1$ be the characteristic vector of $\mathfrak{A}^{(1)}_{1,j_1}$, i.e., the vector at the top right corner of $\mathfrak{A}^{(1)}_{1,j_1}$. Since $\mathcal{A}$ is semisimple there exist a link $f^{(1)}_2$ for $e^{(1)}_1$ in some later summand of $\mathcal{A}$. For notational convenience we assume that $f^{(1)}_2$ belongs to some summand $\mathfrak{A}^{(1)}_{2,j_2}$ of $\mathfrak{A}_2$ and $\varphi^{(1)}_{1,j_1}$ is the mixing embedding from $\mathfrak{A}^{(1)}_{1,j_1}$ into $\mathfrak{A}^{(1)}_{2,j_2}$ that maps $e^{(1)}_1$ onto two copies, say $e^{(1)}_{(1,1)}$ and $e^{(1)}_{(1,2)}$, linked by $f^{(1)}_2$. Now let $e^{(1)}_2$ be the characteristic matrix unit of $\mathfrak{A}^{(1)}_{2,j_2}$ and let $f^{(1)}_3$ be a link for $e^{(1)}_2$ in $\mathfrak{A}^{(1)}_{3,j_3}$ and $\varphi^{(1)}_{2,j_2}$ the corresponding linking mapping. This way we construct a path $\Gamma^{(1)}$ for $\varinjlim(\mathfrak{A}_i, \varphi_i)$. If $\mathrm{summ}(\Gamma^{(1)})$ contains all the summands for all the finite-dimensional algebras $\mathfrak{A}_i$ we stop. Otherwise we chose a direct summand of $\mathfrak{A}$ not in $\mathrm{summ}(\Gamma^{(1)})$ and we repeat the process described earlier. This way we define inductively a sequence $\{\Gamma^{(a)}\}_{a\in\mathbb{N}}$ of paths such that all maps involved with the nodes of the path are mixing and the union

$$\bigcup_{a\in\mathbb{N}} \mathrm{summ}(\Gamma^{(a)})$$

contains all direct summands of $\mathfrak{A}$.

Given any path $\Gamma^{(a)}$ defined above, we now construct a pure state ω_a as follows: start with any unit vector, i.e., normalized sum of diagonal matrix units, $\xi^{(a)}_1$ in the central projection determined by $\mathfrak{A}^{(a)}_{1,j_1}$. Let $\omega^{(a)}_1$ be the vector state on $\mathfrak{A}_1$ determined by $\xi^{(a)}_1$, i.e., $\omega^{(a)}_1(A) = \langle A\xi^{(a)}_1, \xi^{(a)}_1\rangle$. Notice that $\xi^{(a)}_1$, being a sum of diagonal matrix units, it is mapped by $\varphi^{(1)}_{1,j_1}$ onto several copies in $\mathfrak{A}_2$. Two of them, say $\zeta^{(a)}_1$ and $\zeta^{(a)}_2$, are majorized by $e^{(a)}_{(1,1)}$

and $e^{(a)}_{(1,2)}$, respectively. Let

$$\xi_2^{(a)} = \frac{1}{\sqrt{2}}\big(\zeta_1^{(a)} + \zeta_2^{(a)}\big)$$

and let $\omega_2^{(a)}$ be the vector state on $\mathfrak{A}_2$ determined by $\xi_2^{(a)}$. Consequently, find two copies $\zeta_1^{(a)}$ and $\zeta_2^{(a)}$ of the image of $\xi_2^{(a)}$ in $\mathfrak{A}_3$ subordinated by $e^{(a)}_{(2,1)}$ and $e^{(a)}_{(2,2)}$, etc. This way we construct inductively a sequence $\{\omega_a\}_{a\in\mathbb{N}}$ of vector states. Since the states ω_a are supported in single summands of $\mathfrak{A}$, they are pure states and hence their direct limit ω_a is also pure.

Let $(\tau_a, \mathcal{H}_a, g_a)$ be the GNS representation induced by ω_a, $a \in \mathbb{N}$, i.e., $\omega_a(A) = \langle \tau_a(A)g_a, g_a\rangle$, $A \in \mathfrak{A}$. By Lemma 2.7 the representations τ_a are mutually inequivalent. Therefore, Corollary 10.3.9 in [**30**] shows that the representation $\tau = \bigoplus_{a\in\mathbb{N}} \tau_a$ maps $\mathfrak{A}$ onto a weakly dense subalgebra of $\bigoplus_{a\in\mathbb{N}} \mathcal{B}(\mathcal{H}_a)$. Since $\bigcup_{a\in\mathbb{N}} \mathrm{summ}(\Gamma^{(a)})$ contains all the direct summands from $\mathfrak{A}$, an easy argument shows that τ is faithful and so isometric. Therefore, Kaplansky's Theorem shows that τ satisfy the requirements of Definition 2.1 for $\mathfrak{A}$. It only remains to show that τ satisfies the same requirements for $\mathcal{A}$.

For each $i \in \mathbb{N}$ we construct a contractive map $\Phi_i : \mathfrak{A}_i \to \mathcal{A}_{i+1}$ as follows: let u be a matrix unit in $\mathfrak{A}_i$. If u does not belong to any of the nodes $\mathfrak{A}^{(a)}_{i,j_i}$ associated with the paths $\Gamma^{(a)}$, we set $\Phi_i(u) = 0$. If $u \in \mathfrak{A}^{(a)}_{i,j_i}$ for some $a \in \mathbb{N}$, we consider the subordinates u_1 and u_2 of u majorized by $e^{(a)}_{(i,1)}$ and $e^{(a)}_{(i,2)}$ respectively. Define $\Phi_i(u)$ to be the matrix unit with initial projection the initial projection of u_2 and final projection that of u_1. Since the matrix units $e^{(a)}_{(i,1)}$ and $e^{(a)}_{(i,2)}$ are linked in $\mathcal{A}_{i+1}$, we have that $\Phi_i(u) \in \mathcal{A}_{i+1}$. Moreover,

$$\omega_a\big(B\,(A - 2\Phi_i(A))\,C\big) = 0$$

for any $A, B, C \in \mathfrak{A}_i$. The weak density of the 2-ball of $\tau(\mathcal{A})$ in $\bigoplus_{a\in\mathbb{N}} \mathcal{B}(\mathcal{H}_a)_1$ now follows from the above equation and the fact that the collection of all vectors of the form

$$\bigoplus_{a\in\mathbb{N}} \tau_a(A_a)g_a, \quad A_a \in \bigcup_{i\in\mathbb{N}} \mathfrak{A}_i,$$

where all but finitely many A_a equal 0, is a dense subset of the space $\bigoplus_{a\in\mathbb{N}} \mathcal{H}_a$. □

The techniques of the previous theorem are also applicable to inductive limits of *block* upper triangular matrices with mixing embeddings, thus showing that such algebras are operator semisimple.

A function $f : \mathcal{X} \longrightarrow \mathbb{C}$ is said to be *unitary* if $|f(x)| = 1$, for all $x \in \mathcal{X}$.

Theorem 2.9. *Let $\mathcal{X}$ be a compact Hausdorff space and let $\mathcal{A} \subseteq \mathrm{C}(\mathcal{X})$ be a norm closed algebra of continuous functions containing the constant functions. If the unitary functions in $\mathcal{A}$ separate the points of $\mathcal{X}$, then $\mathcal{A}$ is an operator semisimple algebra.*

Proof. Let $\mathcal{X} = \{x_i \mid i \in \mathbb{I}\}$ and for each $i \in \mathbb{I}$, consider the one-dimensional representation $(\mathcal{H}_i, \tau_i, z_i)$ such that $f(x_i) = \langle \tau_i(f)z_i, z_i\rangle, f \in \mathcal{A}$. Clearly, the representation $\tau = \bigoplus_{i\in\mathbb{I}} \tau_i$ is the reduced atomic representation of $\mathrm{C}(\mathcal{X})$. Hence, the strong closure of $\tau(\mathrm{C}(\mathcal{X}))$ equals the algebra $\mathcal{D}$ of all diagonal matrices on $\bigoplus_{i\in\mathbb{I}} \mathcal{H}_i$, i.e., $\mathcal{D} = l^\infty(\mathbb{I})$. We are to show that the strong closure of $\tau(\mathcal{A})_1$ equals $\mathcal{D}_1$.

Claim. Given $x_{i_1}, x_{i_2}, \dots x_{i_n} \in \mathcal{X}$ and a unimodular number w, there exists F in the strong closure of $\tau(\mathcal{A})_1$ such that $\langle Fz_{i_1}, z_{i_1}\rangle = w$ and $\langle Fz_{i_k}, z_{i_k}\rangle = 1$, $k = 2, 3, \dots, n$.

In order to prove the claim we make use of Möbius maps; if $g \in \mathcal{A}$ has norm 1 and m is a Möbius transformation, then the composite function $m \circ g$ also belongs to $\mathcal{A}_1$. For the proof, for each $2 \leq l \leq n$ we choose a unitary function g_l that separates x_{i_1} from x_{i_l}. We apply a Möbius map to each g_l to get g'_l such that $g'_l(x_{i_1}) = 1$ and either $g'_l(x_{i_k}) = 1$ or $e^{i\pi\alpha_{kl}}$ with $\alpha_{kl} \in [1/n, 2/n]$. If $g = \prod_l g'_l$ then $g(x_{i_1}) = 1$ and $g(x_{i_l}) = e^{\pi i\alpha_l}$, where $\alpha_l \in [1/n,\ 2-2/n]$. Now another application of a Möbius map produces a function $f_s \in \mathcal{A}$ such that $f_s(x_{i_1}) = w$ and $|f_s(x_{i_l}) - 1| \leq 1/s$, $l = 2, \dots, n$. Any weak limit of $\{\tau(f_s)\}_{s\in\mathbb{N}}$ proves the claim.

Since the strong closure of $\tau(\mathcal{A})_1$ is closed under multiplication, repeated use of the claim shows that given $x_{i_1}, x_{i_2}, \dots x_{i_n} \in \mathcal{X}$ and a unimodular n-tuple $(w_1, w_2, \dots, w_n)$, there exists F in the strong closure of $\tau(\mathcal{A})_1$ such that $\langle Fz_{i_j}, z_{i_j}\rangle = w_j$. A convexity argument now shows the desired equality. □

Corollary 2.10. *The Hardy space* H^∞ *and the disc algebra* $\mathrm{A}(\mathbb{D})$ *are operator semisimple.*

Proof. Both satisfy the requirements of Theorem 2.9 [**24**, page 174]. □

One of the pleasing features of the representations τ in the proofs of Theorems 2.8 and 2.9 is that they extend to a $*$-representation of their enveloping C*-algebra. (We coin the term C*-*semisimple* for an operator semisimple algebra admitting such a representation.) Using the fact that the spatial norm on the tensor product of two C*-algebras is minimal, one can easily conclude that the spatial tensor product of C*-semisimple algebras is also C*-semisimple. Hence tensor products between C*-algebras, semisimple TAF algebras, Douglas algebras and irreducible algebras are operator semisimple thus providing additional examples of such algebras.

A subalgebra $\mathcal{A}$ of a C*-algebra $\mathfrak{A}$ is said to be *Dirichlet* if it satisfies $\overline{\mathcal{A} + \mathcal{A}^*} = \mathfrak{A}$. The prototypical example of a Dirichlet algebra is the disc algebra, as a subalgebra of the continuous functions on the circle. The term originates from the theory of functions. It has also been studied in the context of nonselfadjoint operator algebras by Arveson [**9**], Muhly and Solel [**35**] and others. All strongly maximal TAF algebras are easily seen to be Dirichlet.

Proposition 2.11. *Suppose that $\mathcal{A}$ is a Dirichlet subalgebra of an infinite-dimensional simple* C*-*algebra* $\mathfrak{A}$. *If $\mathcal{A}$ is operator semisimple, the unit ball of $\mathcal{A}$ does not have any nontrivial compact faces.*

Proof. Assume that the unit ball of $\mathcal{A}$ has a nontrivial compact face. Then Theorem 2.6 shows that $\mathcal{A}$ contains a nonzero atom P. Hence,

$$P\mathfrak{A}P = P\big(\overline{\mathcal{A}+\mathcal{A}^*}\big)P \subseteq \overline{P\mathcal{A}P} + \overline{(P\mathcal{A}P)^*} = \mathbb{C}P$$

and so $\mathfrak{A}$ contains a nonzero atom. But then, Theorem 2.6 contradicts the simplicity of $\mathfrak{A}$. □

Combining Proposition 2.11 with Theorem 2.8 we obtain that the unit ball of the familiar standard, alternation and $A(\mathbb{Q},\nu)$ algebras contain no nontrivial compact faces. (We do not know if the same is true for the unit ball of the refinement algebra.) Proposition 2.11 also applies to semicrossed products $\mathrm{C}(\mathcal{X}) \times_\alpha \mathbb{Z}^+$ corresponding to minimal actions α on a compact metric space $\mathcal{X}$. In [**14**] it is shown that such semicrossed products are operator primitive. Since these are Dirichlet subalgebras of simple C*-algebras their unit ball contains no nontrivial compact faces.

3. Spectral obstructions for the existence of compact faces

In this section we obtain noncommutative analogs of Corollary 1.7. A free semigroup algebra is the weakly closed algebra generated by n isometries with orthogonal ranges. The central example for these algebras is the "noncommutative Toeplitz algebra" $\mathcal{L}_n$ generated by the left regular representation of the free semigroup on n letters. The study of $\mathcal{L}_n$ was initiated by Popescu [**41**, **42**, **43**] in the context of dilation theory. A detailed analysis of $\mathcal{L}_n$ is contained in the papers of Davidson and Pitts [**16**, **17**], Kribbs [**33**] and Arias and Popescu [**7**, **8**], which develop the analytic structure. Apart from being a good example, it turns out that $\mathcal{L}_n$ is also a model for free semigroup algebras: the Structure Theorem in [**15**] shows that every free semigroup algebra has a 2×2 lower triangular form where the first column is a slice of a von Neumann algebra and the 22 entry is an algebra isomorphic to $\mathcal{L}_n$. Special classes of free semigroup algebras are important in the classification of certain representations of the Cuntz–Toeplitz algebra [**16**] and the construction of wavelets [**10**]

A continuous factor representation of the Cuntz algebra induces a nonatomic free semigroup algebra. Moreover, $\mathcal{L}_n$ is also nonatomic. Our next result shows that the unit ball of such free semigroup algebras contains no nontrivial geometrically compact elements. Specifically:

Theorem 3.1. *If $\mathcal{A}$ is a free semigroup algebra, the following statements are equivalent:*

(i) *The unit ball of $\mathcal{A}$ has nontrivial compact faces.*

(ii) *The unit ball of $\mathcal{A}$ has nontrivial finite-dimensional faces.*
(iii) *$\mathcal{A}$ contains nonzero geometrically compact elements.*
(iv) *$\mathcal{A}$ contains an atom.*

Proof. As in the proof of Theorem 2.6, we only need to verify that (iii) implies (iv). Let $\mathcal{A}$ be a free semigroup algebra with no atoms. By Theorem 2.6 in [**15**], $\mathcal{A}$ decomposes via a projection P in $\mathcal{A}$ as $\mathcal{A} = \mathcal{M}P + P^\perp \mathcal{A} P^\perp$, where $\mathcal{M}$ is the von Neumann algebra generated by $\mathcal{A}$ and $P^\perp \mathcal{A}|_{P^\perp \mathcal{H}}$ is isomorphic to $\mathcal{L}_n$. Let π be the isomorphism from $P^\perp \mathcal{A}|_{P^\perp \mathcal{H}}$ onto $\mathcal{L}_n$.

By way of contradiction assume that A is a nontrivial geometrically compact element of $\mathcal{A}$. Then Proposition 1.4 shows that

$$(A\mathcal{B}(\mathcal{H})_{1/2}A) \cap \mathcal{A} \subseteq \mathrm{cp}(\{B\}); \tag{6}$$

and so the norm closure of $A\mathcal{A}_1 A$ is norm compact.

We now claim that $P^\perp A P^\perp = 0$. Indeed, notice that

$$P^\perp A P^\perp (P^\perp \mathcal{A} P^\perp)_1 P^\perp A P^\perp \subseteq P^\perp \left(A\mathcal{A}_1 A\right) P^\perp$$

is norm compact. Therefore, by applying π, we obtain a nonzero element $\hat{A} := \pi(P^\perp A P^\perp)$ of $\mathcal{L}_n$ such that the closure of $\hat{A}(\mathcal{L}_n)_1 \hat{A}$ is norm compact. Then either $A = 0$ or, arguing as in the proof of Theorem 1.6, we conclude that the spectrum of $\hat{A}$ is countable. However, [**16**, Corollary 1.8] asserts that the spectrum of every nonscalar element of $\mathcal{L}_n$ is connected and contains more than one point. Since $\mathcal{L}_n$ is infinite-dimensional, $\hat{A} = 0$.

The claim above, combined with the first paragraph of the proof, shows that $A = AP$. Now let $I - Q$ be the span of all atoms in $\mathcal{M}$. It is well-known that Q is a central projection in $\mathcal{M}$ and that $Q\mathcal{M}Q$ is nonatomic. Since $\mathcal{A}$ contains no atoms, $P\mathcal{M}P$ is nonatomic and so $P \subseteq Q$. Hence

$$A\mathcal{M}_1 A = AP\mathcal{M}_1 A = A(Q\mathcal{M}Q)_1 A.$$

At the same time,

$$A\mathcal{M}A = A\mathcal{M}AP \subseteq \mathcal{M}P \subseteq \mathcal{A};$$

therefore, by (6), A is a geometrically compact element of the nonatomic algebra $Q\mathcal{M}Q$. Hence Theorem 2.6 implies $A = 0$, a contradiction. □

In spite of Theorem 3.1, the unit ball of a free semigroup algebra $\mathcal{A}$ always contains nontrivial faces. Indeed, if $\mathcal{A}$ contains a slice of a von Neumann algebra, this follows from Theorem 2.6. So it suffices to consider only the case where $\mathcal{A} = \mathcal{L}_n$.

Proposition 3.2. *The unit ball of $\mathcal{L}_n$ contains nontrivial faces.*

Proof. It is enough to prove that the boundary of the unit ball of $\mathcal{L}_n$ contains composite points. Indeed, for any such point A, the face $\mathcal{F}(A)$ is not a singleton or else A is an extreme point.

The wandering subspaces for any of the left creation operators L_i span the Fock space $\mathcal{H}$. We therefore define an H^∞ functional calculus for L_i, which is well-defined, linear and isometric, as follows: if $\varphi \in \mathrm{H}^\infty$ then $\varphi(L_i) := \lim p_n(L_i)$, where $\{p_n\}_{n\in\mathbb{N}}$ is any bounded sequence of polynomials converging to φ.

Let φ be a composite point on the boundary of the unit ball of H^∞ (the existence of such a φ follows from the description of the extreme points of H^∞ in [**24**]). Then $\varphi(L_i)$ is the desired composite point. □

The lack of nontrivial compact faces for the unit ball of $\mathcal{L}_n$ is due to the fact that the spectrum of any nonscalar operator in $\mathcal{L}_n$ is infinite and connected. In [**28**], Hoover, Peters and Wogen study the spectral properties of algebras that contain no zero divisors. The spectrum of any operator in such an algebra $\mathcal{A}$ is necessarily connected. However, $\mathcal{A}$ may contain nonzero quasinilpotents, so the arguments in the proof of Theorem 3.1 do not apply in this generality. Nevertheless, the unit ball of such an algebra $\mathcal{A}$ does not contain any nontrivial finite-dimensional faces. Indeed, $\mathcal{A}$ does not contain any nonzero nilpotent operators, so no finite-dimensional subalgebras apart from the trivial one. The conclusion now follows from Theorem 1.8.

If $\mathcal{A}$ is a Banach algebra and α an automorphism of $\mathcal{A}$, the semicrossed product $\mathcal{A} \times_\alpha \mathbb{Z}^+$ is the enveloping Banach algebra of $l^1(\mathcal{A}, \mathbb{Z}^+, \alpha)$ with respect to the class of contractive Hilbert space representations. An interesting feature of the class of algebras studied by Hoover, Peters and Wogen in [**28**] is that it is closed under semicrossed products by $\mathbb{Z}^+$. (If $\mathcal{A} \times_\alpha \mathbb{Z}^+$ contains x and y such that $xy = 0$, then the lowest Fourier coefficients of x and y are necessarily zero divisors in $\mathcal{A}$.) Hence:

Theorem 3.3. *Let $\mathcal{A}$ be an algebra that contains no zero divisors and let α be an automorphism of $\mathcal{A}$. Then the unit ball of $\mathcal{A} \times_\alpha \mathbb{Z}^+$ does not contain any nontrivial finite-dimensional faces.*

The theorem above applies in particular to the algebras $\mathrm{A}(\mathbb{D}) \times_\alpha \mathbb{Z}^+$ and $\mathrm{H}^\infty \times_\alpha \mathbb{Z}^+$, studied in [**27**].

Notice that $\mathrm{C}(\mathcal{X})$, $\mathcal{X}$ compact metric space, contains zero divisors and so Theorem 3.3 does not apply in this case. Crossed products of the form $\mathrm{C}(\mathcal{X}) \times_\alpha \mathbb{Z}^+$ were studied in the previous section with the use of representation theory. They can also be studied with the use of Corollary 1.10 as follows:

Theorem 3.4. *Let $\mathcal{X}$ be a compact connected metric space and let φ be a homeomorphism of $\mathcal{X}$ with a dense set of recurrent points. Then the unit ball of $\mathrm{C}(\mathcal{X}) \times_\varphi \mathbb{Z}^+$ has no finite-dimensional faces.*

Proof. By [**20**], the algebra $\mathrm{C}(\mathcal{X}) \times_\varphi \mathbb{Z}^+$ is semisimple. We claim that $\mathrm{C}(\mathcal{X}) \times_\varphi \mathbb{Z}^+$ contains no nontrivial idempotents.

Indeed, let $Q \in \mathrm{C}(\mathcal{X}) \times_\varphi \mathbb{Z}^+$ be an idempotent and let $Q = \sum_{i=0}^\infty f_i t^i$, $f_i \in \mathrm{C}(\mathcal{X})$ be its Fourier expansion. Since Q is an idempotent, f_0 is an idempotent. But $\mathcal{X}$ is connected and so f_0 is either 0 or I.

If $f_0 = 0$ then the lowest Fourier coefficient in the expansion of Q^2 is of order 2 and so $f_1 = 0$. Consequently, the lowest Fourier coefficient in the expansion of Q^2 is of order 4, so $f_2 = f_3 = 0$. Repeated applications of this argument show that $f_n = 0$, $n \geq 0$ and so $Q = 0$.

If $f_0 = I$ then the second Fourier coefficient of Q^2 is $2f_1$. Hence $f_1 = 0$. Repeated applications of this argument show that the rest of the Fourier coefficients are equal to 0 and so $Q = I$ in this case. This proves the claim.

Since $\mathrm{C}(\mathcal{X}) \times_\varphi \mathbb{Z}^+$ contains no nontrivial idempotents, the conclusion follows from Corollary 1.10. □

4. Concluding remarks

The operator semisimple and C*-semisimple algebras can be thought as special cases of diagonal algebras. An operator algebra $\mathcal{A}$ is said to be *block diagonalizable* if there exists a bicontinuous (but not necessarily isometric) representation τ, satisfying the rest of the properties in Definition 2.1. Several results of Section 2, such as Lemmas 2.2, 2.3 and one direction in Theorem 2.4, are valid in this more general context.

By the Wedderburn–Artin Theorem, all finite-dimensional semisimple algebras are block diagonalizable. However, there are semisimple operator algebras that fail that property, as the following example shows:

Example 4.1. A semisimple operator algebra that is not block diagonalizable.

Let $\mathcal{L} = \{0, M, N, I\}$, where M, N are closed subspaces satisfying

$$M \cap N = M^\perp \cap N^\perp = 0,$$

so that the sum $M + N$ is not closed. Let $\mathcal{A} = \operatorname{Alg}\mathcal{L}$ be the algebra of all operators leaving both M and N invariant. By an old result of Longstaff, a rank one operator $A \in \mathcal{A}$ is of the form $R = e \otimes f$, where either $e \in M^\perp$ and $f \in N$ or $e \in N^\perp$ and $f \in M$. The rank one subalgebra $\mathcal{R}(\mathcal{L})$ of $\operatorname{Alg}\mathcal{L}$ consists of all sums of rank one operators in $\operatorname{Alg}\mathcal{L}$. In [**39**] it is shown that $\mathcal{R}(\mathcal{L})$ is weakly dense in $\operatorname{Alg}\mathcal{L}$ (rank one density).

The semisimplicity of $\mathcal{A}$ follows from [**31**]. We are to show that $\mathcal{A}$ is not block diagonalizable.

By way of contradiction assume that $(\tau_a, \mathcal{H}_a)$, $a \in \mathbb{A}$, is the family of representations of $\mathcal{A}$ implementing the operator semisimplicity and let $\tau = \bigoplus_{a\in\mathbb{A}} \tau_a$. Notice that $R\mathcal{A}R$ is one-dimensional, for any rank one operator $R \in \mathcal{A}$, and so $\tau(R)$ is also a rank one operator. Since $\mathcal{R}(\mathcal{L})$ is dense in $\operatorname{Alg}\mathcal{L}$, $\tau(\mathcal{A})$ contains the set $\mathcal{K}$ of all compact operators in the diagonal algebra

$\bigoplus_{a\in\mathbb{A}} \mathcal{B}(\mathcal{H}_a)$. The restriction of τ^{-1} on $\mathcal{K}$ is similar to a $*$-representation. Therefore there exists an invertible operator S such that

$$\mathcal{R}(S\mathcal{L}) \subseteq S\tau^{-1}(\mathcal{K})S^{-1} \subseteq \operatorname{Alg} S\mathcal{L}.$$

However, $S\tau^{-1}(\mathcal{K})S^{-1}$ is selfadjoint and so the density of $\mathcal{R}(S\mathcal{L})$ in $\operatorname{Alg} S\mathcal{L}$ implies that $\operatorname{Alg} S\mathcal{L}$ is selfadjoint. Hence $S\mathcal{L}$ is orthocomplemented and so the sum $M + N$ is closed, a contradiction.

In this paper we addressed the problem of existence for nontrivial finite-dimensional faces. The problem of characterizing which operators belong to such faces is much harder. Indeed, by the Krein-Milman Theorem such faces are the closed convex hull of their extreme points and therefore one needs to have a good understanding of the extreme points for the unit ball. This is also corroborated by the following:

Proposition 4.2. *Let $\mathcal{X}$ be a normed space and let x be an element of its unit ball. If x belongs to a finite-dimensional face, there exists an extreme point e for the unit ball of $\mathcal{X}$ and an element f with finite geometric rank such that $x = e + \lambda f$, for some $\lambda \in \mathbb{R}$.*

Proof. Since $\mathcal{F}$ is finite-dimensional, $\mathcal{F}(x)$ is also finite-dimensional and by the Krein-Milman Theorem it contains an extreme point e. Moreover there exists an element f in the affine hull of $\mathcal{F}(x)$ such that $x = e + f'$. By Lemma 1.2, the affine hull of $\mathcal{F}(x)$ equals $[\operatorname{cp}(\{x\})]$. Since $\operatorname{cp}(\{x\})$ is convex, a moment's reflection shows that $[\operatorname{cp}(\{x\})]$ consists of multiples of $\operatorname{cp}(\{x\})$ and so there exists an element $f \in \operatorname{cp}(\{x\})$ such that $f' = \lambda f$, for some $\lambda \in \mathbb{R}$. However,

$$\operatorname{cp}^{(2)}(\{f\}) \subseteq \operatorname{cp}^{(3)}(\{x\}) = \operatorname{cp}(\{x\}),$$

which shows that f has finite geometric rank, as desired. □

Even though we have a complete understanding of the elements with finite geometric rank, we are far from characterizing the extreme points for operator semisimple algebras. For instance, the problem of characterizing the extreme points of the standard algebra is open [**26**]. Also compare the nature of extreme points in the unit ball of a C*-algebra with that of H^∞ [**24**, page 138]. However, the situation in operator primitive algebras is much better.

Theorem 4.3. *Let $\mathcal{A}$ be an operator algebra containing the compact operators and let $A \in \mathcal{A}$ with $\|A\| = 1$. Then the operator A belongs to a finite-dimensional face of $\mathcal{A}_1$ if and only if*

$$\dim\big((I - AA^*)\mathcal{A}(I - A^*A)\big) < \infty.$$

Proof. Assume first that $\dim\big((I - AA^*)\mathcal{A}(I - A^*A)\big) < \infty$, so both $I - AA^*$ and $I - A^*A$ are finite rank operators. Hence if P and Q are the range

projections of $(I - A^*A)^{1/2}$ and $(I - AA^*)^{1/2}$ respectively, then $P\mathcal{A}Q \subseteq \mathcal{A}$ is finite-dimensional.

Let $B \in \mathrm{cp}(\{A\})$. Then [**34**, Lemma 1] shows that there exist bounded operators S and T such that

$$B = S(I - A^*A)^{1/2} = (I - AA^*)^{1/2}T$$

and so $B = QBP$. Therefore $\mathrm{cp}(\{A\})$ is a finite-dimensional, and Lemma 1.2 shows that A belongs to the finite-dimensional face $\mathcal{F}(A)$.

Conversely, assume that $\mathcal{F}(A)$ is finite-dimensional face and so, by Theorem 1.8, $\mathrm{S}\left(\mathcal{F}(A)\right) = \mathrm{cp}(\{A\})$ is a finite-dimensional hereditary subalgebra of $\mathcal{A}$. A moment's reflection shows that there exist finite-dimensional projections P, Q such that $\mathrm{cp}(\{A\}) = P\mathcal{A}Q$.

We now claim that

$$(I - P) \cap (I - A^*A)^{1/2}(\mathcal{H}) = (I - Q) \cap (I - AA^*)^{1/2}(\mathcal{H}) = \{0\}.$$

Indeed, by way of contradiction assume that one of the above intersections, say the first one, is nontrivial. Consider a unit vector $(I - AA^*)^{1/2} f \in I - P$ and let e be a vector of norm $1/2$ such that $(I - A^*A)^{1/2} e \neq 0$. Then [**34**, Corollary 1] shows that the rank one operator

$$R = (I - A^*A)^{1/2} e \otimes (I - AA^*)^{1/2} f = (I - AA^*)^{1/2}(e \otimes f)(I - A^*A)^{1/2}$$

belongs to $\mathrm{cp}(\{A\})$ and so our earlier observations show that $R = PRQ$, a contradiction that proves the claim.

From the claim above it follows that

$$(I - P) \cap (I - A^*A)(\mathcal{H}) = (I - Q) \cap (I - AA^*)(\mathcal{H}) = \{0\}.$$

Since both $I - P$ and $I - Q$ are of finite codimension, we conclude that the ranges of both $I - A^*A$ and $I - AA^*$ are finite-dimensional, as promised. □

Arguments similar to the ones above lead to a generalization of Kadison's characterization of the extreme points of the unit ball of a C*-algebra [**29**]. Indeed one can show that a contraction A belongs to a finite-dimensional face of a C*-algebra $\mathcal{A}$ if and only if $\dim\big((I - AA^*)\mathcal{A}(I - A^*A)\big) < \infty$. (This was first shown to us by Anoussis with a different proof.)

Another example that illustrates the complexities associated with extreme points in nonselfadjoint operator algebras is the following: let $\mathcal{A}$ be an operator algebra acting on a Hilbert space $\mathcal{H}$ and consider the operator algebra

$$\mathcal{T}(\mathcal{A}) = \left\{ \begin{pmatrix} \lambda & 0 \\ f & A \end{pmatrix} \;\Big|\; A \in \mathcal{A}, f \in \mathcal{H}, \lambda \in \mathbb{C} \right\}$$

with the obvious multiplications.

Proposition 4.4. *The rank one operator $R = (1, 0) \otimes (0, f)$ is an extreme point for the unit ball of $\mathcal{T}(\mathcal{A})$ if and only if f is a separating unit vector for $\mathcal{A}^*$.*

Proof. First assume f is a separating unit vector for $\mathcal{A}^*$, and let $X \in \mathcal{T}(\mathcal{A})$ be such that $\|R \pm X\| = 1$. Then

$$X^*X \leq I - (1,0) \otimes (1,0)$$

and

$$XX^* \leq I - (0,f) \otimes (0,f).$$

The first inequality shows that $\mathcal{X} \in \mathcal{A}$. The second one shows that the range of $\mathcal{X}$ is orthogonal to f and so $\langle Xg, f\rangle = 0$, $\forall g \in \mathcal{H}$. Hence, $\langle g, X^*f\rangle = 0$, $\forall g \in \mathcal{H}$, and so $X^*f = 0$. Since f is separating for $\mathcal{A}^*$, $X = 0$ and so R is an extreme point.

Reversing the arguments above we obtain the other direction in the Theorem. □

The identification of separating vectors for an operator algebra $\mathcal{A}$, i.e., the cyclic vectors for $\mathcal{A}'$, is an important and difficult problem. Most notable is the work in [**21**] for the adjoint of the unilateral shift.

Note that the algebras of the form $\mathcal{T}(\mathcal{A})$, when $\mathcal{A}^*$ admits separating vectors, show that Theorem 2.4 fails for algebras that are not operator semisimple. Indeed, by way of contradiction assume that an operator $T \in \mathcal{T}(\mathcal{A})$ is geometrically compact if and only if there exists an isometric representation φ of $\mathcal{T}(\mathcal{A})$ such that $\varphi(T)$ is a compact operator. This applies in particular to any operator of the form $T = (e,0) \otimes (0,f)$ and so such an operator is geometrically compact. This contradicts however Proposition 4.4.

Acknowledgements. The author has benefited from many discussions with Mihalis Anoussis. He also acknowledges the assistance of Ken Davidson in the proof of Theorem 2.9.

References

[1] C. Akemann and J. Anderson, *Lyapunov theorems for operator algebras*, Mem. Amer. Math. Soc., **94**(**458**) (1991), MR 1086563 (92e:46113), Zbl 0769.46036.

[2] C.A. Akemann and G.K. Pedersen, *Facial structure in operator algebra theory*, Proc. London Math. Soc., **64** (1992), 418–448, MR 1143231 (93c:46106), Zbl 0759.46050.

[3] J.C. Alexander, *Compact Banach algebras*, Proc. London Math. Soc., **18** (1968), 1–18, MR 0229040 (37 #4618), Zbl 0184.16502.

[4] E. Alfsen and E. Effros, *Structure in real Banach spaces* I, II, Ann. of Math., **96** (1972), 98–128; ibid., **96** (1972), 129–173, MR 0352946 (50 #5432), Zbl 0248.46019.

[5] M. Anoussis and E. Katsoulis, *Compact operators and the geometric structure of* C*-*algebras*, Proc. Amer. Math. Soc., **124** (1996), 2115–2122, MR 1322911 (96i:46068), Zbl 0857.46034.

[6] M. Anoussis and E. Katsoulis, *Compact operators and the geometric structure of nest algebras*, Indiana U. Math. J., **45** (1996), 1175–1191, MR 1444482 (98e:47066a), Zbl 0883.47036.

[7] A. Arias and G. Popescu, *Factorization and reflexivity on Fock spaces*, Integral Equations Operator Theory, **23** (1995), 268–286, MR 1356335 (97e:47066), Zbl 0842.47026.

[8] A. Arias and G. Popescu, *Noncommutative interpolation and Poisson transforms*, Israel J. Math., **115** (2000), 205–234, MR 1749679 (2001i:47021), Zbl 0967.47045.

[9] W. Arveson, *Analyticity in operator algebras*, Amer. J. Math., **89** (1967), 578–642, MR 0223899 (36 #6946), Zbl 0183.42501.

[10] O. Bratteli and P. Jorgensen, *Iterated function systems and permutation representations of the Cuntz algebra*, Mem. Amer. Math. Soc., **139**(**663**), 1999, MR 1469149 (99k:46094a), Zbl 0935.46057.

[11] J. Conway, *A Course in Functional Analysis* (2nd ed.), Graduate Texts in Mathematics, **96**, Springer-Verlag, 1990, MR 1070713 (91e:46001), Zbl 0706.46003.

[12] K. Davidson, C**-Algebras by Example*, Fields Institute Monographs, **6**, American Mathematical Society, 1996, MR 1402012 (97i:46095), Zbl 0958.46029.

[13] K. Davidson and E. Katsoulis, *Primitive limit algebras and* C**-envelopes*, Adv. Math., **170**(**2**) (2002), 181–205, MR 1932328 (2003h:47140), Zbl 1011.46048.

[14] K. Davidson and E. Katsoulis, *work in progress.*

[15] K. Davidson, E. Katsoulis and D. Pitts, *The structure of free semigroup algebras*, J. Reine Angew. Math., **533** (2001), 99–125, MR 1823866 (2002a:47107), Zbl 0967.47047.

[16] K. Davidson and D. Pitts, *Invariant subspaces and hyper-reflexivity for free semigroup algebras*, Proc. London Math. Soc., **78** (1999), 401–430, MR 1665248 (2000k:47005), Zbl 0997.46042.

[17] K. Davidson and D. Pitts, *The algebraic structure of non-commutative analytic Toeplitz algebras*, Math. Ann., **311** (1998), 275–303, MR 1625750 (2001c:47082), Zbl 0939.47060.

[18] K. Davidson and D. Pitts, *Nevanlinna–Pick interpolation for non-commutative analytic Toeplitz algebras*, Integral Equations Operator Theory, **31** (1998), 321–337, MR 1627901 (2000g:47016), Zbl 0917.47017.

[19] A.P. Donsig, *Semisimple triangular* AF *algebras*, J. Funct. Anal., **111** (1993), 323–349, MR 1203457 (94b:46084), Zbl 0808.47031.

[20] A.P. Donsig, A. Katavolos and A. Manoussos, *The Jacobson radical for analytic crossed products*, J. Funct. Anal., **187** (2001), 129–145, MR 1867344 (2002k:46170), Zbl 0998.46034.

[21] R.G. Douglas, H.S. Shapiro and A.L. Shields, *Cyclic vectors and invariant subspaces for the backward shift operator*, Ann. Inst. Fourier (Grenoble), **20** (1970), 37–76, MR 0270196 (42 #5088), Zbl 0186.45302.

[22] C. Edwards and G. Ruttimann, *On the facial structure of the unit balls in a* GL*-space and its dual*, Math. Proc. Cambridge Philos. Soc., **98** (1985), 305–322, MR 0795896 (87a:46114), Zbl 0577.46007.

[23] C. Edwards and G. Ruttimann, *On the facial structure of the unit balls in a* JBW**-triple and its predual*, J. London Math. Soc., **38** (1988), 317–332, MR 0966303 (90b:46129), Zbl 0621.46043.

[24] K. Hoffman, *Banach Spaces of Analytic Functions*, Dover Publications Inc., 1988, MR 1102893 (92d:46066), Zbl 0734.46033.

[25] T. Hudson and E. Katsoulis, *Primitive triangular* UHF *algebras*, J. Funct. Anal., **160** (1998), 1–27, MR 1658724 (99m:46135), Zbl 0923.46053.

[26] T. Hudson, E. Katsoulis and D. Larson, *Extreme points in triangular* UHF *algebras*, Trans. Amer. Math. Soc., **349** (1997), 3391–3400, MR 1407493 (97m:47058), Zbl 0883.47035.

[27] T. Hoover, *Isomorphic operator algebras and conjugate inner functions*, Michigan Math. J., **39** (1992), 229–237, MR 1162033 (93h:47056), Zbl 0771.47013.

[28] T. Hoover, J. Peters and W. Wogen, *Spectral properties of semicrossed products*, Houston J. Math., **19** (1993), 649–660, MR 1251615 (94k:47067), Zbl 0795.46053.

[29] R. Kadison, *Isometries of operator algebras*, Ann. of Math., **54** (1951), 325–338, MR 0043392 (13,256a), Zbl 0045.06201.

[30] R.V. Kadison and J.R. Ringrose, *Fundamentals of the theory of operator algebras*, Vol. II, Academic Press, London, 1983, MR 0859186 (88d:46106), Zbl 0601.46054.

[31] A. Katavolos and E. Katsoulis, *Semisimplicity in operator algebras and subspace lattices*, J. London Math. Soc., **42** (1990), 365–372, MR 1083452 (92b:47066), Zbl 0739.47019.

[32] T. Kato, *Notes on some inequalities for linear operators*, Math. Ann., **125** (1952), 208–212, MR 0053390 (14,766e), Zbl 0048.35301.

[33] D.W. Kribs, *Factoring in non-commutative analytic Toeplitz algebras*, J. Operator Theory, **45**(**1**) (2001), 175–193, MR 1823067 (2002g:47151), Zbl 0996.47065.

[34] R. Moore and T. Trent, *Extreme points of certain operator algebras*, Indiana U. Math. J., **36** (1987), 645–650, MR 0905616 (89d:47103), Zbl 0653.47023.

[35] P. Muhly and B. Solel, *Dilations for representations of triangular algebras* Bull. London Math. Soc., **21** (1989), 489–495, MR 1005829 (90i:47045), Zbl 0721.46034.

[36] G.J. Murphy, C**-Algebras and Operator Theory*, Academic Press, Inc., Boston, MA, 1990, MR 1074574 (91m:46084), Zbl 0714.46041.

[37] J. Orr and J. Peters, *Some representations of* TAF *algebras*, Pacific. J. Math., **167** (1995), 129–161, MR 1318167 (96c:46055), Zbl 0822.46066.

[38] T. Palmer, *Banach Algebras and the General Theory of $*$-Algebras, I: Algebras and Banach Algebras*, Encyclopedia of Mathematics and its applications, **49**, Cambridge University Press, 1994, MR 1270014 (95c:46002), Zbl 0809.46052.

[39] M. Papadakis, *On hyperreflexivity and rank one density for non-*CSL *algebras*, Studia Math., **98** (1991), 11–17, MR 1110095 (92g:47062), Zbl 0755.47030.

[40] G. Pedersen, C**-Algebras and their Automorphism Groups*, London Mathematical Society Monographs, **14**, Academic Press Inc. London-New York, 1979, MR 0548006 (81e:46037), Zbl 0416.46043.

[41] G. Popescu, *Characteristic functions for infinite sequences of noncommuting operators*, J. Operator Theory, **22** (1989), 51–71, MR 1026074 (91m:47012), Zbl 0703.47009.

[42] G. Popescu, *Multi-analytic operators and some factorization theorems*, Indiana Univ. Math. J., **38** (1989), 693-710, MR 1017331 (90k:47019), Zbl 0661.47020.

[43] G. Popescu, *Multi-analytic operators on Fock spaces*, Math. Ann., **303** (1995), 31–46, MR 1348353 (96k:47049), Zbl 0835.47015.

[44] S.C. Power, *Limit Algebras: An Introduction to Subalgebras of C*-Algebras*, Pitman Research Notes in Math., **278**, Longman, London, 1993, MR 1204657 (94g:46001), Zbl 0917.46001.

[45] M. Rørdam, *Advances in the theory of unitary rank and regular approximation*, Ann. of Math., **128** (1988), 153–172, MR 0951510 (90c:46072), Zbl 0659.46052.

[46] J. Rovnyak, *Operator-valued analytic functions of constant norm*, Czechoslovak Math. J., **39** (1989), 165–168, MR 0983493 (90f:47019), Zbl 0686.47013.

[47] K. Ylinen, *A note on the compact elements of C^*-algebras*, Proc. Amer. Math. Soc., **35** (1972), 305–306, MR 0296716 (45 #5775), Zbl 0257.46085.

Received June 3, 2002 and revised November 19, 2003. This research was partially supported by a grant from ECU.

Department of Mathematics
East Carolina University
Greenville NC 27858
E-mail address: KatsoulisE@mail.ecu.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

GENERALIZED SKEW DERIVATIONS CHARACTERIZED BY ACTING ON ZERO PRODUCTS

Tsiu-Kwen Lee

Let A be a prime ring whose symmetric Martindale quotient ring contains a nontrivial idempotent. Generalized skew derivations of A are characterized by acting on zero products. Precisely, if $g, \delta\colon A \to A$ are additive maps such that $\sigma(x)g(y) + \delta(x)y = 0$ for all $x, y \in A$ with $xy = 0$, where σ is an automorphism of A, then both g and δ are characterized as specific generalized σ-derivations on a nonzero ideal of A.

1. Results

Let B be a ring with a subring A. An additive map $\delta\colon A \to B$ is called a derivation if $\delta(xy) = \delta(x)y + x\delta(y)$ for all $x, y \in A$. In a recent paper Jing, Lu and Li proved the following result [**6**, Theorem 6]:

Let $\mathcal{B}$ be a standard operator algebra in a Banach space X containing the identity operator I and let $\delta\colon \mathcal{B} \to \mathcal{B}$ be a linear map such that $\delta(AB) = \delta(A)B + A\delta(B)$ for any pair $A, B \in \mathcal{B}$ with $AB = 0$. Then $\delta(AB) = \delta(A)B + A\delta(B) - A\delta(I)B$ for all $A, B \in \mathcal{B}$. If in addition $\delta(I) = 0$, then δ is a derivation.

The result says that an additive map on a standard operator algebra is almost a derivation if it satisfies the expansion formula of derivations on pairs of elements with zero product. Since standard operator algebras involve many idempotents, from this point of view Chebotar, Ke and P.-H. Lee studied maps acting on zero products in the context of prime rings [**2**]. To state their results precisely we must first fix some notation.

Throughout, unless specially stated, A will denote a prime ring with center Z, extended centroid C and symmetric Martindale quotient ring Q. The maximal right and left quotient rings of A will be denoted by Q_{mr} and Q_{ml}, respectively. See [**1**] for details. Theorem 2 of [**2**] says this:

Let $\delta\colon A \to A$ be an additive map such that $\delta(x)y + x\delta(y) = 0$ for $x, y \in A$ with $xy = 0$. Suppose that Q contains a nontrivial idempotent e such that $eA \cup Ae \subseteq A$.

(I) *If $1 \in A$, then $\delta(xy) = \delta(x)y + x\delta(y) - \lambda xy$ for all $x, y \in A$, where $\lambda = \delta(1) \in Z$. In particular, if $\delta(1) = 0$, then δ is a derivation of A.*

(II) *If* $\deg A \geq 3$, *there exists* $\lambda \in C$ *such that* $\delta(xy) = \delta(x)y + x\delta(y) - \lambda xy$ *for all* $x, y \in A$.

Generalized derivations and σ*-derivations* (or *skew derivations*) are two natural generalizations of derivations, and are defined as follows. Let σ be an automorphism of A. An additive map $\delta\colon A \to Q_{ml}$ is called a σ*-derivation* if $\delta(xy) = \sigma(x)\delta(y) + \delta(x)y$ for all $x, y \in A$. Basic examples are derivations and $\sigma - 1$. Given $b \in A$, the map $\delta\colon x \in A \mapsto \sigma(x)b - bx$ obviously defines a σ-derivation, called the *inner* σ*-derivation* defined by b.

An additive mapping $g\colon A \to Q_{ml}$ is a *generalized derivation* if there exists a derivation $d\colon A \to Q_{ml}$ such that $g(xy) = xg(y) + d(x)y$ for all $x, y \in A$. As basic examples we mention derivations, generalized inner derivations (maps $x \mapsto ax + xb$ for $a, b \in A$) and left A-module mappings from A into itself. From this one easily sees that a map δ as in Theorem 2 of [**2**] (see bottom of previous page) is indeed a generalized derivation. In this paper we will generalize that theorem from a different point of view. We start with a definition of generalized skew derivations, generalizing both skew derivations and generalized derivations.

An additive map $g\colon A \to Q_{ml}$ is called a *generalized* σ*-derivation*, where σ be an automorphism of A, if there exists an additive map $\delta\colon A \to Q_{ml}$ such that $g(xy) = \sigma(x)g(y) + \delta(x)y$ for all $x, y \in A$. It is clear that δ is uniquely determined by g, which is called the associated additive map of g. It is easy to check that δ is always a σ-derivation (see [**10**]). We are now in a position to state our main result:

Theorem 1.1. *Let* A *be a prime ring with symmetric Martindale quotient ring* Q. *Suppose that* Q *contains a nontrivial idempotent and that* $g, \delta\colon A \to A$ *are additive maps. If* $\sigma(x)g(y) + \delta(x)y = 0$ *for all* $x, y \in A$ *with* $xy = 0$, *where* σ *is an automorphism of* A, *there exist a nonzero ideal* N *of* A *and a* σ*-derivation* $d\colon A \to Q$ *such that*

$$g(x) = d(x) + \sigma(x)b \quad \textit{and} \quad \delta(x) = d(x) + ax$$

for all $x \in N$, *where* $b \in Q_{ml}$ *and* $a \in Q_{mr}$. *In addition, we can take* $N = A$ *if* $eA \cup Ae \subseteq A$ *for some nontrivial idempotent* $e \in Q$.

The proof of the theorem depends on both the Lie structure of rings and the theory of functional identities, and will be given in the next section. As an immediate consequence of Theorem 1.1 we have the following generalization of Chebotar, Ke and P.-H. Lee's theorem [**2**, Theorem 2]:

Corollary 1.2. *Let* A *be a prime ring with extended centroid* C *and symmetric Martindale quotient ring* Q. *Suppose that* Q *contains a nontrivial idempotent. If* $\delta\colon A \to A$ *is an additive map such that* $x\delta(y) + \delta(x)y = 0$ *for* $x, y \in A$ *with* $xy = 0$, *then there exists a nonzero ideal* N *of* A *such that*

$\delta(x) = d(x) + \lambda x$ for all $x \in N$, where $\lambda \in C$ and $d\colon A \to Q$ is a derivation. In addition, we can take $N = A$ if $eA \cup Ae \subseteq A$ for some nontrivial idempotent $e \in Q$.

The following corollary gives Jing, Lu and Li's theorem [**6**, Theorem 6] in the context of prime rings:

Corollary 1.3. *Let A be a prime ring with extended centroid C and $c \in A$. Suppose that A possesses a nontrivial idempotent. If $\delta\colon A \to A$ is an additive map such that $x\delta(y) + \delta(x)y + xcy = 0$ for $x, y \in A$ with $xy = 0$, then there exist a derivation $d\colon A \to AC$ and $\mu \in C$ such that $\delta(x) = d(x) + (\mu - c)x$ for all $x \in A$.*

Proof. By assumption, we have $x(\delta(y)+cy)+\delta(x)y = 0$ for all $x, y \in A$ with $xy = 0$. In view of Theorem 1.1, there exist a derivation $d\colon A \to Q$, $a \in Q_{mr}$ and $\mu \in Q_{ml}$ such that $\delta(x) = d(x) + ax$ and $\delta(x) + cx = d(x) + x\mu$ for all $x \in A$. Choose a dense right ideal ρ and a dense left ideal λ of A such that $a\rho \subseteq A$ and $\lambda\mu \subseteq A$. Let $x \in \rho$, $z \in A$ and $y \in \lambda$. Then $xzy \in \rho A\lambda \subseteq \rho \cap \lambda$. Thus $(a + c)x, y\mu \in A$ and $((a + c)x)zy = xz(y\mu)$. By Martindale's Lemma [**11**], y and $y\mu$ are C-dependent for $y \in \lambda$. It is now easy to prove that $\mu \in C$. Thus $a = \mu - c$ follows and so $\delta(x) = d(x) + (\mu - c)x$ for all $x \in A$. In particular, we have $d(A) \subseteq AC$, proving the corollary.

The next application is to generalized polynomial identities. An additive map $f\colon A \to A$ is called an *elementary operator* if there exist finitely many $a_i, b_i \in AC$ such that $f(x) = \sum_i a_i x b_i$ for all $x \in A$.

Corollary 1.4. *Let A be a prime ring with extended centroid C. Suppose that its symmetric Martindale quotient ring Q contains a nontrivial idempotent. If f and g are two elementary operators of A satisfying $xf(y) + g(x)y = 0$ for $x, y \in A$ with $xy = 0$, then there exist $a, b, q \in AC$ and such that $f(y) = [q, y] + yb$ and $g(x) = [q, x] + ax$ for all $x, y \in A$.*

Proof. Since f, g are elementary operators, there exist finitely many a_i, b_i, $c_i, d_i \in AC$ such that $f(x) = \sum_i a_i x b_i$ and $g(x) = \sum_j c_j x d_j$ for all $x \in A$. In view of Theorem 1.1, there exist a nonzero ideal N of A, a derivation $d\colon A \to Q$ and elements $a \in Q_{mr}$, $b \in Q_{ml}$ such that

$$\textstyle\sum_i a_i y b_i = d(y) + yb \quad \text{and} \quad \sum_j c_j x d_j = d(x) + ax$$

for all $x, y \in N$. It is well-known that d can be uniquely extended to a derivation of Q_{ml} into Q_{ml} and $\sum_i a_i y b_i = d(y) + yb$ for all $y \in Q_{ml}$ (see, for instance, [**9**, Theorem 2]). In particular, we set $y = 1$, implying that $b = \sum_i a_i b_i \in AC$. An analogous argument proves $a \in AC$, so $d(AC) \subseteq AC$. Applying Kharchenko's Theorem ([**7**, Lemma 1] or [**8**, Theorem 2]), we conclude that d is X-inner; that is, there exists $q \in Q$ such that $d(y) = [q, y]$ for all $y \in A$. Thus $qy - yq = \sum_i a_i y b_i - yb$ for all $y \in A$. Applying

Martindale's Lemma [**11**], q lies in the C-linear span of the elements b_i's, b and 1. Thus $q \in AC + C$. Since, for $\beta \in C$, $[q + \beta, y] = [q, y]$ for all $y \in A$, we may take $q \in AC$, proving the corollary.

The following example shows that the existence of nontrivial idempotents in Q is essential to Theorem 1.1.

Example 1.5. Let A be a prime ring, not a domain, with center Z and let

$$\mathcal{M}_A = \{a \in A \mid xay = 0 \text{ whenever } x, y \in A \text{ with } xy = 0\}.$$

Suppose that $\mathcal{M}_A$ is noncentral. Choose an element $c \in \mathcal{M}_A \setminus Z$. Let $f, g\colon A \to A$ be additive maps defined by $f(x) = cxc$ and $g(x) = x$ for all $x \in A$. Then $xf(y) + g(x)y = 0$ for $x, y \in A$ with $xy = 0$. We claim that f, g cannot assume the forms given in Theorem 1.1. Indeed, suppose there exist a derivation $d\colon A \to Q$ and a nonzero ideal N of A such that

$$f(x) = d(x) + xb \quad \text{and} \quad g(x) = d(x) + ax$$

for all $x \in N$, where $a \in Q_{mr}$ and $b \in Q_{ml}$. Then $d(x) = (1 - a)x$ for all $x \in A$, implying $d = 0$. Thus $xb = cxc$ for all $x \in A$. Applying Martindale's Lemma [11], we see that $c \in C$, the extended centroid of A, and so $c \in Z$, a contradiction.

Such prime rings do exist. One example is due to Dubrovin [**3**]. Another is $A = K\{x, y\}/(x^2)$ [**5**, pp. 105–108], where $K\{x, y\}$ is the free algebra over a field K in two noncommuting indeterminates x and y. In this example, A is a prime ring and $Kx + xAx + (x^2)/(x^2)$ coincides with the set of all elements of A with square zero. Let $\bar{x} = x + (x^2) \in A$. Then $a\bar{x}b = 0$ whenever $a, b \in A$ with $ab = 0$. Thus $\bar{x}$ lies in $\mathcal{M}_A$ and is noncentral.

2. Proof of Theorem 1.1

Lemma 2.1. *Let I be a nonzero ideal of A and let $f\colon I \to Q_{ml}$ be a left I-module map. Then there exists $q \in Q_{ml}$ such that $f(x) = xq$ for all $x \in I$.*

Proof. Notice that $Q_{ml}I$ is a dense left ideal of Q_{ml}. We define the map $\widetilde{f}\colon Q_{ml}I \to Q_{ml}$ by $\sum_i x_i a_i \mapsto \sum_i x_i f(a_i)$, where $x_i \in Q_{ml}, a_i \in I$. Then $\widetilde{f}$ is well-defined. Indeed, let $\sum_i x_i a_i = 0$, where $x_i \in Q_{ml}, a_i \in I$. Choose a dense left ideal J of A such that $Jx_i \subseteq I$ for each i. Then, for $y \in J$, we have

$$0 = f\big(y \textstyle\sum_i x_i a_i\big) = f\big(\sum_i (yx_i)a_i\big) = \sum_i yx_i f(a_i) = y\big(\sum_i x_i f(a_i)\big),$$

implying $\sum_i x_i f(a_i) = 0$. Thus $\widetilde{f}$ is well-defined. It is clear that $\widetilde{f}$ is a left Q_{ml}-module map extending f. We remark that the maximal left quotient ring of Q_{ml} coincides with itself. Thus there exists a $q \in Q_{ml}$ such that $\widetilde{f}(z) = zq$ for all $z \in Q_{ml}I$. In particular, $f(x) = xq$ for all $x \in I$. This proves the lemma.

Lemma 2.2. *Let I be a nonzero ideal of A and let $f, g\colon I \to Q_{ml}$ be two additive maps. Suppose that $f(x)y + \sigma(x)g(y) = 0$ for all $x, y \in I$, where σ is an automorphism of A. Then there exists $a \in Q_{ml}$ such that $f(x) = \sigma(x)a$ and $g(y) = -ay$ for all $x, y \in I$.*

Proof. Let $x, y, z \in I$. By assumption,

$$f(zx)y + \sigma(zx)g(y) = 0 \quad \text{and} \quad \sigma(z)(f(x)y + \sigma(x)g(y)) = 0.$$

Thus $(f(zx) - \sigma(z)f(x))y = 0$ and so $f(zx) = \sigma(z)f(x)$ since A is prime. Note that σ can be uniquely extended to an automorphism of Q_{ml}. Consider the map $\phi\colon I \to Q_{ml}$ defined by $\phi(x) = \sigma^{-1}(f(x))$ for $x \in I$. Then ϕ is a left I-module map. By Lemma 2.1, there exists $b \in Q_{ml}$ such that $\phi(x) = xb$ for all $x \in I$. That is, $f(x) = \sigma(x)a$ for all $x \in I$, where $a = \sigma(b) \in Q_{ml}$. It is clear that $g(y) = -ay$ for all $y \in I$. This proves the lemma.

Although the next lemma has a more general version, for our purposes we need only the following special form:

Lemma 2.3. *Let $d\colon M \to Q$ be a σ-derivation, where M is a nonzero ideal of A and σ is an automorphism of A. Assume that there exists a nonzero ideal J of A such that $d(m)J + Jd(m) \subseteq A$ for all $m \in M$. Then d can be uniquely extended to a σ-derivation from A into Q.*

Proof. Replacing J by $M \cap J$, we may assume from the start that $J \subseteq M$. Let $a \in A$. Define a map $\psi_a\colon MJ \to A$ by the rule

$$\psi_a\big(\textstyle\sum_i m_i x_i\big) = \sum_i d(am_i)x_i - \sum_i \sigma(a)d(m_i)x_i \in A,$$

where $m_i \in M$ and $x_i \in J$. We claim that ψ_a is well-defined. Indeed, $\sum_i m_i x_i = 0$ implies $\sum_i (am_i)x_i = 0$, so

$$\begin{aligned} 0 = \textstyle\sum_i d((am_i)x_i) &= \textstyle\sum_i (\sigma(am_i)d(x_i) + d(am_i)x_i) \\ &= \textstyle\sum_i d(am_i)x_i - \sum_i \sigma(a)d(m_i)x_i, \end{aligned}$$

since

$$\begin{aligned} \textstyle\sum_i \sigma(am_i)d(x_i) &= \textstyle\sum_i \sigma(a)\sigma(m_i)d(x_i) \\ &= \textstyle\sum_i \sigma(a)(d(m_i x_i) - d(m_i)x_i) \\ &= \sigma(a)d\left(\textstyle\sum_i m_i x_i\right) - \sum_i \sigma(a)d(m_i)x_i \\ &= -\textstyle\sum_i \sigma(a)d(m_i)x_i. \end{aligned}$$

The claim is proved. It is clear that ψ_a is a right A-module map. Thus ψ_a is defined by an element $\widetilde{d}(a) \in Q_r$, the right Martindale quotient ring of A. That is, $\widetilde{d}\colon A \to Q_r$ has the following property: $\widetilde{d}(a)m = d(am) - \sigma(a)d(m)$ for $a \in A$ and $m \in M$. Let $a, b \in A$ and $m \in M$. Then $\widetilde{d}(ab)m = d(abm) - \sigma(ab)d(m)$. On the other hand, since $bm \in M$ we have

$$\begin{aligned} (\sigma(a)\widetilde{d}(b) + \widetilde{d}(a)b)m &= \sigma(a)(d(bm) - \sigma(b)d(m)) + d(abm) - \sigma(a)d(bm) \\ &= d(abm) - \sigma(ab)d(m). \end{aligned}$$

Thus $\widetilde{d}(ab) = \sigma(a)\widetilde{d}(b) + \widetilde{d}(a)b$, proving that $\widetilde{d}$ is a σ-derivation. Clearly, $\widetilde{d}(m) = d(m)$ for all $m \in M$, so $\widetilde{d}$ is an extension of d. Notice that $J\sigma(M)$ is a nonzero ideal of A. Let $a \in A$ and $m \in M$. Then $d(ma) = \sigma(m)\widetilde{d}(a)+d(m)a$ and so

$$J\sigma(M)\widetilde{d}(a) \subseteq Jd(M) + Jd(M)A \subseteq A,$$

implying that $\widetilde{d}(a) \in Q$. Hence, $\widetilde{d}\colon A \to Q$ and the lemma is proved.

We are now ready to prove the main theorem stated in §1.

Theorem 1.1. *Let A be a prime ring with symmetric Martindale quotient ring Q. Suppose that Q contains a nontrivial idempotent and that $g, \delta\colon A \to A$ are additive maps. If $\sigma(x)g(y) + \delta(x)y = 0$ for all $x, y \in A$ with $xy = 0$, where σ is an automorphism of A, there exist a nonzero ideal N of A and a σ-derivation $d\colon A \to Q$ such that*

$$g(x) = d(x) + \sigma(x)b \quad \text{and} \quad \delta(x) = d(x) + ax$$

for all $x \in N$, where $b \in Q_{ml}$ and $a \in Q_{mr}$. In addition, we can take $N = A$ if $eA \cup Ae \subseteq A$ for some nontrivial idempotent $e \in Q$.

Proof. Let e be a nontrivial idempotent of Q. Choose a nonzero ideal I of A such that $Ie+eI \subseteq A$. We consider the additive subgroup E of Q generated by the set $\{f \in Q \mid If + fI \subseteq A \text{ and } f^2 = f\}$. Then $e \in E$. We claim that $0 \neq I^2[E, E]I^2 \subseteq E + E^2$.

We follow Herstein's argument [**4**, Proof of Lemma 1.3]. Let $f \in E$ and $x \in I$. Then $f + fx(1 - f)$, $f + (1 - f)xf \in E$ and so

$$[f, x] = (f + fx(1 - f)) - (f + (1 - f)xf) \in E.$$

Thus $[E, I] \subseteq E$. Indeed, $[E, E]I^2 \subseteq [E, EI^2] + E[E, I^2] \subseteq E + E^2$, and so

$$I^2[E, E]I^2 \subseteq [I^2, [E, E]I^2] + [E, E]I^4 \subseteq [I^2, E + E^2] + E + E^2.$$

Since $[I^2, E + E^2] \subseteq E + E^2$, we see that $I^2[E, E]I^2 \subseteq E + E^2$. Finally, $0 \neq [e, e + eI^2(1 - e)] \subseteq [E, E]$, proving our claim.

Let $x, y \in I$ and $f = f^2 \in E$. Then xf, $(1-f)y$, $x(1-f)$, fy belong to A. By assumption,

$$\begin{aligned} &\delta(xf)(1 - f)y + \sigma(xf)g((1 - f)y) = 0 \qquad \text{and} \\ &\delta(x(1 - f))fy + \sigma(x(1 - f))g(fy) = 0. \end{aligned} \tag{1}$$

Solve the two equations by Lemma 2.2: there exist two unique elements $u, v \in Q_{ml}$, depending on f, such that

$$\begin{aligned} &\delta(xf)(1 - f) = \sigma(x)u, \quad \sigma(f)g((1 - f)y) = -uy \qquad \text{and} \\ &\delta(x(1 - f))f = \sigma(x)v, \quad \sigma(1 - f)g(fy) = -vy. \end{aligned} \tag{2}$$

Thus, by (2), we see that

$$\delta(x)f - \delta(xf) = \sigma(x)(v - u) \ \text{ and } \ \sigma(f)g(y) - g(fy) = (v - u)y. \tag{3}$$

By (3) and the definition of E, there exists an additive map $d\colon E \to Q_{ml}$ such that

$$\delta(x)m - \delta(xm) = -\sigma(x)d(m) \quad \text{and} \quad \sigma(m)g(y) - g(my) = -d(m)y \tag{4}$$

for all $x, y \in I$ and all $m \in E$. Let $m_1, m_2 \in E$; then

$$\delta(x)m_2 - \delta(xm_2) = -\sigma(x)d(m_2) \text{ and } \delta(x)m_1 - \delta(xm_1) = -\sigma(x)d(m_1) \tag{5}$$

for all $x \in I$. Let $x \in I^2$; then $xm_1 \in I$. By (5) we have

$$\begin{aligned} \delta(xm_1)m_2 - \delta(xm_1m_2) &= -\sigma(xm_1)d(m_2) \qquad \text{and} \\ -\delta(xm_1)m_2 + \delta(x)m_1m_2 &= -\sigma(x)d(m_1)m_2. \end{aligned}$$

Thus

$$\delta(x)m_1m_2 - \delta(x(m_1m_2)) = -\sigma(x)(\sigma(m_1)d(m_2) + d(m_1)m_2). \tag{6}$$

On the other hand, let $y \in I^2$; then $m_2y \in I$. By (4) we have

$$\begin{aligned} \sigma(m_1)g(m_2y) - g(m_1m_2y) &= -d(m_1)m_2y, \\ \sigma(m_1)\sigma(m_2)g(y) - \sigma(m_1)g(m_2y) &= -\sigma(m_1)d(m_2)y, \end{aligned}$$

implying that

$$\sigma(m_1m_2)g(y) - g((m_1m_2)y) = -(d(m_1)m_2 + \sigma(m_1)d(m_2))y. \tag{7}$$

This means that d can be extended to $E + E^2$ in such a way that

$$\delta(x)m - \delta(xm) = -\sigma(x)d(m) \quad \text{and} \quad \sigma(m)g(y) - g(my) = -d(m)y$$

for all $x, y \in I^2$ and all $m \in E + E^2$. Moreover, $d(m_1m_2) = \sigma(m_1)d(m_2) + d(m_1)m_2$ for all $m_1, m_2 \in E$. Repeating the argument above, we can extend d to $E + E^2 + E^3 + E^4$ in such a way that

$$\delta(x)m - \delta(xm) = -\sigma(x)d(m) \quad \text{and} \quad \sigma(m)g(y) - g(my) = -d(m)y \tag{8}$$

for all $x \in I^4$ and all $m \in E + E^2 + E^3 + E^4$. Moreover,

$$d(uv) = \sigma(u)d(v) + d(u)v \tag{9}$$

for all $u, v \in E + E^2$. Let $M = I^2[E, E]I^2 \neq 0$. Then M is a nonzero ideal of A contained in $E + E^2$.

Let $m \in M$. By (8) we see that $\sigma(I^4)d(m) \subseteq A$ and $d(m)I^4 \subseteq A$. Thus $d(m) \in Q$ follows, showing that $d\colon M \to Q$ is a σ-derivation satisfying

$$\delta(x)m - \delta(xm) = -\sigma(x)d(m) \quad \text{and} \quad \sigma(m)g(y) - g(my) = -d(m)y \tag{10}$$

for all $x, y \in I^4$ and all $m \in M$. Set $J = \sigma(I^4) \cap I^4$. Then J is a nonzero ideal of A and, moreover, $d(m)J + Jd(m) \subseteq A$ for all $m \in M$. In view of

Lemma 2.3, d can be uniquely extended to a σ-derivation from A into Q. Thus we can rewrite (10) as

$$\begin{aligned} \delta(xm) - d(xm) &= (\delta(x) - d(x))m, \\ \sigma(m)(g(y) - d(y)) &= g(my) - d(my) \end{aligned} \tag{11}$$

for all $x, y \in I^4$ and all $m \in M$. Consider the map $\phi\colon I^4 \to Q$ defined by

$$x \in I^4 \mapsto \phi(x) = \delta(x) - d(x).$$

By (11), ϕ is a right M-module map. Thus it is a right A-module map by the primeness of A. By Lemma 2.1, there exists $a \in Q_{mr}$ such that $\phi(x) = ax$ and so $\delta(x) = d(x) + ax$ for all $x \in I^4$. By an analogous argument, there exists $b \in Q_{ml}$ such that $g(x) = d(x) + \sigma(x)b$ for all $x \in I^4$. We are now done by setting $N = I^4$.

Suppose, in addition, that $eA \cup Ae \subseteq A$ for some nontrivial idempotent $e \in Q$. Then $I = A$ in our construction. Moreover, $EA + AE \subseteq A$. Therefore, (11) remains true for all $x \in A$. So the map $\phi\colon I^4 \to Q$ can be replaced by $\phi\colon A \to Q$. Now our conclusion holds trivially. This proves the theorem.

Acknowledgments. The author would like to thank the referee for her/his useful comments. The research was supported in part by a grant from NSC of R.O.C. (Taiwan).

References

[1] K.I. Beidar, W.S. Martindale III and A.V. Mikhalev, *Rings with Generalized Identities*, Monographs and Textbooks in Pure and Applied Mathematics, **196**, Marcel Dekker, New York, 1996, MR 1368853 (97g:16035), Zbl 0847.16001.

[2] M.A. Chebotar, W.-F. Ke and P.-H. Lee, *Maps characterized by actions on zero products*, Pacific J. Math., **216** (2004), 217–228.

[3] N.I. Dubrovin, *An example of a chain primitive ring with nilpotent elements* (Russian), Mat. Sb. (N.S.), **120**(**162**) (1983), no. 3, 441–447 , MR 0691988 (84f:16012), Zbl 0543.16003.

[4] I.N. Herstein, *Topics in Ring Theory*, University of Chicago Press, Chicago, 1969, MR 0271135 (42 #6018), Zbl 0232.16001.

[5] I.N. Herstein, *Rings with Involution*, Univ. of Chicago Press, Chicago, 1976, MR 0442017 (56 #406), Zbl 0343.16011.

[6] W. Jing, S. Lu and P. Li, *Characterisations of derivations on some operator algebras*, Bull. Austral. Math. Soc., **66** (2002), 227–232, MR 1932346 (2003f:47059), Zbl 1035.47019.

[7] V.K. Kharchenko, *Differential identities of prime rings*, Algebra i Logika, **17** (1978), 220–238; Engl. Transl., Algebra and Logic, **17** (1978), 154–168, MR 0541758 (81f:16025), Zbl 0423.16011.

[8] V.K. Kharchenko, *Differential identities of semiprime rings*, Algebra i Logika, **18** (1979), 86–119; Engl. Transl., Algebra and Logic, **18** (1979), 58–80, MR 0566776 (81f:16052), Zbl 0464.16027.

[9] T.-K. Lee, *Semiprime rings with differential identities*, Bull. Inst. Math. Acad. Sinica, **20** (1992), 27–38, MR 1166215 (93e:16039), Zbl 0769.16017.

[10] T.-K. Lee and K.-S. Liu, *Generalized skew derivations with algebraic values of bounded degree*, preprint.

[11] W.S. Martindale III, *Prime rings satisfying a generalized polynomial identity*, J. Algebra, **12** (1969), 576–584, MR 0238897 (39 #257), Zbl 0175.03102.

Received June 19, 2003 and revised August 21, 2003.

Department of Mathematics
National Taiwan University
Taipei 106
Taiwan
E-mail address: tklee@math.ntu.edu.tw

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

AN INVERSE PROBLEM FOR THE TRANSPORT EQUATION IN THE PRESENCE OF A RIEMANNIAN METRIC

Stephen R. McDowall

The stationary linear transport equation models the scattering and absorption of a low-density beam of neutrons as it passes through a body. In Euclidean space, to a first approximation, particles travel in straight lines. Here we study the analogous transport equation for particles in an ambient field described by a Riemannian metric where, again to first approximation, particles follow geodesics of the metric. We consider the problem of determining the scattering and absorption coefficients from knowledge of the albedo operator on the boundary of the domain. Under certain restrictions, the albedo operator is shown to determine the geodesic ray transform of the absorption coefficient; for "simple" manifolds this transform is invertible and so the coefficient itself is determined. In dimensions 3 or greater, we show that one may then obtain the collision (or scattering) kernel.

1. Introduction

The stationary linear transport equation models the time-independent scattering of a low-density beam of particles off a higher-density material. The term "linear" refers to the fact that the equation models only scattering of particles from the material and assumes that the density of particles is low enough that particle-to-particle interaction may be neglected. If (x, v) is a point in phase space we denote by $f(x, v)$ the density of particles at position x with velocity v. While f is strictly speaking a density, for large numbers of particles it is reasonable to represent f as an L^1 function. In free-space, f satisfies the transport equation

$$-v \cdot \nabla_x f(x, v) - \sigma_a(x, v) f(x, v) + \int_V k(x, v', v) f(x, v')\, dv' = 0. \tag{1}$$

The first term describes the straight-line motion of a particle that does not interact with the material. The second term represents the loss of a particle at (x, v) due to scattering to another velocity or due to absorption, quantified by the function $\sigma_a(x, v)$. The final term accounts for the production of a particle at (x, v) due to scattering from other directions; the kernel $k(x, v', v)$

represents the probability of a particle at (x, v') scattering to (x, v). The reader is referred to [**RS**] for a detailed explanation.

We are concerned here with the situation of particles moving in an ambient field represented by a Riemannian metric. The principal departure from (1) is that in the absence of interaction, a particle will follow the geodesics of the metric. We shall study the problem of determining the absorption coefficient $\sigma_a(x, v)$ and the collision kernel $k(x, v', v)$ from knowledge of the positions and velocities of particles entering and leaving a bounded body.

Let $M \subset \mathbb{R}^n$, $n \geq 2$, be a bounded domain with C^∞ boundary; let g be a Riemannian metric on M. We shall put restrictions on the metric g in due course. Define the incoming and outgoing bundles on ∂M as

$$\Gamma_\pm = \big\{(x, v) \in TM \mid x \in \partial M,\ \pm\langle v, \nu\rangle_{g_x} > 0\big\},$$

where ν is the outward unit normal vector to ∂M and $\langle \cdot, \cdot \rangle_{g_x}$ is the inner product with respect to the metric g. We shall also use $\| \cdot \|_{g_x}$ to denote the norm with respect to g. If $(x, v) \in TM$ we shall denote by $\gamma_{(x,v)}(t)$ the geodesic satisfying $\gamma_{(x,v)}(0) = x$ and $\dot\gamma_{(x,v)}(0) = v$; we introduce the compressed notation

$$\vec\gamma_{(x,v)}(t) = (\gamma_{(x,v)}(t), \dot\gamma_{(x,v)}(t)).$$

For $v \neq 0$, define the time-to-boundary functions $\tau_\pm : TM \to \mathbb{R}^+$ by

$$\tau_\pm(x, v) = \min\big\{t > 0 \mid \gamma_{(x,v)}(\pm t) \in \partial M\big\}$$

and set $\tau(x, v) = \tau_-(x, v) + \tau_+(x, v)$. While in general $\tau_\pm$ might be infinite, we will place restrictions on the metric that ensure that $\tau_\pm$ are well-defined and finite.

With these preparations we are able to generalize Equation (1) and state the results of the paper. Denote by $\mathcal{D}$ the derivative along the geodesic flow,

$$\mathcal{D}f(x, v) = \frac{\partial}{\partial t}\Big|_{t=0} f(\gamma_{(x,v)}(t), \dot\gamma_{(x,v)}(t)).$$

If $(x^i, y^i)_{i=1}^n$ are local coordinates for TM with the (y^i) with respect to the natural basis $\big(\frac{\partial}{\partial x^i}\big)$, we have in these coordinates

$$\mathcal{D}f = \frac{\partial f}{\partial x^i} y^i + \frac{\partial f}{\partial y^i}(-y^j y^k \Gamma^i_{jk}),$$

where Γ^i_{jk} are the Christoffel symbols of the Levi-Civita connection of g. The transport equation (1) is replaced in our setting by

$$-\mathcal{D}f(x, v) - \sigma_a(x, v) f(x, v) + \int_{T_xM} k(x, v', v) f(x, v')\, dv'_x = 0. \tag{2}$$

The measure dv_x is the Euclidean volume form on T_xM determined by the metric g_x at x (see Definition 2.1). Given a function $f_-(x, v)$ on Γ_- let f be

the solution to (2) with boundary condition $f\big|_{\Gamma_-} = f_-$, assuming it exists. One may then define the *albedo operator*

$$\mathcal{A} : f_- \mapsto f\big|_{\Gamma_+}.$$

The inverse problem addressed here is the unique determination of $\sigma_a(x, v)$ and $k(x, v', v)$ for all $x \in M$ and $v, v' \in T_xM$, from the knowledge of $\mathcal{A}$. The main results are the content of Theorems 4.2 and 4.4, which state that under the assumption of simplicity of the metric (see below) and *a priori* assumptions (3), (4) on the coefficients themselves, we can determine σ_a in dimensions $n \geq 2$ and k in dimensions $n \geq 3$.

We remark that if we assume that $k(x, v', v) = 0$ for all $\|v'\|_{g_x} \neq \|v\|_{g_x}$ — that is, that all scattering occurrences preserve *speed* — then we may consider the problem on the unit sphere bundle ΩM of M. We replace $\Gamma_\pm$ by their equivalents restricted to unit tangent vectors, and the integration in (2) is taken over $\Omega_x M$, the unit sphere in T_xM. The analysis and results of this paper remain valid in this setting. Indeed, under this assumption on k we may consider $\Gamma_\pm$ defined to include tangent vectors of lengths $0 < a \leq \|v\| \leq b < \infty$ and in (2) integrate over the corresponding annular region in T_xM. Once again the results remain valid in this setting.

When the ambient metric is Euclidean, the inverse problem for (1) was considered in [**CS2**] and we shall approach the problem here in the same manner. In [**CS2**] it is shown that the most singular part of the distributional kernel of the albedo operator determines the x-ray transform of the absorption coefficient. The next most singular part determines the collision kernel k in dimensions 3 and greater. Here we shall show that the first term determines the *geodesic ray transform* of σ_a, namely the integrals of σ_a along geodesics of g. In order to recover σ_a we make the restrictive assumption that the metric g be "simple." This means that M is strictly convex with respect to g and that for any $x \in \overline{M}$ the exponential map $\operatorname{Exp}_x : \operatorname{Exp}_x^{-1}(\overline{M}) \to \overline{M}$ is a diffeomorphism. This assumption, together with the full set of geodesics joining boundary points, ensures that the ray transform is invertible (see [**BG**] and [**M**], and [**Sh1**] for an extensive treatment of the ray transform).

In the Riemannian setting, an inverse source problem for the stationary transport equation is addressed in [**Sh2**] (see also [**Sh1**]), where $k(x, v, v') = k(x, \langle v, v' \rangle)$ is assumed to depend on the angle between v and v' and the object of interest is the reconstruction of an isotropic source term. In [**F**] the (time-dependent) radiative transfer equation is derived for a medium with spatially varying refractive index and with scattering kernel k independent of position. Such a refractive index is represented by a Riemannian metric.

More is known when the metric is assumed to be Euclidean. The time-dependent inverse problem was treated in [**CS1**] and the stationary case in [**CS2**]. Stability estimates based on this work were obtained under certain restrictions in [**W**].

The stationary transport equation is used to model the absorption and scattering of near-infrared light; the determination of the absorption and scattering properties of a medium from the measurement of the response to such transmitted light is known as optical tomography and has been applied to the problems of medical imaging ([**A**]). For two-dimensional domains, recent works include [**T2**], where a homogeneous collision kernel k that is a function of two independent directions was shown to be uniquely determined; [**SU**], in which the assumption on homogeneity is dropped and k is assumed to be small relative to σ_a, with an explicit constant given; and [**T1**], where the smallness is removed in the case of weakly anisotropic scattering. In [**SU**] the authors prove also a stability estimate; further stability results may be found in [**R**]. Note that some kind of smallness does need to be assumed on k, to ensure that the production rate is in some sense less than the absorption rate, thus keeping the energy of the system bounded.

This paper is organized as follows: in Section 2 we state our assumptions precisely, prove solvability of the forward problem given these assumptions, and demonstrate well-definedness of the boundary albedo operator. In Section 3 we construct distribution solutions to (2) with delta-type boundary conditions. This facilitates determination of the distribution kernel of the albedo operator as the sum of three terms of differing singularity strengths. In Section 4 we prove that it is possible to extract from the kernel each of the singular terms and that these determine the absorption coefficient (in all dimensions) and the collision kernel (in dimensions 3 and greater).

2. The forward problem and the albedo operator

We begin by defining the volume form on TM:

Definition 2.1. The *Liouville volume form* is the canonical $2n$-form defined on TM that is preserved under the geodesic flow of g. It is the product of the Riemannian volume form $d\omega(x)$ on the manifold M and the Euclidean volume form dv_x defined in the tangent space T_xM by the metric g_x at $x \in M$ (see [**KH**], for example).

The sense in which the definition above holds is the following: given an orthonormal basis $Y_1, \ldots, Y_n$ for T_xM, extend it to an adapted basis of vector fields in a neighborhood of x by parallel transport with respect to the Levi-Civita connection of g. If dy^j are the forms dual to Y_j, the Liouville form is given by $\sqrt{\det g}\, dx^1 \wedge \cdots \wedge dx^n \wedge dy^1 \wedge \cdots \wedge dy^n$. For a fixed chart one can in principle express this form in terms of the natural basis $\frac{\partial}{\partial x^1}, \cdots \frac{\partial}{\partial x^n}$, and to do so one must solve the system of differential equations describing the evolution of the Y_j in terms of the $\frac{\partial}{\partial x^j}$ as the basepoint x for T_xM varies. The interested reader can consult Equations (6) and (7) on page 47 of [**H**]. It is not possible to express the Liouville form at an arbitrary point in a chart

without reference to a basepoint. We set $L^1(TM) = L^1(TM, dv_x\, d\omega(x))$ and $L^1(T_xM) = L^1(T_xM, dv_x)$.

We assume that the pair (σ_a, k) is admissible (see [**RS**]), namely that

$$\begin{cases} 0 \le \sigma_a \in L^\infty(TM), \\ 0 \le k(x, v', \cdot) \in L^1(T_xM) \text{ for a.e. } (x, v') \in TM, \\ \sigma_p(x, v') = \int_{T_xM} k(x, v', v)\, dv_x \in L^\infty(TM). \end{cases} \tag{3}$$

Notice that the operators $f \mapsto \sigma_a f$ and $f \mapsto \int_{T_xM} k(x, v', v) f(x, v')\, dv'_x$ are bounded on $L^1(TM)$.

Equation (2) may not be uniquely solvable, so we shall make the following subcriticality assumptions (see [**RS**]), which ensure that the problem is well-posed:

$$\|\tau\sigma_a\|_{L^\infty(TM)} < \infty \qquad \text{and} \qquad \|\tau\sigma_p\|_{L^\infty(TM)} < 1. \tag{4}$$

Furthermore, even in the Euclidean setting, without further restriction on σ_a the albedo operator does not uniquely determine σ_a (see [**CS2**]). To remove this lack of uniqueness, we assume that σ_a depends only on the *speed* $\|v\|_{g_x}$, not on the direction:

$$\sigma_a(x, v) = \sigma_a(x, \|v\|_{g_x}). \tag{5}$$

In order to pose the boundary value problem we must specify the volume form on $\{(x, v) \mid x \in \partial M,\ v \in T_xM\}$, in particular on $\Gamma_\pm$. If $(x, v) \in TM$, define $(x', v') \in \Gamma_-$ and $t \ge 0$ by

$$x' = \gamma_{(x,v)}(-\tau_-(x, v)), \qquad v' = \dot\gamma_{(x,v)}(-\tau_-(x, v)), \qquad t = \tau_-(x, v),$$

and define $F : \Gamma_- \times \mathbb{R} \to TM$ by

$$F(x', v', t) = (\gamma_{(x',v')}(t), \dot\gamma_{(x',v')}(t)).$$

Define the product measure $d\mu(x', v')\, dt$ on $\Gamma_- \times \mathbb{R}$ by

$$d\mu(x', v')\, dt = F^*(dv_x\, d\omega(x)),$$

where F^* is the pullback map defined by F. This measure is similarly defined on Γ_+. We denote $L^1(\Gamma_\pm) = L^1(\Gamma_\pm, d\mu(x', v'))$. The following lemma is immediate.

Lemma 2.2. *If $f \in L^1(TM)$ then*

$$\int_M \int_{T_xM} f(x, v) dv_x d\omega = \int_{\Gamma_\pm} \int_0^{\tau_\mp(x',v')} f(\gamma_{(x',v')}(t), \dot\gamma_{(x',v')}(t))\, dt\, d\mu(x', v').$$

We define the function space $\mathcal{W}$ in which we shall prove solvability of (2) for boundary functions $f_- \in L^1(\Gamma_-)$. We shall see below that functions f in $\mathcal{W}$ have a well-defined trace in $L^1(\Gamma_+)$. Let

$$\mathcal{W} = \{f \mid \mathcal{D}f \in L^1(TM),\ \tau^{-1} f \in L^1(TM)\},$$

with norm

$$\|f\|_{\mathcal{W}} = \|\mathcal{D}f\|_{L^1(TM)} + \|\tau^{-1}f\|_{L^1(TM)}.$$

The following trace theorem is proven in [**CS2**] for the Euclidean metric and we repeat a sketch of the proof in our setting.

Theorem 2.3. *If $f(x,v) \in \mathcal{W}$ then*

$$\|f|_{\Gamma_\pm}\|_{L^1(\Gamma_\pm, d\mu)} \leq \|\mathcal{D}f\|_{L^1(TM)} + \|\tau^{-1}f\|_{L^1(TM)}.$$

Proof. Observe that if $h, h' \in L^1([0,a])$, for some $a > 0$, then

$$|h(0)| \leq \|h'\|_{L^1([0,a])} + \frac{1}{a}\|h\|_{L^1([0,a])}.$$

Let f be as in the hypothesis of the theorem and define

$$h(t,x',v') = f(\vec{\gamma}_{(x',v')}(t)), \quad (x',v') \in \Gamma_-, \quad 0 \leq t \leq \tau_+(x',v').$$

Then

$$\begin{aligned}
\mathcal{D}f(\vec{\gamma}_{(x',v')}(t)) &= \frac{\partial}{\partial s}\Big|_{s=0} f(\vec{\gamma}_{\vec{\gamma}_{(x',v')}(t))}(s)) \\
&= \frac{\partial}{\partial s}\Big|_{s=0} f(\vec{\gamma}_{(x',v')}(t+s)) \\
&= \partial_t h(t,x',v'),
\end{aligned}$$

so $\partial_t h \in L^1(TM)$; therefore, by Fubini's theorem, $\partial_t h \in L^1\big([0,\tau_+(x',v')]\big)$ for a.e. (x',v'). Next,

$$\begin{aligned}
\int_0^{\tau_+(x',v')} |h(s,x',v')|\,ds &\leq \int_0^{\tau_+(x',v')} \int_0^s |\partial_t h(t,x',v')|\,dt\,ds \\
&\leq \tau_+(x',v')\|\partial_t h\|_{L^1([0,\tau_+(x',v')])},
\end{aligned}$$

so that $h(t,x',v')$ is also in $L^1\big([0,\tau_+(x',v')]\big)$ for a.e. (x',v'). Thus

$$\begin{aligned}
|f(x',v')| &= |f(\vec{\gamma}_{(x',v')}(0)| \\
&\leq \int_0^{\tau_+(x',v')} |\mathcal{D}f(\vec{\gamma}_{(x',v')}(t))|\,dt \\
&\qquad + \frac{1}{\tau_+(x',v')} \int_0^{\tau_+(x',v')} |f(\vec{\gamma}_{(x',v')}(t))|\,dt.
\end{aligned}$$

Integrating this inequality over Γ_- gives the result. The proof for Γ_+ is similar. □

Theorem 2.3 gives that the trace operator, $f \mapsto f|_{\Gamma_\pm}$, is continuous from $\mathcal{W}$ into $L^1(\Gamma_\pm, d\mu)$.

We proceed now to solve the boundary value problem (2) with $f|_{\Gamma_-} = f_- \in L^1(\Gamma_-, d\mu)$ and shall do so by reformulating the problem as an integral equation. First,

$$\begin{aligned} \mathcal{D}f(x,v) &= g(x,v) \\ f|_{\Gamma_-} &= 0 \end{aligned}$$

has solution

$$f(x,v) = \int_0^{\tau_-(x,v)} g(\vec{\gamma}_{(x,v)}(s - \tau_-(x,v)))\, ds.$$

To see this, we must calculate the seemingly intractable

$$\mathcal{D}f(x,v) = \frac{\partial}{\partial t}\Big|_{t=0} \int_0^{\tau_-(\vec{\gamma}_{(x,v)}(t))} g\big(\vec{\gamma}_{(\vec{\gamma}_{(x,v)}(t))}\big(s - \tau_-(\vec{\gamma}_{(x,v)}(t))\big)\big)\, ds.$$

This simplifies considerably when one observes that

$$\vec{\gamma}_{(\vec{\gamma}_{(x,v)}(t))}(s) = \vec{\gamma}_{(x,v)}(s+t) \tag{6}$$

and

$$\tau_-(\vec{\gamma}_{(x,v)}(t)) = t + \tau_-(x,v). \tag{7}$$

Thus

$$\begin{aligned} \mathcal{D}f &= \frac{\partial}{\partial t}\Big|_{t=0} \int_0^{t+\tau_-(x,v)} g\big(\vec{\gamma}_{(x,v)}(s - \tau_-(x,v))\big)\, ds \\ &= g(\vec{\gamma}_{(x,v)}(0)) = g(x,v). \end{aligned}$$

Lemma 2.4. *Let f_- be a function on Γ_-. Then*

$$\begin{aligned} \mathcal{D}f(x,v) + \sigma_a(x,v) f(x,v) &= 0 \text{ in } TM \\ f|_{\Gamma_-} &= f_- \end{aligned}$$

has solution Jf_-, where

$$Jf_-(x,v) = E(x,v,0,-\tau_-(x,v)) f_-\big(\vec{\gamma}_{(x,v)}(-\tau_-(x,v))\big),$$

and where

$$E(x,v,s,t) = \exp\left(\int_s^t \sigma_a(\vec{\gamma}_{(x,v)}(p))\, dp\right).$$

Proof. We compute

$$\begin{aligned} \mathcal{D}E(x,v,0,-\tau_-(x,v)) &= \frac{\partial}{\partial t}\Big|_{t=0} E\big(\vec{\gamma}_{(x,v)}(t), 0, -\tau_-(\vec{\gamma}_{(x,v)}(t))\big) \\ &= \frac{\partial}{\partial t}\Big|_{t=0} \exp\left(\int_0^{-t-\tau_-(x,v)} \sigma_a(\vec{\gamma}_{(x,v)}(p+t)\, dp\right) \\ &= -E(x,v,0,-\tau_-(x,v))\, \sigma_a(x,v) \end{aligned}$$

and

$$\begin{aligned}\mathcal{D}f_-(\vec{\gamma}_{(x,v)}(-\tau_-(x,v))) &= \frac{\partial}{\partial t}\Big|_{t=0} f_-\big(\vec{\gamma}_{(\vec{\gamma}_{(x,v)}(t))}\big(-\tau_-(\vec{\gamma}_{(x,v)}(t))\big)\big)\\ &= \frac{\partial}{\partial t}\Big|_{t=0} f_-\big(\vec{\gamma}_{(x,v)}(-\tau_-(x,v))\big) = 0. \qquad \square\end{aligned}$$

Proposition 2.5. *If $\|\tau\sigma_a\|_{L^\infty} < \infty$ then $J : L^1(\Gamma_-, d\mu) \to \mathcal{W}$ with*

$$\|Jf_-\|_{\mathcal{W}} \le (1 + \|\tau\sigma_a\|_{L^\infty})\|f_-\|_{L^1(\Gamma_-,d\mu)}.$$

The proof is carried out as in [**CS2**]. Set

$$T_0 f = -\mathcal{D}f - \sigma_a f, \qquad T_1 f(x,v) = \int_{T_xM} k(x,v',v) f(x,v')\, dv',$$

and put $T = T_0 + T_1$. We wish to solve $T_0 f + T_1 f = 0$ with $f|_{\Gamma_-} = f_-$. To this end, multiply $T_0 f = -T_1 f$ by $E(x,v,-\tau_-(x,v),0)$. Then

$$\begin{aligned}\mathcal{D}(E(x,v,-\tau_-(x,v),0)f(x,v)) &= -E(x,v,-\tau_-(x,v),0)\, T_0 f(x,v)\\ &= E(x,v,-\tau_-(x,v),0)\, T_1 f(x,v)\end{aligned}$$

has solution

$$\begin{aligned}&E(x,v,-\tau_-(x,v),0)f(x,v)\\ &\quad= \int_0^{\tau_-(x,v)} E(\vec{\gamma}_{(x,v)}(t-\tau_-(x,v)),-t,0)(T_1 f)\big(\vec{\gamma}_{(x,v)}(t-\tau_-(x,v))\big)\, dt\\ &\quad= \int_0^{\tau_-(x,v)} E(x,v,-\tau_-(x,v),t-\tau_-(x,v))(T_1 f)\big(\vec{\gamma}_{(x,v)}(t-\tau_-(x,v))\big)\, dt,\end{aligned}$$

where we have used the identities (6) and (7). Define K by

$$\begin{aligned}Kf(x,v) &= -E^{-1}(x,v,-\tau_-(x,v),0)\int_0^{\tau_-(x,v)} E(x,v,-\tau_-(x,v),t-\tau_-(x,v))\\ &\qquad\qquad \cdot (T_1 f)\big(\vec{\gamma}_{(x,v)}(t-\tau_-(x,v))\big)\, dt\\ &= -\int_0^{\tau_-(x,v)} E(x,v,0,t-\tau_-(x,v))(T_1 f)\big(\vec{\gamma}_{(x,v)}(t-\tau_-(x,v))\big)\, dt.\end{aligned}$$

Taking into account the boundary condition $f|_{\Gamma_-} = f_-$,

$$f(x,v) = -Kf(x,v) + E(x,v,0,-\tau_-(x,v))f_-\big(\vec{\gamma}_{(x,v)}(-\tau_-(x,v))\big),$$

that is,

$$(I+K)f = Jf_-. \tag{8}$$

Consider the unbounded operators $\mathbf{T}_0 f = T_0 f$ and $\mathbf{T}f = Tf$ with domains

$$\begin{aligned}D(\mathbf{T}_0) &= \big\{f \in L^1(TM) \mid T_0 f \in L^1(TM),\ f|_{\Gamma_-} = 0\big\},\\ D(\mathbf{T}) &= \big\{f \in L^1(TM) \mid Tf \in L^1(TM),\ f|_{\Gamma_-} = 0\big\}.\end{aligned}$$

Formally,

$$\mathbf{T}_0^{-1} f(x,v) = -\int_0^{\tau_-(x,v)} E(x,v,0,t-\tau_-(x,v)) f\big(\vec{\gamma}_{(x,v)}(t-\tau_-(x,v))\big)\, dt$$

and, again formally, $K = \mathbf{T}_0^{-1} T_1$. In the following proposition we show that on the appropriate spaces these formal statements are precise. Once again, the arguments mirror those in [**CS2**].

Proposition 2.6.

(i) *The operators $\tau^{-1}\mathbf{T}_0^{-1}$ and $T_1\tau$ are bounded on $L^1(TM)$.*
(ii) *$K = \mathbf{T}_0^{-1} T_1$ is bounded on $L^1(TM, \tau^{-1} dv_x\, d\omega(x))$, the operator norm of K is bounded by $\|\tau\sigma_p\|_{L^\infty} < 1$, and so $I + K$ is invertible on this space.*
(iii) *Equation (8) and hence (2) is uniquely solvable for $f_- \in L^1(\Gamma_-, d\mu)$ with solution $f \in \mathcal{W}$.*
(iv) *The albedo operator $\mathcal{A} : L^1(\Gamma_-, d\mu) \to L^1(\Gamma_+, d\mu)$ is a bounded map.*
(v) *The operator $\tau^{-1}\mathbf{T}^{-1}$ is bounded on $L^1(TM)$.*

Proof. First, if $f \in L^1(TM)$ then

$$\begin{aligned}
&\|\tau^{-1}\mathbf{T}_0^{-1} f\|_{L^1(TM)} \\
&\quad \le \int_{\Gamma_-} \int_0^{\tau_+(x',v')} \frac{1}{\tau} \left| \int_0^{\tau_-(\vec{\gamma}_{(x',v')}(t))} f\big(\vec{\gamma}_{(\vec{\gamma}_{(x',v')}(t))}\big(s - \tau_-(\vec{\gamma}_{(x',v')}(t))\big)\big)\, ds \right| dt\, d\mu \\
&\quad \le \int_{\Gamma_-} \int_0^{\tau_+(x',v')} \frac{1}{\tau_+(x',v')} \int_0^{\tau_+(x',v')} \big| f(\vec{\gamma}_{(x',v')}(s)) \big|\, ds\, dt\, d\mu \\
&\quad = \int_{\Gamma_-} \int_0^{\tau_+(x',v')} \big| f(\vec{\gamma}_{(x',v')}(s)) \big|\, ds\, d\mu = \|f\|_{L^1(TM)}.
\end{aligned}$$

Next,

$$\|T_1\tau f\|_{L^1(TM)} \le \left\| \frac{1}{\tau}\sigma_p \right\|_{L^\infty(TM)} \|f\|_{L^1(TM)}, \tag{9}$$

by (4). It follows that $K = \mathbf{T}_0^{-1} T_1$ and

$$\begin{aligned}
\|\tau^{-1} K f\|_{L^1(TM)} = \|\tau^{-1}\mathbf{T}_0^{-1} T_1 f\|_{L^1(TM)} &\le \|T_1 f\|_{L^1(TM)} \\
&\le \|\tau\sigma_p\|_{L^\infty(TM)} \|\tau^{-1} f\|_{L^1(TM)} \\
&< \|\tau^{-1} f\|_{L^1(TM)},
\end{aligned} \tag{10}$$

by (4). Thus $(I + K)$ is invertible on $L^1(TM, \tau^{-1} dv_x\, d\omega(x))$ and (8) has solution $f = (I + K)^{-1} J f_-$. We now show that $f \in \mathcal{W}$ and so has a

well-defined trace. We have

$$\begin{aligned}\|\tau_- f\|_{L^1(TM)} &\le \big(1 - \|\tau\sigma_p\|_{L^\infty(TM)}\big)^{-1}\|\tau^{-1}Jf_-\|_{L^1(TM)}\\ &\le \big(1 - \|\tau\sigma_p\|_{L^\infty(TM)}\big)^{-1}\|f_-\|_{L^1(\Gamma_-,d\mu)},\end{aligned}$$

where the last inequality follows from

$$\begin{aligned}\|\tau^{-1}Jf\|_{L^1(TM)} &\le \int_M\int_{T_xM}\tau^{-1}(x,v)\big|f_-\big(\vec{\gamma}_{(x,v)}(-\tau_-(x,v))\big)\big|\,dv_x\,d\omega(x)\\ &= \int_{\Gamma_-}\int_0^{\tau_+(x',v')}\tau_+^{-1}(x',v')|f_-(x',v')|\,dt\,d\mu(x',v')\\ &= \|f_-\|_{L^1(\Gamma_-,d\mu)}.\end{aligned}\tag{11}$$

Since $Tf = 0$ and $\mathcal{D}f = -\sigma_a f + T_1 f$, we have

$$\|\mathcal{D}f\|_{L^1(TM)} \le \big(\|\tau\sigma_a\|_{L^\infty(TM)} + \|\tau\sigma_p\|_{L^\infty(TM)}\big)\|\tau^{-1}f\|_{L^1(TM)} < \infty$$

from the previous estimate. Thus $f \in \mathcal{W}$, and applying Theorem 2.3 we have $f|_{\Gamma_+} = \mathcal{A}f_- \in L^1(\Gamma_+, d\mu)$ with

$$\begin{aligned}\|\mathcal{A}f\|_{L^1(\Gamma_+,d\mu)} &\le \|\mathcal{D}f\|_{L^1(TM)} + \|\tau^{-1}f\|_{L^1(TM)}\\ &\le \big(1 + \|\tau\sigma_a\|_{L^\infty(TM)} + \|\tau\sigma_p\|_{L^\infty(TM)}\big)\|\tau^{-1}f\|_{L^1(TM)}\\ &\le \big(1 + \|\tau\sigma_a\|_{L^\infty} + \|\tau\sigma_p\|_{L^\infty}\big)\big(1 - \|\tau\sigma_p\|_{L^\infty}\big)^{-1}\|f_-\|_{L^1(\Gamma_-,d\mu)}.\end{aligned}$$

Finally, (v) follows by setting

$$\mathbf{T}^{-1} = (I+K)^{-1}\mathbf{T}_0^{-1} : L^1(TM) \to L^1(TM, \tau^{-1}dv_x\,d\omega(x)). \qquad \square$$

3. Singular solutions and the kernel of $\mathcal{A}$

We now solve (2) in the sense of distributions with a singular boundary condition:

$$\begin{gathered}-\mathcal{D}f(x,v) - \sigma_a(x,v)f(x,v) + \int_{T_xM} k(x,v'',v)f(x,v'')\,dv_x'' = 0,\\ f|_{\Gamma_-} = \delta_{\{\hat{x},\hat{v}\}}(x',v'),\end{gathered}\tag{12}$$

with $(\hat{x},\hat{v}) \in \Gamma_-$. Here, $\delta_{\{\hat{x},\hat{v}\}}(x',v')$ is a distribution on Γ_- defined by

$$(\delta_{\{\hat{x},\hat{v}\}},\varphi) = \int_{\Gamma_-}\delta_{\{\hat{x},\hat{v}\}}(x',v')\varphi(x',v')\,d\mu(x',v') = \varphi(\hat{x},\hat{v}).$$

Let $\varphi_- \in C_0^\infty(\Gamma_-)$ and φ be the solution to

$$\begin{gathered}-\mathcal{D}\varphi(x,v) - \sigma_a(x,v)\varphi(x,v) + \int_{T_xM} k(x,v,v')\varphi(x,v')\,dv_x' = 0,\\ \varphi|_{\Gamma_-} = \varphi_-.\end{gathered}\tag{13}$$

Then

$$\varphi = (I+K)^{-1}J\varphi_- = J\varphi_- - KJ\varphi_- + \mathbf{T}^{-1}T_1KJ\varphi_-. \tag{14}$$

We shall analyze the three terms in this expression for φ, determining the distribution kernel of the solution operator $\varphi_- \mapsto \varphi$ of (13), which solves (12) in the sense of distributions. This is the content of the following three propositions:

Proposition 3.1. *For the first term in* (14) *we have*

$$J\varphi_-(x,v) = \int_{\Gamma_-} f_0(x,v,x',v')\,\varphi_-(x',v')\,d\mu(x',v'), \tag{15}$$

with

$$f_0(x,v,x',v') = \int_0^{\tau_+(x',v')} E(x,v,0,-\tau_-(x,v))\,\delta_{(x,v)}(\vec{\gamma}_{(x',v')}(t))\,dt.$$

Proof. If $\psi \in C_0^\infty(TM)$ then

$$\begin{aligned}
&(J\varphi_-,\psi)\\
&\quad= \int_{\Gamma_-}\int_0^{\tau_+(x',v')} J\varphi_-(\vec{\gamma}_{(x',v')}(t))\psi(\vec{\gamma}_{(x',v')}(t))\,dt\,d\mu(x',v')\\
&\quad= \int_{\Gamma_-}\varphi_-(x',v')\int_0^{\tau_+(x',v')} E(\vec{\gamma}_{(x',v')}(t),0,-t)\psi(\vec{\gamma}_{(x',v')}(t))\,dt\,d\mu(x',v')
\end{aligned}$$

(since $-\tau_-(\vec{\gamma}_{(x',v')}(t)) = -t$),

$$\begin{aligned}
&= \int_M\int_{T_xM}\int_{\Gamma_-}\varphi_-(x',v')\int_0^{\tau_+(x',v')} E(x,v,0,-\tau_-(x,v))\\
&\qquad\qquad\cdot\delta_{(x,v)}(\vec{\gamma}_{(x',v')}(t))\,dt\,d\mu(x',v')\,\psi(x,v)\,dv_x\,d\omega(x),
\end{aligned}$$

where $\delta_{(\overline{x},\overline{v})}$ is the distribution on TM defined by

$$(\delta_{(\overline{x},\overline{v})},\varphi) = \int_M\int_{T_xM}\delta_{(\overline{x},\overline{v})}(x,v)\varphi(x,v)\,dv_x\,d\omega = \varphi(\overline{x},\overline{v}).$$

Thus,

$$J\varphi_-(x,v) = \int_{\Gamma_-} f_0(x,v,x',v')\varphi(x',v')\,d\mu(x',v'),$$

with f_0 as given in the statement of the proposition. □

Due to our assumptions on the metric, the following parallel translation map is globally well-defined: given $(x,v) \in TM$ and $y \in M$ denote by $\mathcal{P}(v;x,y) : T_xM \to T_yM$ the parallel translation of v along the (unique) geodesic joining x and y.

Proposition 3.2. *For the second term in* (14) *we have*

$$KJ\varphi_-(x,v) = \int_{\Gamma_-} f_1(x,v,x',v')\,\varphi_-(x',v')\,d\mu(x',v') \tag{16}$$

with

$$\begin{aligned} f_1(x,v,x',v') &= \int_0^{\tau_+(x',v')} \int_0^{\tau_-(x,v)} E(x,v,0,s-\tau_-(x,v))E(x',v',0,r) \\ &\quad \cdot k\big(\vec{\gamma}_{(x',v')}(r), \mathcal{P}(\dot\gamma_{(x,v)}(s-\tau_-(x,v)); \gamma_{(x,v)}(s-\tau_-(x,v)), \gamma_{(x',v')}(r))\big) \\ &\quad \cdot \delta_{\{\gamma_{(x,v)}(s-\tau_-(x,v))\}}(\gamma_{(x',v')}(r))\,ds\,dr. \end{aligned}$$

Proof. Consider, for $\psi \in C_0^\infty(TM)$,

$$(KJ\varphi_-, \psi) \;=\; \int_{\Gamma_-} \int_0^{\tau_+(x',v')} KJ\varphi_-(\vec{\gamma}_{(x',v')}(t))\,\psi(\vec{\gamma}_{(x',v')}(t))\,dt\,d\mu(x',v').$$

First, using identities (6) and (7) together with the fact that $\tau_-(x',v') = 0$ when $(x',v') \in \Gamma_-$, we get

$$KJ\varphi_-(\vec{\gamma}_{(x',v')}(t)) = -\int_0^t E(x',v',t,s) T_1 J\varphi_-(\vec{\gamma}_{(x',v')}(s))\,ds,$$

$$\begin{aligned} T_1 J\varphi_-(\vec{\gamma}_{(x',v')}(s)) = \int_{T_{\gamma_{(x',v')}(s)}M} & k(\gamma_{(x',v')}(s), w, \dot\gamma_{(x',v')}(s)) \\ & \cdot J\varphi_-(\gamma_{(x',v')}(s), w)\,dw_{\gamma_{(x',v')}(s)} \end{aligned}$$

and

$$\begin{aligned} J\varphi_-(\gamma_{(x',v')}(s), w) = -E(\gamma_{(x',v')}(s), w, -\tau_-(\gamma_{(x',v')}(s), w), 0) \\ \cdot \varphi_-\big(\vec{\gamma}_{(\gamma_{(x',v')}(s),w)}(-\tau_-(\gamma_{(x',v')}(s), w))\big). \end{aligned}$$

Thus,

$$\begin{aligned} &(KJ\varphi_-, \psi) \\ &\quad = \int_{\Gamma_-} \int_0^{\tau_+(x',v')} \int_0^t \int_{T_{\gamma_{(x',v')}(s)}M} E(x',v',t,s) \\ &\qquad \cdot E\big(\gamma_{(x',v')}(s), w, -\tau_-(\gamma_{(x',v')}(s), w), 0\big) k(\gamma_{(x',v')}(s), w, \dot\gamma_{(x',v')}(s)) \\ &\qquad \cdot \varphi_-(\bar{x}, \bar{v})\,\psi(\vec{\gamma}_{(x',v')}(t))\,dw\,ds\,dt\,d\mu(x',v') \end{aligned}$$

where $(\bar{x}, \bar{v}) = \big(\vec{\gamma}_{(\gamma_{(x',v')}(s),w)}(-\tau_-(\gamma_{(x',v')}(s), w))\big)$; see Figure 1 on the next page.

We now perform the change of variables from (x',v',t) to (x,v), where $x = \gamma_{(x',v')}(t)$ and $v = \dot\gamma_{(x',v')}(t)$. Then $t = \tau_-(x,v)$,

$$\gamma_{(x',v')}(s) = \gamma_{(x,v)}(s-\tau_-(x,v)), \quad \text{and} \quad \dot\gamma_{(x',v')}(s) = \dot\gamma_{(x,v)}(s-\tau_-(x,v)).$$

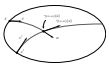


Figure 1.

Using parallel translation we may now introduce a delta distribution to obtain

$$
\begin{aligned}
&(KJ\varphi_-,\psi)\\
&\quad=\int_M\int_{T_xM}\int_0^{\tau_-(x,v)}\int_M \delta_{\gamma_{(x,v)}(s-\tau_-(x,v))}(y)\int_{T_yM} E(x,v,0,s-\tau_-(x,v))\\
&\qquad\qquad\cdot E(y,w,-\tau_-(y,w),0)\\
&\qquad\qquad\cdot\ k\big(y,w,\mathcal{P}(\dot\gamma_{(x,v)}(s-\tau_-(x,v));\gamma_{(x,v)}(s-\tau_-(x,v)),y)\big)\\
&\qquad\qquad\cdot\varphi_-\big(\vec\gamma_{(y,w)}(-\tau_-(y,w))\big)\,\psi(x,v)\,dw_y\,d\omega(y)\,ds\,dv_x\,d\omega(x),
\end{aligned}
$$

where

$$(\delta_x,\varphi)_M=\int_M \delta_x(y)\varphi(y)\,d\omega(y)=\varphi(x).$$

Make another change of variables $(y',w',r)=(\vec\gamma_{(y,w)}(-\tau_-(y,w)),\tau_-(y,w))$ and interchange the order of integration $dr\,d\mu(y',w')\,ds$ to $ds\,dr\,d\mu(y',w')$. We obtain

$$
\begin{aligned}
&(KJ\varphi_-,\psi)\\
&\quad=\int_M\int_{T_xM}\psi(x,v)\int_{\Gamma_-}\varphi_-(y',w')\int_0^{\tau_+(y',w')}\int_0^{\tau_-(x,v)} E(x,v,0,s-\tau_-(x,v))\\
&\qquad\cdot E(y',w',0,r)\\
&\qquad\cdot k\big(\vec\gamma_{(y',w')}(r),\mathcal{P}(\dot\gamma_{(x,v)}(s-\tau_-(x,v));\gamma_{(x,v)}(s-\tau_-(x,v)),\gamma_{(y',w')}(r))\big)\\
&\qquad\cdot\delta_{(\gamma_{(x,v)}(s-\tau_-(x,v)))}(\gamma_{(y',w')}(r))\,ds\,dr\,d\mu(y',w')\,dv_x\,d\omega(x).
\end{aligned}
$$

Thus we have (16). □

Proposition 3.3. *For the third and final term in* (14) *we have*

$$\mathbf{T}^{-1}T_1KJ\varphi_-(x,v)=\int_{\Gamma_-} f_2(x,v,x',v')\varphi_-(x',v')\,d\mu(x',v')$$

with $f_2\in L^\infty(\Gamma_-;\mathcal{W})$.

Proof. We have

$$
\begin{aligned}
(17)\quad & T_1KJ\varphi_-(x,v) \\
&= \int_{T_xM}\int_0^{\tau_-(x,w')}\int_{T_{\hat{x}}M} k(x,w',v)k\big(\hat{x},w,\dot{\gamma}_{(x,w')}(t-\tau_-(x,w'))\big) \\
&\qquad \cdot E(x,w',0,t-\tau_-(x,w'))E(\hat{x},w,-\tau_-(\hat{x},w),0) \\
&\qquad \cdot \varphi_-\big(\vec{\gamma}_{(\hat{x},w)}(-\tau_-(\hat{x},w))\big)\,dw_{\hat{x}}\,dt\,dw'_x,
\end{aligned}
$$

where $\hat{x} = \gamma_{(x,\,w')}(t-\tau_-(x,\,w'))$. Take the $L^1(TM)$ norm of $T_1KJ\varphi_-(x,v)$. We observe that

$$
\begin{aligned}
(18)\quad & \|T_1KJ\varphi_-(x,v)\|_{L^1(TM)} \\
&\le \|T_1\tau\|_{L^1(TM)\to L^1(TM)}\|\tau^{-1}K\tau\|_{L^1(TM)\to L^1(TM)}\|\tau^{-1}J\varphi_-\|_{L^1(TM)} \\
&\le \|\varphi_-\|_{L^1(\Gamma_-,d\mu)},
\end{aligned}
$$

by (9), (4), (10) and (11). Thus by Fubini's theorem, the integral above defining $T_1KJ\varphi_-(x,v)$ is absolutely convergent for a.e. $(x,v)\in TM$.

In the distributional sense, we make the following computation of the distribution kernel of T_1KJ: for $x\in M$ let $\mathcal{S}(x)=\{\hat{v}\in T_xM \mid \|\hat{v}\|=1\}$ and extend all functions defined on T_xM by zero to be defined on all of $\{r\hat{v} \mid r\ge 0,\ \hat{v}\in\mathcal{S}(x)\}$. Let $(x,v,x',v')\in\Gamma_+\times\Gamma_-$. Then, as in (17),

$$
\begin{aligned}
&(T_1KJ\delta_{(x',v')})(x,v) \\
&\qquad = \int_{T_xM}\int_0^{\tau_-(x,w')}\int_{T_{\hat{x}}M} k(x,w',v)k\big(\hat{x},w,\dot{\gamma}_{(x,w')}(t-\tau_-(x,w'))\big) \\
&\qquad\qquad \cdot E(x,w',0,t-\tau_-(x,w'))E(\hat{x},w,-\tau_-(\hat{x},w),0) \\
&\qquad\qquad \cdot \delta_{(x',v')}\big(\vec{\gamma}_{(\hat{x},w)}(-\tau_-(\hat{x},w))\big)\,dw_{\hat{x}}\,dt\,dw'_x.
\end{aligned}
$$

We shall cease to write out the arguments of the functions E in order to compress the presentation. Change the integration over $w'\in T_xM$ to polar coordinates $(r,\hat{w}')\in\mathbb{R}^+\times\mathcal{S}(x)$ and make the change of variables $y=\hat{x}=\gamma_{(x,\hat{w}')}(t-\tau_-(x,\hat{w}'))$. Due to the assumptions on the metric g this is a global diffeomorphism. Let J be the determinant of the Jacobian of this change of variables. If γ is the unit speed geodesic joining $\gamma(0)=y$ to $\gamma(d)=x$ (for $d>0$), denote by $\hat{w}_x(y)=\dot{\gamma}(d)\in T_xM$ the (unit) tangent vector of γ at x and denote by $\hat{w}(y)=\dot{\gamma}(0)\in T_yM$ the (unit) tangent vector of γ at y. Then

$$
\begin{aligned}
&(T_1KJ\delta_{(x',v')})(x,v) \\
&\qquad = \int_0^\infty\int_M\int_{V(y)} k(x,r\hat{w}_x(y),v)k(y,w,r\hat{w}(y)) \\
&\qquad\qquad \cdot E(\cdot)E(\cdot)\delta_{(x',v')}\big(\vec{\gamma}_{(y,w)}(-\tau_-(y,w))\big)|J|\,dw_y\,d\omega(y)\,r^{n-2}\,dr.
\end{aligned}
$$

Change variables again setting $(y,w) = \vec{\gamma}_{(a',b')}(s)$ and integrate with respect to $d\mu(a',b')$:

$$\begin{aligned}(T_1KJ\delta_{(x',v')})(x,v)&\\ &= \int_0^\infty \int_{\Gamma_-} \int_0^{\tau_+(a',b')} k\big(x, r\hat{w}_x(\gamma_{(a',b')}(s)), v\big)\, k\big(\vec{\gamma}_{(a',b')}(s), r\hat{w}(\gamma_{(a',b')}(s))\big)\\ &\qquad \cdot E(\,\cdot\,)E(\,\cdot\,)\delta_{(x',v')}(a',b')|J|\, ds\, d\mu(a',b')\, r^{n-2}\, dr\\ &= \int_0^\infty \int_0^{\tau_+(x',v')} k\big(x, r\hat{w}_x(\gamma_{(x',v')}(s)), v\big)\, k\big(\vec{\gamma}_{(x',v')}(s), r\hat{w}(\gamma_{(x',v')}(s))\big)\\ &\qquad \cdot E(\,\cdot\,)E(\,\cdot\,)|J|\, ds\, r^{n-2}\, dr.\end{aligned}$$

Claim 3.4. If $\varphi_- \in L^1(\Gamma_-, d\mu)$ then for $(x,v) \in TM$

$$(T_1KJ\varphi_-)(x,v) = \int_{\Gamma_-} (T_1KJ\delta_{(x',v')})(x,v)\varphi_-(x',v')\, d\mu(x',v'). \tag{19}$$

Proof. By the definition of T_1KJ (see (17)), and manipulating the integral as above, we obtain

$$\begin{aligned}(T_1KJ\varphi_-)(x,v)&\\ &= \int_{T_xM} \int_0^{\tau_-(x,w')} \int_{T_{\hat{x}}M} k(x,w',v) k\big(\hat{x}, w, \dot{\gamma}_{(x,w')}(t-\tau_-(x,w'))\big)\\ &\qquad \cdot E(\,\cdot\,)E(\,\cdot\,)\delta_{(x',v')}\big(\vec{\gamma}_{(\hat{x},w)}(-\tau_-(\hat{x},w))\big)\, dw_{\hat{x}}\, dt\, dw'_x\\ &= \int_0^\infty \int_{\Gamma_-} \int_0^{\tau_+(a',b')} k\big(x, r\hat{w}_x(\gamma_{(a',b')}(s)), v\big) k\big(\vec{\gamma}_{(a',b')}(s), r\hat{w}(\gamma_{(a',b')}(s))\big)\\ &\qquad \cdot E(\,\cdot\,)E(\,\cdot\,)\varphi_-(a',b')|J|\, ds\, d\mu(a',b')\, r^{n-2}\, dr.\end{aligned}$$

Relabeling (a',b') as (x',v') and interchanging the order of integration from $ds\, d\mu(x',v')\, dr$ to $ds\, dr\, d\mu(x',v')$ we obtain the right-hand side of (19) as claimed.

From (18) we see that $T_1KJ\varphi_- \in L^1(TM)$ for all $\varphi_- \in L^1(\Gamma_-, d\mu)$. Thus from (19) and the Riesz representation theorem for $L^1(TM)$-valued functionals, the kernel $T_1KJ\delta_{(x',v')}(x,v)$ is in $L^\infty(\Gamma_-(x',v'); L^1(TM))$.

Define $f_2(\cdot,\cdot,x',v') = \mathbf{T}^{-1}T_1KJ\delta_{(x',v')}$, where $\mathbf{T}^{-1}$ acts on the variables (x,v) with (x',v') as parameters. Then

$$\mathbf{T}^{-1}T_1KJ\varphi_-(x,v) = \int_{\Gamma_-} f_2(x,v,x',v')\varphi_-(x',v')\, d\mu(x',v').$$

(The interchange of the operator $\mathbf{T}^{-1}$ and the integration is justified by the continuity of $\mathbf{T}^{-1}$; see [**DU**] for example.) We show that $f_2 \in L^\infty(\Gamma_-; \mathcal{W})$, so that restriction to $(x,v) \in \Gamma_+$ is well-defined. By Proposition 2.6, $\mathbf{T}^{-1}$ is

bounded from $L^1(TM)$ to $L^1(TM, \tau^{-1} dv_x\, d\omega(x))$, so

$$f_2 \in L^\infty\big(\Gamma_-; L^1(TM, \tau^{-1} dv_x\, d\omega(x))\big). \tag{20}$$

Next, $\mathcal{D} f_2 = -\sigma_a f_2 + T_1 f_2 + T_1 K J \delta_{(x',v')}$ and $f_2 \mapsto -\sigma_a f_2 + T_1 f_2$ is bounded from $L^1(TM, \tau^{-1} dv_x\, d\omega(x))$ to $L^1(TM)$ by (3) and (4), so

$$\mathcal{D} f_2 \in L^\infty(\Gamma_-; L^1(TM)). \tag{21}$$ □

To summarize our analysis of the terms in (14), we have shown that (13) has solution

$$\varphi = \int_{\Gamma_-} f(x, v, x', v') \varphi_-(x', v') d\mu(x', v') \tag{22}$$

where the integral is in distributional sense and f is the sum of the three terms given by (15), (16) and f_2 above. Furthermore, f solves (12) in the sense of distributions. In other words, f is the distribution kernel of the solution operator $\varphi_- \mapsto \varphi$ of (13). We proceed now to show that the distribution kernel $\alpha(x, v, x', v')$ (where $(x, v) \in \Gamma_+$, $(x', v') \in \Gamma_-$) of the albedo operator $\mathcal{A}$ is the restriction of f to $(x, v) \in \Gamma_+$.

Theorem 3.5. *The distribution kernel $\alpha(x, v, x', v')$ of $\mathcal{A}$ is expressible as a sum $\alpha = \alpha_0 + \alpha_1 + \alpha_2$, with*

$$\begin{aligned}
\alpha_0 &= E(x, v, -\tau_-(x, v), 0) \delta_{\{\vec{\gamma}_{(x,v)}(-\tau_-(x,v))\}}(x', v') \\
&= \exp\left(\int_{-\tau_-(x,v)}^{0} \sigma_a(\vec{\gamma}_{(x,v)}(r))\, dr \right) \delta_{\{\vec{\gamma}_{(x,v)}(-\tau_-(x,v))\}}(x', v'), \\
\alpha_1 &= \int_0^{\tau_+(x',v')} \int_0^{\tau_-(x,v)} E(x, v, 0, s - \tau_-(x, v)) E(x', v', 0, r) \\
&\qquad \cdot k\big(\vec{\gamma}_{(x',v')}(r), \mathcal{P}(\dot{\gamma}_{(x,v)}(s - \tau_-(x, v)); \gamma_{(x,v)}(s - \tau_-(x, v)), \gamma_{(x',v')}(r))\big) \\
&\qquad \cdot \delta_{\{\gamma_{(x,v)}(s - \tau_-(x,v))\}}(\gamma_{(x',v')}(r))\, ds\, dr, \\
\alpha_2 &\in L^\infty(\Gamma_-; L^1(\Gamma_+, d\mu)).
\end{aligned}$$

Proof. We have formally $\alpha_j = f_j$ restricted to $(x, v) \in \Gamma_+$, $j = 0, 1, 2$. Let $\varphi_- \in C_0^m(\Gamma_-)$ and for $(x, v) \in \Gamma_+$ consider $\varphi_j(x, v) = \int_{\Gamma_-} f_j\, \varphi_-\, d\mu(x', v')$. From (15), changing variables to $(y, w) = \vec{\gamma}_{(x',v')}(t)$, we have

$$\begin{aligned}
&\varphi_0(x, v) \\
&\quad = \int_M \int_{T_y M} E(x, v, -\tau_-(x, v), 0) \delta_{\{x,v\}}(y, w) \varphi_-\big(\vec{\gamma}_{(y,w)}(-\tau_-(y, w))\big)\, dw_y\, d\omega \\
&\quad = E(x, v, -\tau_-(x, v), 0) \varphi_-(\vec{\gamma}_{(x,v)}(-\tau_-(x, v))) \\
&\quad = \int_{\Gamma_-} \alpha_0(x, v, x', v') \varphi_-(x', v')\, d\mu(x', v'),
\end{aligned}$$

which justifies the expression for α_0. The expression for α_1 is nothing more than a restatement of (16) with $(x,v)\in\Gamma_+$. That $\alpha_2\in L^\infty(\Gamma_-;L^1(\Gamma_+,d\mu))$ follows from (20), (21) and Proposition 2.5. □

4. The solution to the inverse problem

In this section we show that from the distribution kernel α of the albedo operator $\mathcal{A}$ we can isolate the terms that differ in strength of singularity. In dimensions 3 and higher, we isolate α_0 and α_1 and show that this yields the absorption coefficient σ_a and the scattering kernel k. We are able to do so because α_0 and α_1 are delta-type singularities in $\Gamma_+\times\Gamma_-$ supported on varieties of differing dimension and α_2 is an L^∞ function. In dimension 2, we isolate α_0 from α, which yields σ_a, but are unable to determine k since in this case α_1 is in fact a locally L^1 function and so not distinguishable from α_2.

We shall use the following global coordinates on TM: fix $p\in M$ and let $\{E_i\}$ be an orthonormal basis for T_pM. Define $E_p:\mathbb{R}^n\to T_pM$ by $E_p(x^1,\dots,x^n)=\sum x^iE_i$. Let $\operatorname{Exp}_p:T_pM\to M$ denote the exponential map based at p. Note that Exp_p provides a diffeomorphism from $\operatorname{Exp}_p^{-1}(M)$ to M due to our assumptions on the metric. Recall that $\mathcal{P}(v;x,p):T_xM\to T_pM$ is the parallel translation of v along the unique geodesic joining x and p. We may now define global coordinates for TM by $\phi:TM\to\mathbb{R}^n\times\mathbb{R}^n$,

$$\phi(x,v)=\big(\phi_1(x),\phi_2(v)\big)=\big((E_p^{-1}\circ\operatorname{Exp}_p^{-1})(x),(E_p^{-1}\circ\mathcal{P})(v;x,p)\big).$$

We first show that one can isolate α_0 from α. Let $\psi\in C_0^\infty(\mathbb{R})$ satisfy $0\le\psi\le1$, $\psi(0)=1$, and $\int\psi(x)\,dx=1$. Define $\psi_\varepsilon(x)=\psi\big(x/\varepsilon\big)$. Now let $\varphi:\Gamma_+\times\Gamma_-\to\mathbb{R}$ be a defining function for the support of the distribution $\delta_{\{\vec\gamma_{(x,v)}(-\tau_-(x,v))\}}(x',v')$; that is, set

$$\begin{aligned}&\varphi(x,v,x',v')\\&\quad=\big\|\phi_1(x')-\phi_1\big(\gamma_{(x,v)}(-\tau_-(x,v))\big)\big\|^2+\big\|\phi_2(v')-\phi_2\big(\dot\gamma_{(x,v)}(-\tau_-(x,v))\big)\big\|^2,\end{aligned}$$

where the norm $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^n$. We have that the support of $\delta_{\{\vec\gamma_{(x,v)}(-\tau_-(x,v))\}}(x',v')$ is the set $\{\varphi(x,v,x',v')=0\}$.

Proposition 4.1. *The following limit holds in $L^1_{\mathrm{loc}}(\Gamma_+,d\mu(x,v))$:*

$$\lim_{\varepsilon\to0}\int_{\Gamma_-}\alpha(x,v,x',v')\,(\psi_\varepsilon\circ\varphi)(x,v,x',v')\,d\mu(x',v')=\exp\left(\int_{-\tau_-(x,v)}^0\sigma_a(\vec\gamma_{(x,v)}(r))\,dr\right)$$

where the integral is meant in the sense of distributions.

Proof. When α is replaced by α_0 the result is immediate. It remains to show that when α is replaced by α_1 and α_2 the limit vanishes. Consider first α_1.

We shall show that the limit vanishes when considered in

$$L^1\big(\{(x,v)\in\Gamma_+ \mid \|v\|_{g_x}\le M\}, d\mu(x,v)\big),$$

for any $M>0$. Let $0\le\chi\in C_0^\infty(\mathbb{R})$. Then

(23)

$$\begin{aligned}
0 &\le \int_{\Gamma_+}\int_{\Gamma_-}\chi(\|v\|_{g_x})\alpha_1(x,v,x',v')(\psi_\varepsilon\circ\varphi)(x,v,x',v')\,d\mu(x',v')\,d\mu(x,v)\\
&\le \int_{\Gamma_+}\int_{\Gamma_-}\int_0^{\tau_+(x',v')}\int_0^{\tau_-(x,v)}\chi(\|v\|_{g_x})\\
&\qquad\cdot k\big(\vec\gamma_{(x',v')}(r),\mathcal{P}(\dot\gamma_{(x,v)}(s-\tau_-(x,v));\gamma_{(x,v)}(s-\tau_-(x,v)),\gamma_{(x',v')}(r))\big)\\
&\qquad\cdot\delta_{\{\gamma_{(x,v)}(s-\tau_-(x,v))\}}(\gamma_{(x',v')}(r))(\psi_\varepsilon\circ\varphi)(x,v,x',v')\\
&\qquad\cdot ds\,dr\,d\mu(x',v')\,d\mu(x,v)\\
&\le \int_M\int_{T_yM}\int_{T_yM}\chi(\|w\|_{g_y})k(y,w',w)\\
&\qquad\qquad\cdot(\psi_\varepsilon\circ\varphi)\big(\vec\gamma_{(y,w)}(\tau_+(y,w)),\vec\gamma_{(y,w')}(-\tau_-(y,w'))\big)\,dw'_y\,dw_y\,d\omega(y),
\end{aligned}$$

where $(y',w')=\vec\gamma_{(x',v')}(r)$, $(y,w)=\vec\gamma_{(x,v)}(s-\tau_-(x,v))$ and we have used the fact that $\|w\|_{g_y}=\|\dot\gamma_{(x,v)}(s-\tau_-(x,v))\|_{g_y}=\|v\|_{g_x}$.

Now

$$\begin{aligned}
&(\psi_\varepsilon\circ\varphi)\big(\vec\gamma_{(y,w)}(\tau_+(y,w)),\vec\gamma_{(y,w')}(-t_-(y,w'))\big)\\
&\qquad=\psi_\varepsilon\Big(\big\|\phi_1\big(\gamma_{(y,w')}(-\tau_-(y,w'))\big)-\phi_1\big(\gamma_{(y,w)}(-\tau_-(y,w))\big)\big\|^2\\
&\qquad\qquad+\|\phi_2\big(\dot\gamma_{(y,w')}(-\tau_-(y,w'))\big)-\phi_2\big(\dot\gamma_{(y,w)}(-\tau_-(y,w))\big)\|^2\Big)\\
&\qquad=\psi_\varepsilon\Big(\big\|\phi_1\big(\gamma_{(y,w')}(-\tau_-(y,w'))\big)-\phi_1\big(\gamma_{(y,w)}(-\tau_-(y,w))\big)\big\|^2\\
&\qquad\qquad+\|\phi_2(w')-\phi_2(w)\|^2\Big),
\end{aligned}$$

so it follows that there is C depending on ψ such that

$$\begin{aligned}
&\operatorname{supp}(\psi_\varepsilon\circ\varphi)\big(\vec\gamma_{(y,w)}(\tau_+(y,w)),\vec\gamma_{(y,w')}(-t_-(y,w'))\big)\\
&\qquad\subset\{(y,w',w)\in M\times T_yM\times T_yM\mid\|w'-w\|_{g_y}<C\varepsilon\}.
\end{aligned}$$

If we set $W_\varepsilon=\big\{(y,w',w)\;\big|\;\|w'-w\|_{g_y}<C\varepsilon\text{ and }\|w\|_{g_y}\in\operatorname{supp}\chi\big\}$, then from (23) we have

$$\begin{aligned}
0&\le\int_{\Gamma_+}\int_{\Gamma_-}\chi(\|v\|_{g_x})\alpha_1(x,v,x',v')(\psi_\varepsilon\circ\varphi)(x,v,x',v')\,d\mu(x',v')\,d\mu(x,v)\\
&\le\int_{W_\varepsilon}\chi(\|v\|_{g_x})k(y,w',w)\,dw'_y\,dw_y\,d\omega(y)
\end{aligned}$$

tending to 0 as $\lambda\to0$, since $\chi(\|v\|_{g_x})k(y,w',w)\in L^1(M\times T_yM\times T_yM)$ and the measure of $W_\varepsilon\to0$ as $\varepsilon\to0$.

Finally, for α_2,

(24)

$$0 \le \int_{\Gamma_+} \left| \int_{\Gamma_-} \chi(\|v\|_{g_x}) \alpha_2(x,v,x',v') (\psi_\varepsilon \circ \varphi)(x,v,x',v') \, d\mu(x',v') \right| d\mu(x,v)$$
$$\le \int_{V_\varepsilon} \chi(\|v\|_{g_x}) |\alpha_2(x,v,x',v')| \, d\mu(x',v') \, d\mu(x,v),$$

tending to 0 as $\lambda \to 0$. Here

$$\operatorname{supp} (\psi_\varepsilon \circ \varphi)(x,v,x',v')$$
$$\subset V_\varepsilon = \{(x,v,x',v') \in \Gamma_+ \times \Gamma_- \mid \|\mathcal{P}(v';x',p) - \mathcal{P}(v;x,p)\|_{g_p} < C\varepsilon$$
$$\text{and } \|v'\|_{g_{x'}}, \ \|v\|_{g_x} \in \operatorname{supp} \chi\}$$

and the limit holds as stated since Theorem 3.5 gives

$$\chi(\|v\|_{g_x}) \, \alpha_2(x,v,x',v') \in L^1(\Gamma_+ \times \Gamma_-, d\mu(x,v) \, d\mu(x',v'))$$

and the measure of V_ε tends to 0 as $\varepsilon \to 0$. □

Theorem 4.2. *Suppose $M \subset \mathbb{R}^n$, $n \ge 2$, is a bounded domain with C^∞ boundary, that g is a simple Riemannian metric on M, and that assumptions* (3) *and* (4) *hold. Then from the kernel of $\mathcal{A}$ we may determine the absorption coefficient $\sigma_a(x,v) = \sigma_a(x, \|v\|_{g_x})$.*

Proof. From $\mathcal{A}$ we of course know α; taking the limit as in Proposition 4.1 we obtain

$$\int_{-\tau_-(x,v)}^{0} \sigma_a(\vec{\gamma}_{(x,v)}(r)) \, dr = \int_{-\tau_-(x,v)}^{0} \sigma_a(\gamma_{(x,v)}(r), \|v_0\|) \, dr$$

for all $(x,v) \in \Gamma_+$, where $\|v_0\| = \|\dot\gamma_{(x,v)}(0)\|$ $(= \|\dot\gamma_{(x,v)}(t)\|$ for all $t)$. That is, we know the integrals of σ_a along the geodesics joining $(x',v') = \vec{\gamma}_{(x,v)}(-\tau_-(x,v)) \in \Gamma_-$ and $(x,v) \in \Gamma_+$. This is the geodesic ray transform of the function σ_a and this transform is invertible. See [**Sh1**]. □

We work now towards determining k by isolating α_1. Fix $(y,w,w') \in M \times T_yM \times T_yM$ with w and w' linearly independent and let

$$Z = \left\{ \gamma_{(y,\overline{w})}(-\tau_-(y,\overline{w})) \mid \overline{w} \in \operatorname{span}\{w,w'\} \right\} \subset \partial M.$$

We let h_1 be a defining function for the set Z as follows: for $z \in Z$ let $\gamma_z(t)$ be the geodesic joining z to y such that $\gamma_z(0) = z$ and $\gamma_z(1) = y$, and define $\pi(\dot\gamma_z(1)) \in T_yM$ to be the orthogonal projection of $\dot\gamma_z(1)$ onto $(\operatorname{span}\{w,w'\})^\perp$. Now define $h_1(z) = \|\pi(\dot\gamma_z(1))\|_{g_y}$. For $z \in \partial M$, we have $h_1(z) = 0$ if and only if $z \in Z$. Now let $\psi^1 \in C_0^\infty(\mathbb{R})$ satisfy $0 \le \psi^1 \le 1$, $\psi^1(0) = 1$, $\int_{\mathbb{R}} \psi^1(x) \, dx = 1$ and define $\psi^1_\rho : \partial M \to \mathbb{R}$ by

$$\psi^1_\rho(z) = \psi^1 \left(\frac{h_1(z)}{\rho} \right);$$

ψ^1_ρ concentrates at Z as ρ approaches 0.

For $0 \le s \le \tau(y,w)$ let $\hat{y}(s) = \gamma_{(y,w)}(s - \tau_-(y,w))$, $\hat{b}(s) = \mathcal{P}(w'; y, \hat{y}(s)) \in T_{\hat{y}(s)}M$, and denote $x^* = \gamma_{(y,w')}(-\tau_-(y,w'))$. See Figure 2.

Figure 2.

Define $h(s) = \gamma_{(\hat{y}(s),\hat{b}(s))}\big(-\tau_-(\hat{y}(s),\hat{b}(s))\big) \in \partial M$. Note that $h(\tau_-(y,w)) = x^*$. It is easily seen that $h(s) \in Z$ for each s, so $h_1(h(s)) = 0$. Denote by $w^* \in T_{x^*}\partial M$ the tangent vector

$$\frac{d}{ds}\Big|_{s=\tau_-(y,w)} h(s);$$

since w and w' are linearly independent, it is readily checked that $w^* \neq 0$. Define $h_2 : \partial M \to \mathbb{R}$ by

$$h_2(x') = \langle \operatorname{Exp}^{-1}_{x^*}(x'), w^* \rangle_{g_{x^*}} + \operatorname{sgn}\big(\langle \operatorname{Exp}^{-1}_{x^*}(x'), w^* \rangle_{g_{x^*}}\big) \|\phi_1(x') - \phi_1(x^*)\|_{\mathbb{R}^n}.$$

The function h_2 has the property that $h_2 = 0$ only at $x' = x^*$. Furthermore, if $s_0 = \tau_-(y,w)$ then

$$\frac{d}{ds}\Big|_{s=s_0} h_2(h(s)) = \|w^*\|_{g_{x^*}} \neq 0.$$

Now let $\psi^2 \in C^\infty_0(\mathbb{R})$ satisfy $0 \le \psi^2 \le 1$, $\psi^2(0) = 1$, $\int_{\mathbb{R}} \psi^2(x)\,dx = 1/\|w^*\|$, and define ψ^2_ε by

$$\psi^2_\varepsilon(x) = \frac{1}{\varepsilon}\psi^2\left(\frac{x}{\varepsilon}\right).$$

We shall need one more approximate identity function. Let $\psi^3 \in C^\infty_0(\mathbb{R}^n)$ satisfy $0 \le \psi^3 \le 0$, $\psi^3(0) = 1$, $\int_{\mathbb{R}^n} \psi^3(x)\,dx = 1$, and define

$$\psi^3_\epsilon(x) = \frac{1}{\epsilon^n}\psi^3\left(\frac{x}{\epsilon}\right).$$

Let

$$W = \{(y, w, w') \in M \times TM \times TM \mid w \text{ and } w' \text{ are linearly independent}\}.$$

Proposition 4.3. *If $n \geq 3$ and $(y, w, w') \in M \times T_yM \times T_yM$ with w and w' linearly independent we have*

$$\lim_{\varepsilon\to 0}\lim_{\rho\to 0}\lim_{\epsilon\to 0}\int_{\Gamma_-} \psi^1_\rho(x')\,\psi^2_\varepsilon(f(x'))\,\psi^3_\epsilon\big(\phi_2(v') - \phi_2(\mathcal{P}(w'; y, x'))\big) \cdot \alpha(\vec{\gamma}_{(y,w)}(\tau_+(y,w)), x', v')\, d\mu(x', v') = E(\vec{\gamma}_{(y,w)}(\tau_+(y,w)), 0, -\tau_+(y,w)) E(\vec{\gamma}_{(y,w')}(\tau_-(y,w')), 0, \tau_-(y,w')) \cdot k(y, w, w'),$$

where the limit holds in $L^1_{\mathrm{loc}}(W)$.

Proof. Replacing α by α_0 and integrating with respect to $d\mu(x', v')$ we obtain a multiple of $\psi^2_\varepsilon\big(h_2(\gamma_{(y,w)}(-\tau_-(y,w)))\big)$ that (for all sufficiently small ε) is zero unless w and w' are linearly dependent.

Substituting α_1 for α, changing variables $(a, b) = \vec{\gamma}_{(x',v')}(r)$, and interchanging the order of integration from $ds\, db_a\, d\omega(a)$ to $db_a\, d\omega(a)\, ds$ we obtain

$$\begin{aligned}\int_0^{\tau(y,w)}\int_M\int_{T_aM} &\psi^1_\rho\big(\gamma_{(a,b)}(-\tau_-(a,b))\big)\,\psi^2_\varepsilon\big(h_2\big(\gamma_{(a,b)}(-\tau_-(a,b))\big)\big) \\ &\cdot \psi^3_\epsilon\Big(\phi_2\big(\dot\gamma_{(a,b)}(-\tau_-(a,b))\big) - \phi_2\big(\mathcal{P}\big(w'; y, \gamma_{(a,b)}(-\tau_-(a,b))\big)\big)\Big) \\ &\cdot E(\cdot)E(\cdot)k\big(a, b, \dot\gamma_{(y,w)}(s-\tau_-(y,w))\big)\,\delta_{\hat y(s)}(a)\, db_a\, d\omega(a)\, ds \\ = \int_0^{\tau(y,w)}\int_{T_{\hat y}M} &\psi^1_\rho\big(\gamma_{(\hat y,b)}(-\tau_-(\hat y,b))\big)\,\psi^2_\varepsilon\big(h_2\big(\gamma_{(\hat y,b)}(-\tau_-(\hat y,b))\big)\big) \\ &\cdot \psi^3_\epsilon\Big(\phi_2\big(\dot\gamma_{(\hat y,b)}(-\tau_-(\hat y,b))\big) - \phi_2\big(\mathcal{P}\big(w'; y, \gamma_{(\hat y,b)}(-\tau_-(\hat y,b))\big)\big)\Big) \\ &\cdot E(\,\cdot\,)E(\,\cdot\,)k\big(\hat y, b, \dot\gamma_{(y,w)}(s-\tau_-(y,w))\big)\, db_{\hat y}\, ds.\end{aligned}$$

Here $\hat y = \hat y(s)$. Now $\phi_2\big(\dot\gamma_{(\hat y,b)}(-\tau_-(\hat y,b))\big) - \phi_2\big(\mathcal{P}\big(w'; y, \gamma_{(\hat y,b)}(-\tau_-(\hat y,b))\big)\big)$ equals $\phi_2(b) - \phi_2\big(\mathcal{P}(w'; y, \hat y(s))\big)$ since our global coordinates ϕ_2 are obtained by parallel translation. Set $\mathcal{P}(w'; y, \hat y(s)) = \hat b(s) \in T_{\hat y(s)}M$. The expression of ψ^2_ϵ becomes $\psi^2_\epsilon\big(\phi_2(b) - \phi_2(\hat b(s))\big)$ and taking the limit as $\epsilon \to 0$ (in L^1) we obtain

$$\int_0^{\tau(y,w)} \psi^2_\varepsilon\big(h_2(h(s))\big)E(\,\cdot\,)E(\,\cdot\,)k\big(\hat y(s), \hat b(s), \dot\gamma_{(y,w)}(s-\tau_-(y,w))\big)\, ds.$$

Set $\widetilde s = h_2(h(s))$; then

$$\frac{d\widetilde s}{ds}\Big|_{s=\tau_-(y,w)} = \|w^*\| \neq 0,$$

and so for sufficiently small ε, for all s in the support of $\psi^2_\varepsilon\big(h_2(h(s))\big)$ we have $\frac{d\widetilde s}{ds} \neq 0$ and we may perform the change of variables from s to $\widetilde s$ in the

above integral. Since $f\big(h(\tau_-(y,w))\big)=0$, setting $s=(f\circ h)^{-1}(\widetilde{s})$ we obtain (for sufficiently small δ)

$$\int_{-\delta}^{\delta}\psi_\varepsilon^1(\widetilde{s})\,E(\,\cdot\,)E(\,\cdot\,)\,k(\hat{y}(s,\hat{b}(s),\dot{\gamma}_{(y,w)}(s-\tau_-(y,w)))\frac{ds}{d\widetilde{s}}\,d\widetilde{s}$$
$$\to E(\,\cdot\,)E(\,\cdot\,)\,k\big(\hat{y}(s(0)),b(s(0)),\dot{\gamma}_{(y,w)}(s(0)-\tau_-(y,w))\big)$$

as $\varepsilon\to 0$, which in turn equals

$$E(\vec{\gamma}_{(y,w)}(\tau_+(y,w)),0,-\tau_+(y,w))E(\vec{\gamma}_{(y,w')}(\tau_-(y,w')),0,\tau_-(y,w'))k(y,w,w').$$

Finally, let $\chi(y,w,w')\in C_0^\infty(W)$. Let

$$G_-=\big\{(x',v')\in\Gamma_-\mid\phi_1(v')-\phi_1(\mathcal{P}(w';y,x'))\in\operatorname{supp}\psi_\epsilon^3,\ w'\in\operatorname{supp}\chi\big\},$$
$$G_+=\big\{(x,v)=\vec{\gamma}_{(y,w)}(\tau_+(y,w))\in\Gamma_+\mid(y,w)\in\operatorname{supp}\chi\big\}.$$

Then

$$\begin{aligned}(25)\quad &\int_M\int_{T_yM}\int_{T_yM}\Big|\int_{\Gamma_-}\psi_\rho^1(x')\,\psi_\varepsilon^2(h_2(x'))\,\psi_\epsilon^3\big(\phi_2(v')-\phi_2(\mathcal{P}(w';y,x'))\big)\\ &\qquad\cdot\chi(y,w,w')\alpha_2(\vec{\gamma}_{(y,w)}(\tau_+(y,w)),x',v')\,d\mu(x',v')\Big|\,dw'_y\,dw_y\,d\omega(y)\\ &\le\frac{1}{\varepsilon\epsilon}\int_{G_+}\int_0^{\tau_-(x,v)}\int_{T_{\gamma_{(x,v)}(s)}M}\int_{G_-}\psi_\rho^1(x')\chi(\vec{\gamma}_{(x,v)}(s),w')\alpha_2(x,v,x',v')\\ &\qquad\cdot d\mu(x',v')\,dw'_{\gamma_{(x,v)}(s)}\,ds\,d\mu(x,v)\\ &\le\frac{1}{\varepsilon\epsilon}\int_{G_-}\int_{G_+}\psi_\rho^1(x')C_\chi(x,v)\tau(x,v)\alpha_2(x,v,x',v')\,d\mu(x,v)\,d\mu(x',v')\end{aligned}$$

where

$$C_\chi(x,v)=\sup_{s\in[0,\tau(x,v)]}\int_{T_{\gamma_{(x,v)}(s)}M}\chi(\vec{\gamma}_{(x,v)}(s),w')\,dw'_{\gamma_{(x,v)}(s)}.$$

The integrand on the last line of (25) is an L^1 function since $\tau(x,v)C_\chi(x,v)$ is bounded for $(x,v)\in\operatorname{supp}C_\chi$, and $\alpha_2\in L^\infty(\Gamma_-;L^1(\Gamma_+,d\mu))$. Since the support of ψ_ρ^1 is a ρ-small neighborhood of a $3n$-dimensional variety in the $4n-2$-dimensional $\Gamma_+\times\Gamma_-$, the integral (25) tends to zero as $\rho\to 0$. This completes the proof of Proposition 4.3. □

Theorem 4.4. *If $M\subset\mathbb{R}^n$, $n\ge 3$, is a bounded domain with C^∞ boundary, g is a simple Riemannian metric on M, and assumptions* (3) *and* (4) *hold then from the kernel of $\mathcal{A}$ we may determine the collision kernel $k(x,v',v)$.*

Proof. From Theorem 4.2 we can determine σ_a from $\mathcal{A}$ and so determine $E(x,v,s,t)$; taking the limit as in Proposition 4.3, we can thus determine $k(x,v',v)$ for $v',v\in T_xM$ linearly independent. Of course k is an L^1 function and so knowing k on such a set of v',v is equivalent to knowing k. □

Acknowledgements. The author would like to thank Gunther Uhlmann for suggesting this problem and for several helpful discussions. We also thank Simon Arridge for the reference [**F**].

References

[A] S. Arridge, *Optical tomography in medical imaging*, Inverse Problems, **15** (1999), R41–R93, MR 1684463 (2000b:78023), Zbl 0926.35155.

[BG] I.N. Bernstein and M.L. Gerver, *Conditions of distinguishability of metrics by hodographs*, Methods and Algorithms of Interpretation of Seismological Information, Computerized Seismology, **13** (1980), Nauka, Moscow, 50–73 (in Russian).

[CS1] M. Choulli and P. Stefanov, *Inverse scattering and inverse boundary value problems for the linear Boltzmann equation*, Comm. Partial Differential Equations, **21**(**5-6**) (1996), 763–785, MR 1391523 (97f:35230), Zbl 0857.35131.

[CS2] M. Choulli and P. Stefanov, *An inverse boundary value problem for the stationary transport equation*, Osaka J. Math. **36**(**1**) (1999), 87–104, MR 1670750 (2000g:35228), Zbl 0998.35064.

[DU] J. Diestel and J.J. Uhl, Jr., *Vector Measures*, Mathematical Surveys and Monographs, **15**, American Mathematical Society, Providence, R.I., 1977, MR 0453964 (56 #12216), Zbl 0369.46039.

[F] H.A. Ferwerda, *The radiative transfer equation for scattering media with a spatially varying refractive index*, J. Opt. A: Pure Appl. Opt., **1** (1999), L1–L2.

[H] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces*, Pure and Applied Mathematics, **80**, Academic Press, New York, 1978, MR 0514561 (80k:53081), Zbl 0451.53038.

[KH] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Encyclopedia of Mathematics and its Applications, **54**, Cambridge University Press, Cambridge, 1995, MR 1326374 (96c:58055), Zbl 0878.58020.

[M] R.G. Mukhometov, *On a problem of reconstructing Riemannian metrics*, Siberian Math. J., **22**(**3**) (1982), 420–433, MR 0621466 (82m:53071), Zbl 0478.53048.

[RS] M. Reed and B. Simon, *Methods of Modern Mathematical Physics,* III: *Scattering Theory*, Academic Press, New York, 1979, MR 0529429 (80m:81085), Zbl 0405.47007.

[R] V.G. Romanov, *Stability estimates in problems of recovering the attenuation coefficient and the scattering indicatrix for the transport equation*, J. Inverse Ill-Posed Probl. **4**(**4**) (1996), 297–305, MR 1403885 (97f:35235), Zbl 0860.35146.

[Sh1] V.A. Sharafutdinov, *Integral Geometry of Tensor Fields*, Inverse and Ill-Posed Problems Series, VSP, The Netherlands, 1994, MR 1374572 (97h:53077), Zbl 0883.53004.

[Sh2] V.A. Sharafutdinov, *The inverse problem of determining the source in the stationary transport equation on a Riemannian manifold*, J. Math. Sci. (New York), **96**(**4**) (1999), 3430–3433, MR 1700653 (2000g:58056), Zbl 0935.58013.

[SU] P. Stefanov and G. Uhlmann, *Optical tomography in two dimensions*, Methods Appl. Anal., **10**(**1**) (2003), 1–9, MR 2014159 (2004h:35224).

[T1] A. Tamasan, *Optical tomography in weakly anisotropic scattering media*, in Inverse problems: theory and applications (Cortona/Pisa, 2002), Contemp. Math. **333**, Amer. Math. Soc., Providence, RI, 2003, 199–207, MR 2032017.

[T2] A. Tamasan, *An inverse boundary value problem in two-dimensional transport*, Inverse Problems, **18**(**1**) (2002), 209–219, MR 1893591 (2003h:35283), Zbl 0995.65146.

[W] J.-N. Wang, *Stability estimates of an inverse problem for the stationary transport equation*, Ann. Inst. H. Poincaré Phys. Théor., **70**(**5**) (1999), 473–495, MR 1697917 (2000e:82042), Zbl 0963.35204.

Received July 30, 2003 and revised April 2, 2004. This project was partially funded by NSF Grants DMS-9970503, DMS-0070488.

Department of Mathematics
Western Washington University
Bellingham WA 98225-9063
E-mail address: stephen.mcdowall@wwu.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

SUFFICIENT CONDITIONS FOR ROBUSTNESS OF ATTRACTORS

C.A. Morales and M.J. Pacifico

A recent problem in dynamics is to determine whether an attractor Λ of a C^r flow X is C^r robust transitive. By an *attractor* we mean a transitive set to which all positive orbits close to it converge. An attractor is C^r robust transitive (or C^r *robust* for short) if it has a neighborhood U such that the set $\bigcap_{t>0} Y_t(U)$ is transitive for every flow Y that is C^r close to X. We give sufficient conditions for robustness of attractors based on the following definitions: an attractor is *singular-hyperbolic* if it has singularities, all of which are hyperbolic, and is partially hyperbolic with volume expanding central direction (Morales, Pacifico and Pujals, 1998). An attractor is C^r *critically robust* if it has a neighborhood U such that $\bigcap_{t>0} Y_t(U)$ is in the closure of the closed orbits of every flow Y C^r close to X. We show that on compact 3-manifolds all C^r critically robust singular-hyperbolic attractors with only one singularity are C^r robust.

1. Introduction

A recent problem in dynamics is to determine whether an attractor Λ of a C^r flow X is C^r robust transitive or not. By an *attractor* we mean a transitive set to which all positive orbits close to it converge. An attractor is C^r robust transitive (or C^r *robust* for short) if it has a neighborhood U such that the set $\bigcap_{t>0} Y_t(U)$ is transitive for every flow Y that is C^r close to X. We give sufficient conditions for robustness of attractors based on the following definitions: an attractor is *singular-hyperbolic* if it has singularities, all of which are hyperbolic, and is partially hyperbolic with volume expanding central direction [**MPP**]. An attractor is C^r *critically robust* if it has a neighborhood U such that $\bigcap_{t>0} Y_t(U)$ is in the closure of the closed orbits is every flow Y C^r close to X. We show that, for flows on compact 3-manifolds, all C^r critically robust singular-hyperbolic attractors with only one singularity are C^r robust.

Let us state our result in a precise way. Hereafter X_t is a flow induced by a C^r vector field X on a compact 3-manifold M. The ω-limit set of $p \in M$ is the accumulation point set $\omega_X(p)$ of the positive orbit of p. An invariant set is *transitive* if it equals $\omega_X(p)$ for some point p on it.

Definition 1.1. A compact set in M is an *attracting set* of X if it can be written in the form $\bigcap_{t>0} X_t(U)$ for some neighborhood U. An *attractor* is a transitive attracting set.

See [**Mi**] where several definitions of attractors are discussed. The central definition of this paper is the following:

Definition 1.2. An attractor of a C^r flow X is C^r robust transitive (or C^r *robust* for short) if it has a neighborhood U such that $\bigcap_{t>0} Y_t(U)$ is a transitive set of Y for every flow Y that is C^r close to X.

Recently the problem of finding sufficient conditions for robustness of attractors was introduced in [**B**] and [**P**]. To deal with it we introduce the following definitions: a compact invariant set Λ of X is *partially hyperbolic* if there are an invariant splitting $T\Lambda = E^s \oplus E^c$ and positive constants K, λ such that:

1. E^s *is contracting*, namely
$$\|DX_t/E^s_x\| \le Ke^{-\lambda t}, \qquad \forall x \in \Lambda,\ \forall t > 0.$$
2. E^s *dominates* E^c, namely
$$\|DX_t/E^s_x\| \cdot \left\|DX_{-t}/E^c_{X_t(x)}\right\| \le Ke^{-\lambda t}, \qquad \forall x \in \Lambda,\ \forall t > 0.$$

The central direction E^c of Λ is said to be *volume expanding* if the additional condition
$$\left|J(DX_t/E^c_x)\right| \ge Ke^{\lambda t}$$
holds for all $x \in \Lambda$ and $t > 0$, where $J(\,\cdot\,)$ means the Jacobian.

Definition 1.3 ([**MPP**])**.** An attractor is *singular-hyperbolic* if it has singularities, all of which are hyperbolic, and is partially hyperbolic with volume expanding central direction.

The most representative example of a C^r robust singular-hyperbolic attractor is the geometric Lorenz attractor [**GW**]. The main result in [**MPP**] claims that C^1 robust nontrivial attractors on compact 3-manifolds are singular-hyperbolic. The converse is false: there are singular-hyperbolic attractors on compact 3-manifolds that are not C^r robust [**MPu**]. The following definition gives a further sufficient condition for robustness of singular-hyperbolic attractors:

Definition 1.4. An attractor of a C^r flow X is C^r *critically robust* if it has a neighborhood U such that $\bigcap_{t>0} Y_t(U)$ is in the closure of the closed orbits of Y, for every flow Y that is C^r close to X.

Hyperbolic attractors on compact manifolds are C^r robust and C^r critically robust for all r. The geometric Lorenz attractor [**GW**] is an example of a singular-hyperbolic attractor with only one singularity which is also C^r robust and C^r critically robust. In general singular-hyperbolic attractors

with only one singularity may be neither C^r robust nor C^r critically robust [**MPu**]. Nevertheless we shall prove that on compact 3-manifolds C^r critically robustness implies C^r robustness among singular-hyperbolic attractors with only one singularity. More precisely:

Theorem A. *A C^r critically robust singular-hyperbolic attractor with only one singularity on compact 3-manifolds is C^r robust.*

This gives explicit sufficient conditions for the robustness of attractors but they *depend on the perturbed flow*. E. Pujals is interested in conditions *depending on the unperturbed flow only*. It would also be interesting to determine whether the conclusion of Theorem A holds if we interchange the roles of robust and critically robust in the statement.

The proof of Theorem A relies on recent work [**MP2**]. We reproduce the necessary results in Section 2 for completeness. The proof of Theorem A is in Section 3.

2. Singular-hyperbolic attracting sets

In this section we describe the results in [**MP2**], omitting some proofs; see [**MP2**] for details. Hereafter X is a C^r flow on a closed 3-manifold M. The closure of B will be denoted by $\mathrm{Cl}(B)$. If A is a compact invariant set of X we denote by $\mathrm{Sing}_X(A)$ the set of singularities of X in A, and by $\mathrm{Per}_X(A)$ the union of the periodic orbits of X in A. A compact invariant set H of X is *hyperbolic* if the tangent bundle over H has an invariant decomposition $E^s \oplus E^X \oplus E^u$ such that E^s is contracting, E^u is expanding and E^X is generated by the direction of X [**PT**]. Stable Manifold Theory [**HPS**] asserts the existence of the stable manifold $W^s_X(p)$ and the unstable manifold $W^u_X(p)$ associated to $p \in H$. These manifolds are respectively tangent to the subspaces $E^s_p \oplus E^X_p$ and $E^X_p \oplus E^u_p$ of T_pM. In particular, $W^s_X(p)$ and $W^u_X(p)$ are well-defined if p belongs to a hyperbolic periodic orbit of X. If O is an orbit of X we write $W^s_X(O) = W^s_X(p)$ and $W^u_X(O) = W^u_X(p)$ for some $p \in O$. We observe that $W^{s(u)}_X(O)$ does not depend on $p \in O$. When $\dim E^s = \dim E^u = 1$ we say that H is of *saddle type*. In this case $W^s_X(p)$ and $W^u_X(p)$ are two-dimensional submanifolds of M. The maps $p \in H \to W^s_X(p)$ and $p \in H \to W^u_X(p)$ are continuous (on compact parts). Moreover, a compact, singular, invariant set Λ of X is *singular-hyperbolic* if all its singularities are hyperbolic and the tangent bundle over Λ has an invariant decomposition $E^s \oplus E^c$ such that E^s is contracting, E^s dominates E^c and E^c is volume expanding (i.e., the Jacobian of DX_t/E^c grows exponentially as $t \to \infty$). Again, Stable Manifold Theory asserts the existence of the strong stable manifold $W^{ss}_X(p)$ associated to $p \in \Lambda$. This manifold is tangent to the subspace E^s_p of T_pM. For all $p \in \Lambda$ we define $W^s_X(p) = \bigcup_{t\in\mathbf{R}} W^{ss}_X(X_t(p))$. If p is regular (i.e., $X(p) \neq 0$) then $W^s_X(p)$ is a

well-defined two-dimensional submanifold of M. The map $p \in \Lambda \to W^s_X(p)$ is continuous (on compact parts) at the regular points p of Λ. A singularity σ of X is *Lorenz-like* if its eigenvalues $\lambda_1, \lambda_2, \lambda_3$ are real and satisfy

$$\lambda_2 < \lambda_3 < 0 < -\lambda_3 < \lambda_1$$

up to some reordering of the eigenvalues. A Lorenz-like singularity σ is hyperbolic, so $W^s_X(\sigma)$ and $W^u_X(\sigma)$ do exist. Moreover, the eigenspace of λ_2 is tangent to a one-dimensional invariant manifold $W^{ss}_X(\sigma)$. This manifold is called the *strong stable manifold* of σ. Clearly $W^{ss}_X(\sigma)$ splits $W^s_X(\sigma)$ into two connected components. We denote by $W^{s,+}_X(\sigma)$ and $W^{s,-}_X(\sigma)$ the two connected components of $W^s_X(\sigma) \setminus W^{ss}_X(\sigma)$.

Let Λ be a singular-hyperbolic set with dense periodic orbits of a three-dimensional flow. It follows from [**MPP**] that every $\sigma \in \mathrm{Sing}_X(\Lambda)$ is Lorenz-like and satisfies $\Lambda \cap W^{ss}_X(\sigma) = \{\sigma\}$. It follows also from [**MPP**] that any compact invariant subset without singularities of Λ is hyperbolic of saddle type. If in addition Λ is attracting, there is for every $p \in \mathrm{Per}_X(\Lambda)$ a $\sigma \in \mathrm{Sing}_X(\Lambda)$ such that

$$W^u_X(p) \cap W^s_X(\sigma) \neq \emptyset.$$

This follows from the methods in [**MP1**].

For every singular-hyperbolic set Λ of a three-dimensional flow X and every Lorenz-like singularity $\sigma \in \mathrm{Sing}_X(\Lambda)$ we define

$$\begin{aligned} P^+ &= \{p \in \mathrm{Per}_X(\Lambda) : W^u_X(p) \cap W^{s,+}_X(\sigma) \neq \emptyset\}, \\ P^- &= \{p \in \mathrm{Per}_X(\Lambda) : W^u_X(p) \cap W^{s,-}_X(\sigma) \neq \emptyset\}, \\ H^+_X &= \mathrm{Cl}(P^+), \\ H^-_X &= \mathrm{Cl}(P^-). \end{aligned}$$

These sets will play an important role.

Lemma 2.1. *Let Λ be a connected, singular-hyperbolic, attracting set with dense periodic orbits and only one singularity σ. Then $\Lambda = H^+_X \cup H^-_X$.*

Next we state a technical lemma to be used later. If S is a submanifold we denote by T_xS the tangent space at $x \in S$. A *cross-section* of X is a compact submanifold Σ transverse to X and diffeomorphic to the two-dimensional square $[0,1]^2$. If Λ is a singular-hyperbolic set of X and $x \in \Sigma \cap \Lambda$, then x is regular and so $W^s_X(x)$ is a two-dimensional submanifold transverse to Σ. In this case we denote by $W^s_X(x, \Sigma)$ the connected component of $W^s_X(x) \cap \Sigma$ containing x. We shall be interested in a special cross-section described as follows: let Λ be a singular-hyperbolic set of a three-dimensional flow X and let $\sigma \in \mathrm{Sing}_X(\Lambda)$. Suppose that the closed orbits contained in Λ are dense in Λ. Then σ is Lorenz-like [**MPP**], and therefore one can describe the flow using the Grobman–Hartman Theorem [**dMP**]. Indeed, we can assume that the flow of X around σ is the linear flow $\lambda_1\partial_{x_1} + \lambda_2\partial_{x_2} + \lambda_3\partial_{x_3}$

in a suitable coordinate system $(x_1, x_2, x_3) \in [-1, 1]^3$ around $\sigma = (0, 0, 0)$. A cross-section Σ of X is *singular* if it corresponds to the submanifolds $\Sigma^+ = \{x_3 = 1\}$ or $\Sigma^- = \{x_3 = -1\}$ in the coordinate system (x_1, x_2, x_3). We denote by l^+ and l^- the curves obtained by intersecting $\{x_2 = 0\}$ with Σ^+ and Σ^-, respectively; these curves are contained in $W_X^{s,+}(\sigma)$ and $W_X^{s,-}(\sigma)$ respectively. We state without proof the following straightforward lemma:

Lemma 2.2. *Let Λ a singular-hyperbolic set with dense periodic orbits of a three-dimensional flow X, and fix $\sigma \in \mathrm{Sing}_X(\Lambda)$. There are singular cross-sections Σ^+, Σ^- as above such that every orbit of Λ passing close to some point in $W_X^{s,+}(\sigma)$ (resp. $W_X^{s,-}(\sigma)$) intersects Σ^+ (resp. Σ^-). If $p \in \Lambda \cap \Sigma^+$ is close to l^+, then $W_X^s(p, \Sigma^+)$ is a vertical curve crossing Σ^+. If $p \in \mathrm{Per}_X(\Lambda)$ and $W_X^u(p) \cap W_X^{s,+}(\sigma) \neq \emptyset$, then $W_X^u(p)$ contains an interval $J = J_p$ intersecting l^+ transversally; and the same is true if we replace $+$ by $-$.*

To prove transitivity we shall use two lemmas:

Lemma 2.3 (Birkhoff's criterion)**.** *Let T be a compact, invariant set of X such that for all open sets U, V intersecting T there is $s > 0$ such that $X_s(U \cap T) \cap V \neq \emptyset$. Then T is transitive.*

Lemma 2.4. *Let Λ be a connected, singular-hyperbolic, attracting set with dense periodic orbits and only one singularity σ. Let U, V be open sets, $p \in U \cap \mathrm{Per}_X(\Lambda)$ and $q \in V \cap \mathrm{Per}_X(\Lambda)$. If $W_X^u(p) \cap W_X^{s,+}(\sigma) \neq \emptyset$ and $W_X^u(q) \cap W_X^{s,+}(\sigma) \neq \emptyset$, there exist $t > 0$ and $z \in W_X^u(p)$ arbitrarily close to $W_X^u(p) \cap W_X^{s,+}(\sigma)$ such that $X_t(z) \in V$. The same is true if we replace $+$ by $-$.*

Let Λ be a singular-hyperbolic set of $X \in \mathcal{X}^r$ satisfying:

1. Λ is connected.
2. Λ is attracting.
3. The closed orbits contained in Λ are dense in Λ.
4. Λ has only one singularity σ.

We note that condition 3 implies

(H1) $\qquad \Lambda = \mathrm{Cl}(\mathrm{Per}_X(\Lambda))$.

Proposition 2.5. *Suppose that, for any given $p, q \in \mathrm{Per}_X(\Lambda)$, either*

1. *$W_X^u(p) \cap W_X^{s,+}(\sigma) \neq \emptyset$ and $W_X^u(q) \cap W_X^{s,+}(\sigma) \neq \emptyset$, or*
2. *$W_X^u(p) \cap W_X^{s,-}(\sigma) \neq \emptyset$ and $W_X^u(q) \cap W_X^{s,-}(\sigma) \neq \emptyset$.*

Then Λ is transitive.

Proof. By Birkhoff's criterion we only need prove that for all open sets U, V intersecting Λ there exists $s > 0$ such that $X_s(U \cap \Lambda) \cap V \neq \emptyset$. For this we proceed as follows: by (H1) there are $p \in \mathrm{Per}_X(\Lambda) \cap U$ and $q \in \mathrm{Per}_X(\Lambda) \cap V$.

First suppose that alternative 1 holds. By Lemma 2.4, there are $z \in W^u_X(p)$ and $t > 0$ such that $X_t(z) \in V$. Since $z \in W^u_X(p)$, we have $w = X_{-t'}(z) \in U$ for some $t' > 0$. Since Λ is an attracting set, $w \in \Lambda$. If $s = t + t' > 0$ we conclude that $w \in (U \cap \Lambda) \cap X_{-s}(V)$ and so $X_s(U \cap \Lambda) \cap V \neq \emptyset$. If alternative 2 of the corollary holds we can find $s > 0$ such that $X_s(U \cap \Lambda) \cap V \neq \emptyset$ in a similar way (replacing + by −). The result follows. □

Proposition 2.6. *If there is a sequence $p_n \in \mathrm{Per}_X(\Lambda)$ converging to some point in $W^{s,+}_X(\sigma)$ such that $W^u_X(p_n) \cap W^{s,-}_X(\sigma) \neq \emptyset$ for all n, then Λ is transitive. The same is true interchanging + and −.*

Proof. Let $p, q \in \mathrm{Per}_X(\Lambda)$ be fixed. Suppose that $W^u_X(p) \cap W^{s,+}_X(\sigma) \neq \emptyset$ and $W^u_X(q) \cap W^{u,-}_X(\sigma) \neq \emptyset$. By Lemma 2.2 we can fix a cross-section $\Sigma = \Sigma^+$ through $W^{s,+}_X(\sigma)$ and an open arc $J \subset \Sigma \cap W^u_X(p)$ intersecting $W^{s,+}_X(\sigma)$ transversally. Again by Lemma 2.2 we can assume that $p_n \in \Sigma$ for every n. Because the direction E^s of Λ is contracting, the size of $W^s_X(p_n)$ is uniformly bounded away from zero. It follows that there is n large so that J intersects $W^s_X(p_n)$ transversally. Applying the Inclination Lemma [**dMP**] to the saturation of $J \subset W^u_X(p)$, and the assumption $W^u_X(p_n) \cap W^{s,-}_X(\sigma) \neq \emptyset$, we conclude that $W^u_X(p) \cap W^{s,-}_X(\sigma) \neq \emptyset$. So alternative 2 of Proposition 2.5 holds; it follows from that proposition that Λ is transitive. □

Proposition 2.7. *If there is $a \in W^u_X(\sigma) \setminus \{\sigma\}$ such that $\sigma \in \omega_X(a)$, then Λ is transitive.*

Proof. Without loss of generality we can assume that there exists z in $\omega_X(a) \cap W^+_X(\sigma)$. If $W^u_X(q) \cap W^{s,-}_X(\sigma) = \emptyset$ for all $q \in \mathrm{Per}_X(\Lambda)$, then $W^u_X(q) \cap W^{s,+}_X(\sigma) \neq \emptyset$ for all $q \in \mathrm{Per}_X(\Lambda)$; see [**MP1**]. Then Λ is transitive by Proposition 2.5 since alternative 1 holds for all $p, q \in \mathrm{Per}_X(\Lambda)$. So we can assume that there is $q \in \mathrm{Per}_X(\Lambda)$ such that $W^u_X(q) \cap W^{s,-}_X(\sigma) \neq \emptyset$. It follows from the dominating condition of the singular-hyperbolic splitting of Λ that the intersection $W^u_X(q) \cap W^{s,-}_X(\sigma)$ is transversal. This allows us to choose a point in $W^u_X(q)$ arbitrarily close to $W^{s,-}_X(\sigma)$ on the side of $W^u_X(q) \cap W^{s,-}_X(\sigma)$ accumulating a. Since Λ is attracting and satisfies (H1), one can find a sequence $p_n \in \mathrm{Per}_X(\Lambda)$ converging to $z \in W^{s,+}_X(\sigma)$ such that for all n there is p'_n in the orbit of p_n such that the sequence p'_n converges to some point in $W^{s,-}_X(\sigma)$. Now suppose for a contradiction that Λ is not transitive. Then Proposition 2.6 implies

$$W^u_X(p'_n) \cap W^{s,+}_X(\sigma) = \emptyset \quad \text{and} \quad W^u_X(p_n) \cap W^{s,-}_X(\sigma) = \emptyset$$

for n large. But $W^u_X(p_n) = W^u_X(p'_n)$ since p'_n and p_n are in the same orbit of X. So $W^u_X(p_n) \cap (W^{s,+}_X(\sigma) \cup W^{s,-}_X(\sigma)) = \emptyset$. However

$$W^u_X(p_n) \cap W^s_X(\sigma) = \emptyset,$$

a contradiction since $\mathrm{Sing}_X(\Lambda) = \{\sigma\}$. We conclude that Λ is transitive. □

Theorem 2.8. *If Λ is not transitive, there is for all $a \in W^u_X(\sigma)\setminus\{\sigma\}$ a periodic orbit O of X with positive expanding eigenvalues such that $a \in W^s_X(O)$.*

Proof. Fix $a \in W^u_X(\sigma)\setminus\{\sigma\}$, and assume that $\omega_X(a)$ is *not* a periodic orbit. We will obtain a contradiction once we prove that if $p, q \in \mathrm{Per}_X(\Lambda)$ then p, q satisfy one of the two alternatives in Proposition 2.5. To prove this we proceed as follows: as noted before, both $W^u_X(p)$ and $W^u_X(q)$ intersect $W^s_X(\sigma)$ (see [**MP1**]). Then we can assume

$$W^u_X(p) \cap W^{s,+}_X(\sigma) \neq \emptyset \quad \text{and} \quad W^u_X(q) \cap W^{s,-}_X(\sigma) \neq \emptyset.$$

By using this and the linear coordinate around σ, it is easy to construct an open interval $I = I_a$, contained in a suitable cross-section $\Sigma = \Sigma_a$ of X containing a, and such that $I\setminus\{a\}$ is formed by two intervals $I^+ \subset W^u_X(p)$ and $I^- \subset W^u_X(q)$. Observe that the tangent vector of I is contained in $E^c \cap T\Sigma_a$. Proposition 2.7 implies that $\sigma \notin \omega_X(a)$, since Λ is not transitive. It follows that $H = \omega_X(a)$ is a hyperbolic set of saddle type; see [**MPP**]. As in [**M**] one proves that H is one-dimensional, so Bowen's Theory of hyperbolic one-dimensional sets [**Bw**] applies. In particular we can choose a family of cross-sections $\mathcal{S} = \{S_1, \ldots, S_r\}$ of small diameter such that H is the flow-saturate of $H \cap \operatorname{int}\mathcal{S}'$, where $\mathcal{S}' = \bigcup S_i$ and $\operatorname{int}\mathcal{S}'$ is the interior of $\mathcal{S}'$. Also, $I \subset \Lambda$ since Λ is attracting. Recall that the tangent direction of I is contained in E^c. Since E^c is volume expanding and H is nonsingular, the Poincaré map induced by X on $\mathcal{S}'$ is expanding along I. As in [**MP1**, p. 371] we can find $\delta > 0$ and a open arc sequence $J_n \subset \mathcal{S}'$ in the positive orbit of I with length bounded away from 0 such that there is a_n *in the positive orbit of* a contained in the interior of J_n. We can fix $S = S_i \in \mathcal{S}$ in order to assume that $J_n \subset S$ for every n. Let $x \in S$ be a limit point of a_n. Then $x \in H \cap \operatorname{int}\mathcal{S}'$. Because I is tangent to E^c, the interval sequence J_n converges to an interval $J \subset W^u_X(x)$ in the C^1 topology ($W^u_X(x)$ exists since $x \in H$ and H is hyperbolic). J is not trivial since the length of J_n is bounded away from 0. If a_n were in $W^s_X(x)$ for n large we would conclude that x is periodic by [**MP1**, Lemma 5.6], a contradiction since $\omega_X(a)$ is not periodic. We conclude that for infinitely many values of n, $a_n \notin W^s_X(x)$. Since $J_n \to J$ and Λ has strong stable manifolds of uniformly size, there exists

$$z_n \in \big(W^s_X(a_{n+1}) \cap S\big) \cap \big(J_n \setminus \{a_n\}\big)$$

for all n large. For every n let J^+_n and J^-_n be the two connected components of $J_n\setminus\{a_n\}$, with J^+_n in the positive orbit of I^+ and J^-_n in the positive orbit of I^-. Clearly, either $z_n \in J^+_n$ or $z_n \in J^-_n$.

If $z_n \in J^+_n$ there is $v_{n+1} \in \mathrm{Per}_X(\Lambda) \cap S$ close to a_{n+1} such that

$$W^s_X(v_{n+1}) \cap J^+_n \neq \emptyset \quad \text{and} \quad W^s_X(v_{n+1}) \cap J^-_{n+1} \neq \emptyset.$$

Since v_{n+1} is periodic, it follows from [**MPP**] that $W^u_X(v_{n+1})$ intersects $W^{s,+}_X(\sigma)$ or $W^{s,-}_X(\sigma)$. The choice of v_{n+1} implies that its orbit passes close

to a point in $W_X^{s,-}(\sigma)$. Since Λ is not transitive we conclude that $W_X^u(v_{n+1})$ intersects $W_X^{s,-}(\sigma)$. Since $W_X^s(v_{n+1}) \cap J_n^+$ is transversal, the Inclination Lemma then implies $W_X^u(p) \cap W_X^{s,-}(\sigma) \neq \emptyset$. Hence

$$W_X^u(p) \cap W_X^{s,-}(\sigma) \neq \emptyset \quad \text{and} \quad W_X^u(q) \cap W_X^{s,-}(\sigma) \neq \emptyset.$$

If $z_n \in J_n^-$ we can prove by similar arguments that

$$W_X^u(p) \cap W_X^{s,+}(\sigma) \neq \emptyset \quad \text{and} \quad W_X^u(q) \cap W_X^{s,+}(\sigma) \neq \emptyset.$$

These alternatives yield the desired contradiction. Therefore $\omega_X(a) = O$ for some periodic orbit O of X. To finish we prove that the expanding eigenvalue of O is positive. Suppose by contradiction that it is not. Fix a cross-section Σ intersecting O in a single point p_0. This section defines a Poincaré map $\Pi : \operatorname{Dom}\Pi \subset \Sigma \to \Sigma$ of which p_0 is a hyperbolic fixed-point. The assumption implies that $D\Pi(p_0)$ has negative expanding eigenvalue. Because $p_0 \in \operatorname{Per}_X(\Lambda)$, it follows from [**MPP**] that $W_X^u(p_0)$ intersects $W_X^{s,+}(\sigma)$ or $W_X^{s,-}(\sigma)$. We shall assume the former case since the proof in the latter is similar. We claim that $W_X^u(p) \cap W_X^{s,+}(\sigma) \neq \emptyset$ for all $p \in \operatorname{Per}_X(\Lambda)$. Indeed, let $p \in \operatorname{Per}_X(\Lambda)$ be fixed. Again $W_X^u(p)$ intersects $W_X^{s,+}(\sigma)$ or $W_X^{s,-}(\sigma)$. In the first case we are done. So we can assume that $W_X^u(p) \cap W_X^{s,-}(\sigma) \neq \emptyset$. By flow-saturating this intersection we obtain an interval $K \subset W_X^u(p) \cap \Sigma$ intersecting $W_X^s(p_0, \Sigma)$ transversally. At the same time, there is an interval $J \subset W_X^{s,+}(\sigma) \cap \Sigma$ intersecting $W_X^u(p_0, \Sigma)$ transversally. Since the expanding eigenvalue of $D\Pi(p_0)$ is negative the Inclination Lemma implies that the backward iterates $\Pi^{-n}(J)$ of J accumulate on $W_X^s(p_0, \Sigma)$ *in both sides*. Because K has transversal intersection with $W_X^s(p_0\Sigma)$ we conclude that one such backward iterate intersects K, and this yields $W_X^u(p) \cap W_X^{s,+}(\sigma) \neq \emptyset$ as desired, proving the claim. The claim together with Proposition 2.5 implies that Λ is transitive, yielding the contradiction needed to complete the proof of theorem. □

Hereafter we shall assume that Λ is not transitive. Let $a \in W_X^s(\sigma) \setminus \{\sigma\}$ be fixed. By Theorem 2.8, $a \in W_X^s(O)$ for some periodic orbit O with positive expanding eigenvalue. This last property implies that the unstable manifold $W_X^u(O)$ of O is a cylinder with generating curve O. Then O separates $W_X^u(O)$ into two connected components, which we denote by $W^{u,+}, W^{u,-}$ according to the following convention (see Figure 1): there is an interval $I = I_a$, contained in a suitable cross-section of X and containing a, such that if I^+, I^- are the connected components of $I \setminus \{a\}$ then $I^+ \subset W_X^u(p)$ and $I^- \subset W_X^u(q)$ for some periodic points $p, q \in \Lambda$ (recall that Λ is not transitive). In addition I is tangent to the central direction E^c of Λ (see Figure 1). Since $a \in W_X^s(O)$ and I is tangent to E^c, the flow of X carries I to an interval I' transverse to $W_X^s(O)$ at a. Note that the flow carries I^+ and I^- into I_0^+ and I_0^- respectively.

Definition 2.9. We denote by $W^{u,+}$ the connected component of $W^u \setminus O$ that is accumulated (via the Inclination Lemma and the Strong λ Lemma [**dMP**, **D**]) by the positive orbit of I_0^+. We denote $W^{u,-}$ the connected component of $W_X^u(O) \setminus O$ accumulated by the positive orbit of I_0^-.

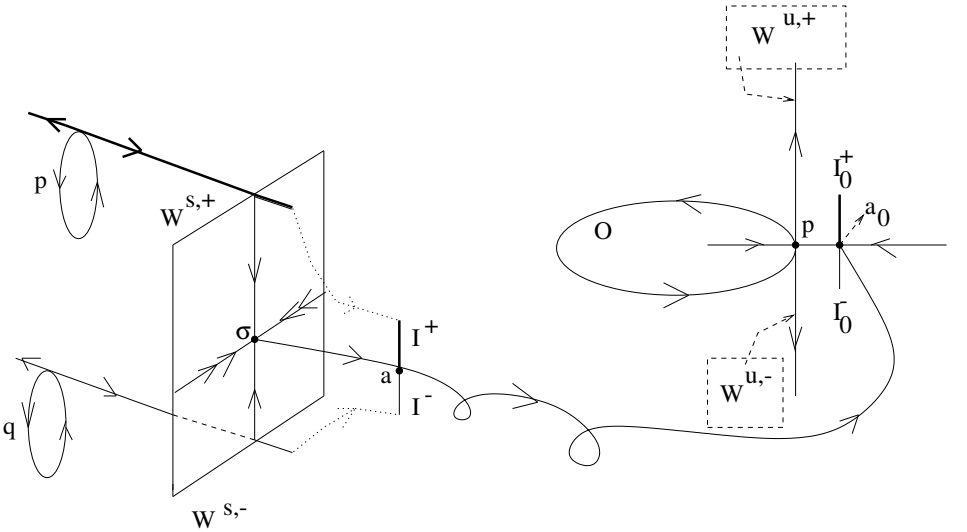


Figure 1. Definition of $W^{u,+}$ and $W^{u,-}$.

It can easily be proved using the Strong λ Lemma [**D**] that this definition does not depend on p, q, J_p, J_q.

Proposition 2.10. *$W^{u,+} \cap W_X^{s,-}(\sigma) = \emptyset$ and $W^{u,+} \cap W_X^{s,+}(\sigma) \neq \emptyset$. The same is true interchanging $+$ and $-$.*

Proof. For simplicity set $W = W^{u,+}$. First we prove that $W \cap W_X^{s,-}(\sigma) = \emptyset$. Suppose for a contradiction that $W \cap W_X^{s,-}(\sigma) \neq \emptyset$. Since this last intersection is transversal, there is an interval $J \subset W_X^{s,-}(\sigma)$ intersecting W transversally. Now, fix a cross-section $\Sigma = \Sigma^+$ as in Lemma 2.2 and let $p \in \mathrm{Per}_X(\Lambda)$ be such that $W_X^u(p) \cap W_X^{s,+}(\sigma) \neq \emptyset$. Then there is an small interval $I \subset W_X^u(p) \cap \Sigma$ transversal to $\Sigma \cap W_X^{s,+}(\sigma)$. By the definition of $W = W^{u,+}$ (Definition 2.9), the positive orbit of I accumulates on W. Since J is transversal to W the Inclination Lemma implies that the positive orbit of I intersects J. This proves $W_X^u(p) \cap W_X^{s,-}(\sigma) \neq \emptyset$ for all $p \in \mathrm{Per}_X(\Lambda)$. It follows that alternative 2 of Proposition 2.5 holds for all p, q, which contradicts the nontransitivity of Λ and proves that $W \cap W_X^{s,-}(\sigma) = \emptyset$, as desired. Now suppose for a contradiction that $W \cap W_X^{s,+}(\sigma) = \emptyset$. Since $W \cap W_X^{s,-}(\sigma) = \emptyset$ we obtain $W \cap W_X^s(\sigma) = \emptyset$ (see [**MPP**]). But the denseness of periodic orbits and the Inclination Lemma imply that $W \cap W_X^s(\sigma) \neq \emptyset$.

This is a contradiction, which proves that $W \cap W_X^{s,+}(\sigma) \neq \emptyset$. The result follows. □

Proposition 2.11. *$H^+ = \mathrm{Cl}(W^{u,+})$ and $H^- = \mathrm{Cl}(W^{u,-})$.*

Proof. Fix $q \in P^+$, i.e., $W_X^u(q) \cap W_X^{s,+}(\sigma) \neq \emptyset$. Note that $W^{u,+} \cap W_X^{s,+}(\sigma)$ is nonempty by Lemma 2.10. Using (H1) and the Inclination Lemma it is not hard to prove that $W^{u,+}$ accumulates on q; therefore $H^+ \subset \mathrm{Cl}(W^{u,+})$. Conversely let $x \in W^{u,+}$ be fixed. By (H1) and $W^{u,+} \subset \Lambda$ there is z in $\mathrm{Per}_X(\Lambda)$ near x. Choosing z close to x we ensure that $W_X^s(z) \cap W^{u,+} \neq \emptyset$, because stable manifolds have size uniformly bounded away from zero. If $W_X^u(z) \cap W_X^{s,-}(\sigma) \neq \emptyset$, the Inclination Lemma and the fact that $W_X^s(z)$ intersects $W^{u,+}$ imply that $W^u \cap W_X^{s,-}(\sigma) \neq \emptyset$. This contradicts Proposition 2.10, so $W_X^u(z) \cap W_X^{s,-}(\sigma) = \emptyset$. By [**MP1**] we obtain $z \in P^+$, proving that $x \in H^+$ and the lemma. □

Proposition 2.12. *If $z \in \mathrm{Per}_X(\Lambda)$ and $W_X^s(z) \cap W^{u,+} \neq \emptyset$, then*

$$\mathrm{Cl}(W_X^s(z) \cap W^{u,+}) = \mathrm{Cl}(W^{u,+}).$$

The same is true if we replace + by −.

Proof. We have shown that $\mathrm{Cl}(W_X^{s,+}(\sigma) \cap W^{u,+}) = \mathrm{Cl}(W^{u,+})$. Fix $x \in W^{u,+}$. By (H1) there is $w \in \mathrm{Per}_X(\Lambda)$ close to x. In particular $W_X^s(w)$ intersects $W^{u,+}$. If $W_X^u(w) \cap W_X^{s,+}(\sigma) = \emptyset$ then $W_X^u(w) \cap W_X^{s,-}(\sigma) \neq \emptyset$ by [**MP1**]. It follows from the Inclination Lemma that $W^{u,+} \cap W_X^{s,-}(\sigma) \neq \emptyset$, contradicting Proposition 2.10. We conclude that $W_X^u(w) \cap W_X^{s,+}(\sigma) \neq \emptyset$. Note that $W_X^s(w) \cap W^{u,+} \neq \emptyset$ gets close to x as $w \to x$. Since the intersection $W_X^u(w) \cap W_X^{s,+}(\sigma) \neq \emptyset$ is transversal we can apply the Inclination Lemma to find a transverse intersection $W^{u,+} \cap W_X^{s,+}(\sigma)$ close to x. This proves $\mathrm{Cl}(W_X^{s,+}(\sigma) \cap W^{u,+}) = \mathrm{Cl}(W^{u,+})$. Finally we prove $\mathrm{Cl}(W_X^s(z) \cap W^{u,+}) = \mathrm{Cl}(W^{u,+})$. Choose $x \in W^{u,+}$. Since $\mathrm{Cl}(W_X^{s,+}(\sigma) \cap W^{u,+}) = \mathrm{Cl}(W^{u,+})$ there is an interval $I_x \subset W^{u,+}$ arbitrarily close to x such that $I_x \cap W_X^{s,+}(\sigma) \neq \emptyset$. The positive orbit of I_x first passes through a and then accumulates on $W^{u,+}$. But $W_X^s(z) \cap W^{u,+} \neq \emptyset$ by assumption. Since this intersection is transversal the Inclination Lemma implies that the positive orbit of I_x intersects $W_X^s(z)$. By taking the backward flow of the last intersection we get $W_X^s(z) \cap I_x \neq \emptyset$. This proves the lemma. □

Given $z \in \mathrm{Per}_X(\Lambda)$, let $H_X(z)$ be the homoclinic class associated to z.

Proposition 2.13. *If $z \in \mathrm{Per}_X(\Lambda)$ is close to a point in $W^{u,+}$, then*

$$H_X(z) = \mathrm{Cl}(W^{u,+}).$$

The same is true if we replace + by −.

Proof. Let $z \in \mathrm{Per}_X(\Lambda)$ be a point close to one in $W^{u,+}$. It follows from the continuity of the stable manifolds that $W^s_X(z) \cap W^{u,+} \neq \emptyset$. We claim that $H_X(z) = \mathrm{Cl}(W^{u,+})$. Indeed $W^{u,+} \cap W^{s,-}_X(\sigma) = \emptyset$ by Proposition 2.10. This equality and the Inclination Lemma imply that $W^u_X(z) \cap W^s_X(\sigma) \subset W^{s,+}_X(\sigma)$. By Proposition 2.12, $W^s_X(z) \cap W^{u,+}$ is dense in $W^{u,+}$ since $W^s_X(z) \cap W^{u,+} \neq \emptyset$. Let Σ be a cross-section containing p_0 and fix $x \in W^{u,+}$. We can assume $x, z \in \Sigma$. Since $W^u_X(z) \cap W^s_X(\sigma) \neq \emptyset$ and $W^u_X(z) \cap W^s_X(\sigma) \subset W^{s,+}_X(\sigma)$ there is an interval $I \subset W^u_X(z)$ intersecting $W^{s,+}_X(\sigma)$. Then the positive orbit of I yields an interval J close to σ intersecting $W^{s,+}_X(\sigma)$. In addition, the positive orbit of J yields an interval K whose positive orbit *accumulates* $W^{u,+}$ (recall Definition 2.9). Since $W^s_X(z) \cap W^{u,+}$ is dense in $W^{u,+}$ and $x \in W^{u,+}$, the orbit $W^s_X(z)$ passes close to x. The Inclination Lemma applied to the positive orbit of K yields a homoclinic point z' associated to z which is close to x. This proves that $x \in H_X(z)$, so $\mathrm{Cl}(W^{u,+}) \subset H_X(z)$. The opposite inclusion is a direct consequence of the Inclination Lemma applied to $W^s_X(z) \cap W^{u,+} \neq \emptyset$. We conclude that $\mathrm{Cl}(W^{u,+}) = H_X(z)$ as desired. □

Theorem 2.14. *Let Λ be a singular-hyperbolic set of a C^r flow X on a closed three-manifold, where $r \geq 1$. Suppose that the following properties hold:*

1. *Λ is connected.*
2. *Λ is attracting.*
3. *The closed orbits contained in Λ are dense in Λ.*
4. *Λ has a unique singularity σ.*
5. *Λ is not transitive.*

Then H^+ and H^- are homoclinic classes of X.

Proof. Let Λ be a singular-hyperbolic set of X satisfying the theorem's conditions. To prove that H^+ is a homoclinic class it suffices by Proposition 2.11 to prove that $\mathrm{Cl}(W^{u,+})$ is a homoclinic class. By condition 3 of the Theorem we can choose $z \in \mathrm{Per}_X(\Lambda)$ arbitrarily close to a point in $W^{u,+}$. Then $\mathrm{Cl}(W^{u,+}) = H_X(z)$ by Proposition 2.13 and the result follows. □

3. Proof of Theorem A

First we introduce some notations. Hereafter M is a compact 3-manifold and $\mathcal{X}^r$ is the space of C^r flows in M equipped with the C^r topology, $r \geq 1$. The *nonwandering set* of $X \in \mathcal{X}^r$ is the set $\Omega(X)$ of points $p \in M$ such that for all neighborhood U of p and $T > 0$ there is $t > T$ such that $X_t(U) \cap U \neq \emptyset$. An attracting Λ with isolating block U has a continuation $\Lambda(Y)$ for Y C^r close to X defined by $\Lambda(Y) = \bigcap_{t>0} Y_t(U)$. This continuation is then defined when Λ is an attractor. A compact invariant set is *nontrivial* if it is not a

closed orbit of X. Transitive sets for flows are always connected. The proof of Theorem A is based on the following result:

Theorem 3.1. *Let Λ be a singular-hyperbolic set of $X \in \mathcal{X}^r$, $r \geq 1$. Suppose that the following properties hold:*

1. *Λ is connected.*
2. *Λ is attracting.*
3. *The closed orbits contained in Λ are dense in Λ.*
4. *Λ has only one singularity.*
5. *Λ is not transitive.*

Then for every neighborhood U of Λ there is a flow Y that is C^r close to X and such that

$$\Lambda(Y) \not\subset \Omega(Y).$$

To prove this theorem we shall use the following definitions and facts: let $X \in \mathcal{X}^r$ and let Λ be a singular-hyperbolic set of X satisfying the conditions of the theorem. Let σ be the unique singularity of Λ. As mentioned on page 330, σ is Lorenz-like. As in Section 2, $W^{ss}_X(\sigma)$ divides $W^s_X(\sigma)$ into two connected components, which we denote by $W^{s,+}_X(\sigma)$ and $W^{s,-}_X(\sigma)$, or $W^{s,+}, W^{s,-}$ for short. Recall that $\mathrm{Per}_X(\Lambda)$ denotes the union of the periodic orbits of X in Λ. Fix such $a \in W^u_X(\sigma)\setminus\{\sigma\}$. By Theorem 2.8, $\omega_X(a) = O$ for some periodic orbit with positive expanding eigenvalues of X. In particular, $W^{u,+}$ and $W^{u,-}$ are defined (Definition 2.9).

Lemma 3.2. $\mathrm{Cl}(W^{u,+}) \cap W^{s,-} = \emptyset$.

Proof. Suppose for a contradiction that $\mathrm{Cl}(W^{u,+}) \cap W^{s,-} \neq \emptyset$. By Lemma 2.2 there is a singular cross-section Σ^- such that every orbit of Λ passing close to some point in $W^{s,-}_X(\sigma)$ intersects Σ^-. Let $q \in \Lambda$ be periodic such that $W^u_X(q) \cap W^{s,-}_X(\sigma) \neq \emptyset$. Since $\mathrm{Cl}(W^{u,+}) \cap W^{s,-} \neq \emptyset$, we have $\mathrm{Cl}(W^{u,+}) \cap \Sigma^- \neq \emptyset$. Because closed orbits are dense we can prove that $q \in \mathrm{Cl}(W^{u,+})$. It follows that $H^- \subset \mathrm{Cl}(W^{u,+})$, so $\Lambda = \mathrm{Cl}(W^{u,+})$ by Lemma 2.1. Also, since Λ is not transitive, $H^+ = \mathrm{Cl}(W^{u,+})$ (Proposition 2.11) and H^+ is a homoclinic class (Theorem 2.14). Since homoclinic classes are transitive sets we conclude that Λ is transitive, a contradiction. This proves the result. □

Lemma 3.3. *Let D be a fundamental domain of $W^{uu}_X(p_0)$ contained in $W^{u,+}$. There exist a neighborhood V of D and a cross-section Σ^- of X intersecting $W^{s,-}$ satisfying the following properties:*

1. *Every X-orbit's sequence in Λ converging to a point in $W^{s,-}$ intersects Σ^-.*
2. *No positive X-orbit with initial point in V intersects Σ^-.*

Proof. Fix a fundamental domain F of $W^s_X(\sigma)$ and define $F^- = F \cap W^{s,-}_X(\sigma)$ Then there is a compact interval $F' \subset F^-$ such that $\Lambda \cap F^- \subset F'$. By

Lemma 3.2 there is $\epsilon > 0$ such that $B_\epsilon(\mathrm{Cl}(W^{u,+})) \cap B_\epsilon(F') = \emptyset$. Clearly we can choose a cross-section Σ^- of X inside $B_\epsilon(F')$ such that every X-orbit's sequence in Λ converging to some point in $W^{s,-}$ intersects Σ^-. Since $W^{u,+}$ is invariant and $\mathrm{Cl}(W^{u,+}) \cap B_\epsilon(F') = \emptyset$, every positive orbit with initial point in D cannot intersect Σ^-. By using the contracting foliation of Λ we have the same property for every positive trajectory with initial point in a neighborhood V of D. This proves the result. □

Now we define a perturbation (pushing) close to a point $a \in W^u_X(\sigma) \setminus \{\sigma\}$. To this end we fix the following cross-sections:

1. Σ_a, containing a in its interior.
2. $\Sigma' = X_1(\Sigma)$.
3. Σ_0, intersecting O in a single interior point.
4. Σ^+, Σ^-, which intersect $W^{s,+}, W^{s,-}$, respectively, and point toward the side of a.

Every X-orbit intersecting $\Sigma^+ \cup \Sigma^-$ will intersect Σ. Note that there is a well-defined neighborhood $\mathcal{O}$ given by

$$\mathcal{O} = X_{[0,1]}(\Sigma_a).$$

This neighborhood will be the support of the pushing described in the Figures 2 and 3. The pushing in $\mathcal{O}$ yielding the perturbed flow Y of X is obtained in the standard way (see [**dMP**]).

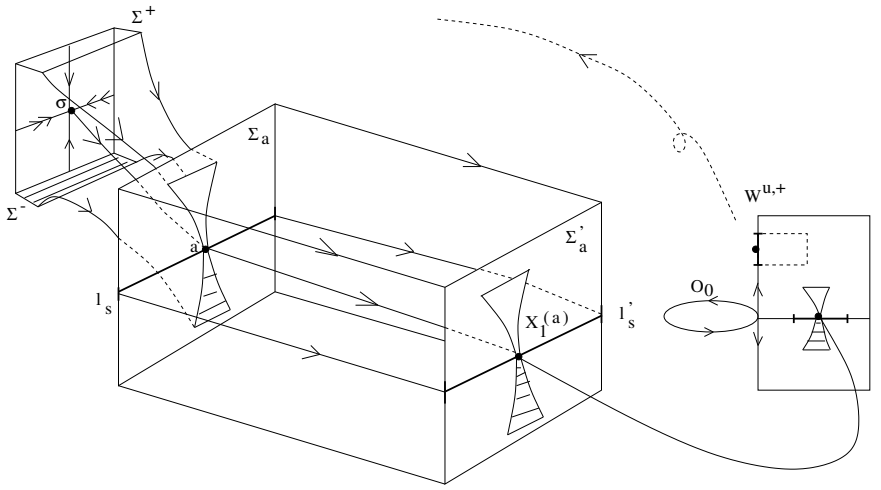


Figure 2. Unperturbed flow X.

We have to prove that $\Lambda(Y) \not\subset \Omega(Y)$ for the perturbed flow Y. For this purpose we observe that by 5 of Theorem 3.1 and Proposition 2.5 we can assume that there q periodic in U such that $W^u_X(q) \cap W^{s,-}_X(\sigma) \neq \emptyset$. We obtain in this way an interval K in $\Sigma^- \cap W^u_X(q)$ crossing Σ^- as in

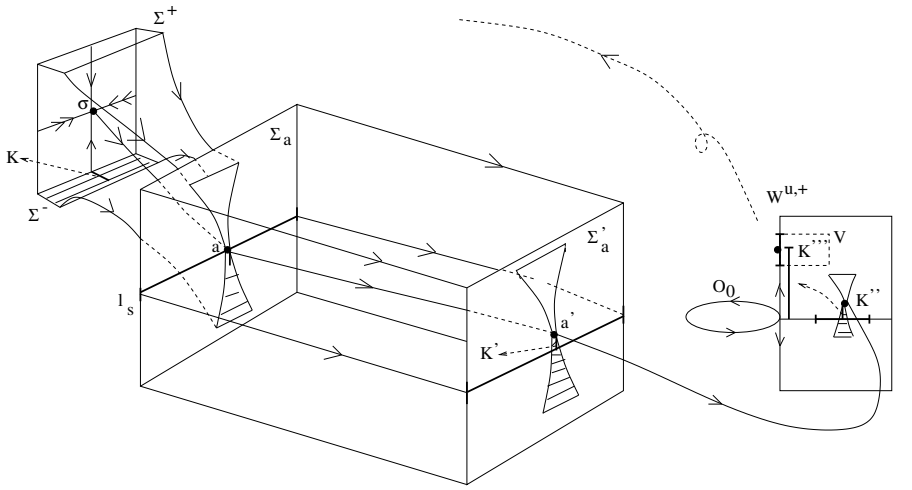


Figure 3. Perturbed flow Y.

Figure 3. The Y-flow carries K to an interval K'' as in Figure 3. Let $q(Y), W^u_Y(q(Y)), K''(Y), \sigma(Y)$ denote the continuation of these objects for the perturbed flow Y. We observe that $K(Y) \subset \Lambda(Y)$ since $\Lambda(Y)$ is an attracting set, $q(Y) \in \Lambda(Y)$ and $K \subset W^u_Y(q(Y))$. Then Theorem 3.1 will follow from the lemma below:

Lemma 3.4. *$K(Y) \not\subset \Omega(Y)$.*

Proof. Suppose for a contradiction that $K(Y) \subset \Omega(Y)$ and pick $x \in \operatorname{Int} K(Y)$, the interior of the interval $K(Y)$. The flow of Y carries the points near x to the neighborhood V depicted in Figure 3. This neighborhood is obtained by saturating a fundamental domain in $W^{u,+}$ by the strong stable manifolds [**HPS**]. Note that there are points x' near x that back up close to x under the forward flow of Y ($x \in K(Y) \subset \Omega(Y)$). In particular, *the positive Y-orbit of x' returns to Σ^-*. At the same time, Lemma 3.3-2 implies that no X-orbit starting in V intersects Σ^-. Since $X = Y$ outside $\mathcal{O}$ we conclude that the positive Y-orbit of x' intersects Σ^+. Afterward this positive orbit passes through the box $\mathcal{O}$ and arrives to V. By repeating the argument we conclude that *the positive Y-orbit of x' never returns to Σ^-*, a contradiction. The lemma is proved. □

Proof of Theorem A. Let Λ be a singular-hyperbolic attractor of a C^r flow X on a compact 3-manifold M. Assume that Λ is C^r critically robust and has a unique singularity σ. Denote by $\Lambda(Y) = \bigcap_{t>0} Y_t(U)$ the continuation of Λ in a neighborhood U of Λ for Y close to X. Denote by $C(Y)$ the union of the closed orbits of a flow Y. Since Λ is C^r critically robust, there is a neighborhood U of Λ such that $\Lambda(Y) \cap C(Y)$ is dense in $\Lambda(Y)$ for every flow

Y that is C^r close to X. Clearly $\Lambda(Y)$ is a singular-hyperbolic set of Y for all Y close to X. Because Λ has a unique singularity, so does $\Lambda(Y)$. Now recall that Λ is an attractor by assumption. In particular, Λ is transitive (recall Definition 1.1). It follows that Λ is connected and so the neighborhood U above can be arranged to be connected. Then $\Lambda(Y)$ is connected as well. Summarizing, $\Lambda(Y)$ is a singular-hyperbolic set of Y satisfying conditions 1–4 of Theorem 3.1. If Λ is not C^r robust, we can find a Y that is is C^r close to X and such that $\Lambda(Y)$ is not transitive. Then $\Lambda(Y)$ satisfies all conditions of Theorem 3.1, and we can find a Y' that is C^r close to Y and such that $\Lambda(Y') \not\subset \Omega(Y')$. This is a contradiction, since $\Lambda(Y') \subset \Omega(Y')$ (recall that $\Lambda(Y') \cap C(Y')$ is dense in $\Lambda(Y')$). This contradiction proves the result. □

References

[B] C. Bonatti, *C^1-generic dynamics: tame and wild behavior*, Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002), 265–277, Higher Ed. Press, Beijing, 2002, MR 1957538 (2003k:37002).

[Bw] R. Bowen, *Symbolic dynamics for hyperbolic flows*, Amer. J. Math., **95** (1973), 429–460, MR 0339281 (49 #4041), Zbl 0282.58009.

[dMP] W. de Melo and J. Palis, *Geometric Theory of Dynamical Systems. An Introduction*, Translated from the Portuguese by A. K. Manning, Springer-Verlag, New York-Berlin, 1982, MR 0669541 (84a:58004), Zbl 0491.58001.

[D] B. Deng, *The Silnikov problem, exponential expansion, strong λ-lemma, C^1-linearization, and homoclinic bifurcation*, J. Differential Equations, **79** (1989), 189–231 , MR 1000687 (90k:58161), Zbl 0674.34040.

[GW] J. Guckenheimer and R.F. Williams, *Structural stability of Lorenz attractors*, Inst. Hautes Études Sci. Publ. Math., **50** (1979), 59–72, MR 0556582 (82b:58055a), Zbl 0436.58018.

[HPS] M. Hirsch, C. Pugh and M. Shub, *Invariant Manifolds*, Lectures Notes in Mathematics **583**, Springer, Berlin, 1977, MR 0501173 (58 #18595), Zbl 0355.58009.

[Mi] J. Milnor, *On the concept of attractor*, Comm. Math. Phys., **99** (1985), 177–195, MR 0790735 (87i:58109a), Zbl 0595.58028; *Correction and remarks: "On the concept of attractor"*, Comm. Math. Phys. **102** (1985), 517–519, MR 0818833 (87i:58109b), Zbl 0602.58030.

[M] C.A. Morales, *Singular-hyperbolic sets and topological dimension*, Dyn. Syst., **18(2)** (2003), 181–189, MR 1994788 (2004g:37037).

[MP1] C.A. Morales and M.J. Pacifico, *Mixing attractors for 3-flows*, Nonlinearity, **14** (2001), 359–378, MR 1819802 (2002a:37036), Zbl 1026.37015.

[MP2] C.A. Morales and M.J. Pacifico, *Transitivity and homoclinic classes for singular-hyperbolic systems*, preprint, 2003, to appear.

[MPP] C.A. Morales, M.J. Pacifico and E.R. Pujals, *On C^1 robust transitive sets for three-dimensional flows*, C.R. Acad. Sci. Paris, Série I (Math.), **326** (1998), 81–86, MR 1649489 (99j:58183), Zbl 0918.58036.

[MPu] C.A. Morales and E.R. Pujals, *Singular strange attractors on the boundary of Morse-Smale systems*, Ann. Sci. Ecole Norm. Sup., **30** (1997), 693–717, MR 1476293 (98k:58137), Zbl 0911.58022.

[PT] J. Palis and F. Takens, *Hyperbolicity and Sensitive Chaotic Dynamics at Homoclinic Bifurcations. Fractal Dimensions and Infinitely Many Attractors*, Cambridge Studies in Advanced Mathematics, **35**, Cambridge University Press, Cambridge, 1993, MR 1237641 (94h:58129), Zbl 0790.58014.

[P] E.R. Pujals, *Tangent bundle dynamics and its consequences*, Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002), 327–338, Higher Ed. Press, Beijing, 2002, MR 1957543 (2004c:37036).

Received April 24, 2003. The work was partially supported by CNPq, FAPERJ and PRONEX/DYN. SYS.

Instituto de Matemática
Universidade Federal do Rio de Janeiro
C.P. 68.530
21945-970 Rio de Janeiro, RJ
Brazil
E-mail address: morales@impa.br

Instituto de Matemática
Universidade Federal do Rio de Janeiro
C.P. 68.530
21945-970 Rio de Janeiro, RJ
Brazil
E-mail address: pacifico@impa.br

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

ALGEBRAIC D-GROUPS AND DIFFERENTIAL GALOIS THEORY

Anand Pillay

We discuss various relationships between the algebraic D-groups of Buium, 1992, and differential Galois theory. In the first part we give another exposition of our general differential Galois theory, which is somewhat more explicit than Pillay, 1998, and where generalized logarithmic derivatives on algebraic groups play a central role. In the second part we prove some results with a "constrained Galois cohomological flavor". For example, if G and H are connected algebraic D-groups over an algebraically closed differential field F, and G and H are isomorphic over some differential field extension of F, then they are isomorphic over some Picard–Vessiot extension of F. Suitable generalizations to isomorphisms of algebraic D-varieties are also given.

1. Introduction

We work throughout with (differential) fields of characteristic zero. In [**8**] the notion of a generalized differential Galois extension (or generalized strongly normal extension) of a differential field was introduced, generalizing Kolchin's theory of strongly normal extensions, which in turn generalized the Picard–Vessiot theory. The idea was to systematically replace algebraic groups over the constants by "finite-dimensional differential algebraic groups", to obtain new classes of extensions of differential fields with a good Galois theory. This idea (almost obvious from the model-theoretic point of view) was implicit in Poizat [**11**] who gave a model-theoretic treatment of the strongly normal theory. However the "correct" definition of a generalized differential Galois extension needed some additional fine-tuning. Nevertheless, our exposition of this general theory in [**8**] was overly model-theoretic, and possibly remained somewhat obscure to differential algebraists. We try to remedy this in the current paper by concentrating on the *differential equations* that have a good Galois theory, very much in the spirit of Section 7, Chapter IV of Kolchin's book [**4**]. The key notion is that of a generalized logarithmic derivative on an algebraic group G over a differential field K (a certain kind of differential rational map from G to its Lie algebra). We

will see that such a generalized logarithmic derivative is essentially equivalent to an algebraic D-group structure on G (in the sense of Buium [**3**]). Our resulting exposition of the generalized differential Galois theory will be equivalent to that in [**8**] when the base field K is *algebraically closed*. The general situation (K not necessarily algebraically closed) can be treated using analogues of the V-primitives from [**4**, IV.10], and we leave the details to others.

Let me now say a little more about the generalized logarithmic derivatives, and how they tie up with the Picard–Vessiot/strongly normal theory. Let us fix a differential field K, and assume for now that the field C_K of constants of K is algebraically closed. A linear differential equation over K, in vector form, is $\partial y = Ay$, where y is a $n \times 1$ column vector of unknowns and A is an $n \times n$ matrix over K. Looking for a fundamental matrix of solutions, one is led to the equation on GL_n: $\partial Y = AY$, where Y is a $n \times n$ matrix of unknowns ranging over GL_n, which we can write as $\partial(Y)Y^{-1} = A$. Now the map $Y \to \partial(Y)Y^{-1}$ is the classical logarithmic derivative, a first-order differential crossed homomorphism from GL_n into its Lie algebra, which is surjective when viewed in a differentially closed overfield of K. A Picard–Vessiot extension of K for the original equation is then a differential field extension $L = K(g)$, where $g \in \mathrm{GL}_n$ is a solution of $\partial(Y)Y^{-1} = A$, and $C_L = C_K$. Such an extension exists, and is unique up to K-isomorphism. The group of (differential) automorphisms of L over K has the structure of an algebraic subgroup of $\mathrm{GL}_n(C_K)$, and there is a Galois correspondence.

In place of GL_n one can consider an arbitrary algebraic group G defined over K (not necessarily linear and not necessarily defined over the constants of K). By a *generalized logarithmic derivative* on G we will mean a first-order differential rational crossed homomorphism μ from G to $L(G)$, defined over K, such that μ is *geometrically surjective* (that is, surjective when viewed in a differentially closed overfield) and such that $\mathrm{Ker}\,(\mu)$, a *finite-dimensional differential algebraic subgroup* of G, is Zariski-dense in G. The analogue of a linear differential equation over K will then be an equation

$$(*) \qquad \mu(x) = a,$$

where x ranges over G and $a \in L(G)(K)$.

Under an additional technical condition on the data, analogous to the requirement that the field of constants of K be algebraically closed, we can define the notion of a differential Galois extension L of K for the equation $(*)$, prove its existence and uniqueness, identify the Galois group, and obtain a Galois correspondence. In the case where G is defined over the constant field C_K *and* μ is the standard logarithmic derivative of Kolchin, we recover Kolchin's strongly normal extensions (see Theorem 6, Section 7, Chapter IV of [**4**]).

For G an algebraic group over the differential field K, an algebraic D-group structure on G is precisely an extension of the derivation ∂ on K to a derivation on the structure sheaf of G, respecting the group operation. Algebraic D-groups belong entirely to algebraic geometry, and Buium [**3**] points out that there is an equivalence of categories between the category of algebraic D-groups and the category of ∂_0-groups, finite-dimensional *differential algebraic* groups. The latter category belongs to Kolchin's *differential algebraic geometry*. On the other hand, there is essentially a one-to-one correspondence between algebraic D-group structures on G and generalized logarithmic derivatives on G. So our general differential Galois theory is in a sense subsumed by the very concept of an algebraic D-group.

Details of the above will be given in Sections 2 and 3, including a "Tannakian" approach and an examination of different manifestations of the differential Galois group.

In Section 4 we will give another relation between algebraic D-groups and the Picard–Vessiot theory: if two algebraic D-groups over an algebraically closed differential field are isomorphic (as D-groups) over some differential field extension of K, then such an isomorphism can be found defined over a Picard–Vessiot extension of K. This uses Kolchin's differential Lie algebra, and strengthens the somewhat artificial results from [**9**]. We will also give some related results on isomorphisms between algebraic D-varieties, using differential jets (higher-dimensional versions of differential tangent spaces).

2. Algebraic D-groups

We will briefly describe the algebraic D-groups of Buium in a matter suitable for our purposes. We also introduce the generalized logarithmic derivative on an algebraic group induced by a given D-group structure. We refer to [**6**] for a discussion of related themes.

Let us fix an ordinary differential field (K, ∂). For convenience we also give ourselves a differential field extension $(\mathcal{U}, \partial)$ of (K, ∂) that is "universal" with respect to (K, ∂). Namely, $\mathcal{U}$ has cardinality $\kappa > |K|$, and for any differential subfield $F < \mathcal{U}$ of cardinality $< \kappa$ and differential extension L of K of cardinality $\leq \kappa$ there is an embedding (as differential fields) of L into $\mathcal{U}$ over K. C_K denotes the field of constants of K, and $\mathcal{C}$ the field of constants of $\mathcal{U}$.

We begin by recalling the tangent bundle of an algebraic variety or group over K (in which the derivation on K plays no role).

Let X be an algebraic variety over K (maybe reducible). The tangent bundle $T(X)$ of X is another algebraic variety over K, with a canonical surjective morphism π (over K) to X, and is defined locally by equations: $\sum_i \partial P/\partial x_i(x_1, \ldots, x_n)v_i = 0$ for P polynomials over K generating the ideal of X over K. If $X = G$ is an algebraic group over K, then $T(G)$ has the

structure of an algebraic group over K such that the canonical projection to G is an (algebraic) group homomorphism. The group operation on $T(G)$ is obtained by differentiating the group operation of G. That is, if $f(-,-)$ is the group operation on G and (g,u) and (h,v) are in $T(G)$ then the product $(g,u)\cdot(h,v)$ equals $(g\cdot h, df_{g,h}(u,v))$. Note that if λ^g, ρ^g denote left and right multiplication by g in G, then we have:

$$(*) \qquad \text{In } T(G), \quad (g,u)\cdot(h,v) = (g\cdot h, d(\lambda^g)_h(v) + d(\rho^h)_g(u)).$$

We will denote $T(G)_e$, the tangent space of G at the identity, by $L(G)$ (for the Lie algebra of G). Then $L(G)$, with its usual vector space group structure, is a normal subgroup of $T(G)$, and we have the exact sequence $0 \to L(G) \to T(G) \to G \to \{e\}$ of algebraic groups (over K). We denote by $i : L(G) \to T(G)$ the natural inclusion, and by $\pi : T(G) \to G$ the canonical surjection above.

Note that G acts on $L(G)$ (denoted $(g,a) \to a^g$) by the adjoint map (differentiating conjugation by $g \in G$ at the identity). And this "coincides" with the action of $T(G)$ on the normal subgroup $L(G)$ by conjugation: for $a \in L(G)$ and $g \in G$, $a^g = xax^{-1}$ for any $x \in T(G)$ such that $\pi(x) = g$.

For $a \in L(G)$, we let l_a and r_a denote the left and right invariant vector fields on G determined by a. Namely for $g \in G$, $l_a(g) = d(\lambda^g)_e(a)$ and $r_a(g) = d(\rho^g)_e(a)$.

A K-rational splitting of $T(G)$ as a semidirect product of G and $L(G)$ is given by either of the equivalent pieces of data:

(a) a K-rational homomorphic section $s : G \to T(G)$ (that is $\pi \circ s = \mathrm{id}$);
(b) a K-rational crossed homomorphism h from $T(G)$ onto $L(G)$ such that $h \circ i = \mathrm{id}$. (By definition $\alpha : T(G) \to L(G)$ is a crossed homomorphism if $\alpha(xy) = \alpha(x) + x\alpha(y)x^{-1}$.)

Note that the set of these K-rational splittings has the structure of a commutative group. For example, using the data in (a), if s_1, s_2 are K-rational homomorphic sections of the tangent bundle, and $s_i(g) = (g, u_i)$ for $i = 1, 2$ then $(s_1 + s_2)(g) = (g, u_1, +u_2)$. The identity element is just the 0-section $s_0(g) = (g, 0)$, which is a K-rational homomorphic section. Let $P_K(G)$ denote this commutative group of K-rational splittings of $T(G)$. We denote the crossed homomorphism from $T(G)$ onto $L(G)$ corresponding to the identity of $P_K(G)$ by $h_0 : T(G) \to L(G)$, and the crossed homomorphism corresponding to the homomorphic section s by h_s.

Remark 2.1.

(i) $h_0(g,u) = d(\rho^{g^{-1}})_g(u)$.
(ii) *More generally, if s is a K-rational homomorphic section of $T(G) \to G$, then* $h_s(g,u) = d(\rho^{g^{-1}})_g(u - s(g))$.

Proof. This follows directly from formula $(*)$ for multiplication in $T(G)$. □

Let us now bring in the differential structure. Assume for now that the algebraic variety X is defined over C_K the field of constants of K. Then it is easy to see that if $a \in X(\mathcal{U})$ then working in local coordinates with respect to a given covering of X by affine varieties over C_K, $(a, \partial(a)) \in T(X)$. If in addition $X = G$ is an algebraic group defined over C_K, then the map $\nabla : G(\mathcal{U}) \to T(G)(\mathcal{U})$ taking g to $(g, \partial(g))$ is a *group embedding*. Let $lD : G(\mathcal{U}) \to L(G)(\mathcal{U})$ be defined by $lD(g) = h_0(g, \partial(g))$. Then lD is a "differential rational" crossed homomorphism defined over K, which is precisely Kolchin's logarithmic derivative. The map lD depends on h_0, and clearly any other $h \in P_K(G)$ gives rise to another differential rational crossed homomorphism (over K) from $G(\mathcal{U})$ onto $L(G)(\mathcal{U})$.

Remark 2.2. *Suppose G is defined over C_K. Then:*

(i) *The (standard) logarithmic derivative $lD : G(\mathcal{U}) \to L(G)(\mathcal{U})$ is given by $lD(g) = d(\rho^{g^{-1}})_g(\partial(g))$.*
(ii) *lD is surjective.*
(iii) *$Ker(lD)$ is precisely $G(\mathcal{C})$.*

Proof. (i) follows from Remark 2.1.

(ii) Let $a \in L(G)(\mathcal{U})$. Then a determines the right invariant vector field $r_a : G(\mathcal{U}) \to T(G)(\mathcal{U})$. As $\mathcal{U}$ is differentially closed, the main result of [**7**] gives $g \in G(\mathcal{U})$ such that $\partial(g) = r_a(g)$, hence (by (i)), $lD(g) = a$.

(iii) is obvious from (i). □

Let us now work in a more general context, dropping our assumption that the variety X is defined over C_K. Then for $a \in X(\mathcal{U})$, $(a, \partial(a))$ may no longer be a point of $T(X)$ but rather a point of another bundle $\tau(X)$ over X, which we now describe. For $P(x_1, \ldots, x_n)$ a polynomial over K, let P^∂ denote the polynomial obtained from P by applying ∂ to its coefficients. (So $P^\partial = 0$ if P is over C_K.) Then $\tau(X)$ is defined locally by equations:

$$\sum_i \partial P/\partial x_i(x_1, \ldots, x_n)v_i + P^\partial(x_1, \ldots, x_n) = 0,$$

where P again ranges over polynomials in the ideal of X over K. (That is, these affine pieces fit together to give an algebraic variety $\tau(X)$ over K, together with a canonical projection from $\tau(X)$ to X.) It is immediate that $(a, \partial a) \in \tau(X)$ for $a \in X(\mathcal{U})$. It is also immediate that for each $a \in X$, $\tau(X)_a$ is a principal homogeneous space for the tangent space $T(X)_a$, where the action is addition (with respect to local coordinates above). Moreover this happens uniformly, making $\tau(X)$ a *torsor* for the tangent bundle $T(X)$. In any case, if X is defined over C_K, then $\tau(X)$ coincides with $T(X)$. By an algebraic D-variety over K we mean a pair (X, s) such that X is an algebraic variety over K and $s : X \to \tau(X)$ is a regular section defined over K.

Now assume again that $X = G$ is an algebraic group over K. Then $\tau(G)$ has the structure of an algebraic group over K such that the canonical projection $\tau(G) \to G$ is a homomorphism. In local coordinates, assuming that multiplication in G is given by the sequence of polynomials $f = (f_i(x_1, \ldots, x_n, y_1, \ldots, y_n))_{i=1,\ldots,n}$ over K, then for $(g, u), (h, v) \in \tau(G)$, the product of (g, u) and (h, v) in $\tau(G)$ is given by

$$\big(g \cdot h,\ df_{(g,h)}(u, v) + (f_1^\partial(g, h),\ \ldots,\ f_n^\partial(g, h))\big).$$

We again have the map $\nabla : G(\mathcal{U}) \to \tau(G)(\mathcal{U})$, given (in local coordinates) by $\nabla(g) = (g, \partial(g))$ and this is a *group embedding*.

So we have two K-algebraic groups $T(G)$ and $\tau(G)$. Although these need not be isomorphic as algebraic groups, they are "differential rationally" isomorphic. The following is left to the reader.

Lemma 2.3. *The map that takes (g, u) to $(g, \partial(g) - u)$ is a (differential rational) isomorphism of groups between $\tau(G)(\mathcal{U})$ and $T(G)(\mathcal{U})$. Although not necessarily rational, it is rational when restricted to the fibers over G. In particular, the above map defines a K-rational isomorphism between the vector groups $\tau(G)_e$ and $T(G)_e = L(G)$.*

Note that we have again an exact sequence

$$0 \to \tau(G)_e \to \tau(G) \to G \to e$$

of algebraic groups over K, which by virtue of the (canonical) isomorphism between $\tau(G)_e$ and $L(G)$ given by Lemma 2.3 can be rewritten as:

$$0 \to L(G) \to \tau(G) \to G \to e.$$

Let us again write i for the (canonical) injection of $L(G)$ in $\tau(G)$, and π for the canonical surjection $\tau(G) \to G$. So if G is defined over C_K this agrees with our earlier notation: $i : L(G) \to T(G)$ and $\pi : T(G) \to G$.

We can consider splittings (as algebraic groups over K) of $\tau(G)$ as a semidirect product of G and $L(G)$. Again each such splitting is determined either by a K-rational homomorphic section $s : G \to \tau(G)$, or a K-rational crossed homomorphism $h : \tau(G) \to L(G)$ such that $h \circ i = \mathrm{id}$ on $L(G)$. We will write h_s for the crossed homomorphism corresponding to the homomorphic section s, and give explicit formulas below.

In any case, we can now define an algebraic D-group.

Definition 2.4. Let G an algebraic group over K. Then an algebraic D-group structure on G over K is precisely a K-rational homomorphic section $s : G \to \tau(G)$. We write the corresponding algebraic D-group as (G, s).

Given an algebraic D-group (G, s) we obtain a generalized logarithmic derivative that we call lD_s, a crossed homomorphism (in the obvious sense) from $G(\mathcal{U})$ to $L(G)(\mathcal{U})$: $lD_s = h_s \circ \nabla$. Here is the analogue to Remarks 2.1 and 2.2.

Remark 2.5. *Let (G, s) be an algebraic D-group over K.*

(i) *For $(g, u) \in \tau(G)$, $h_s(g, u) = d(\rho^{g^{-1}})_g(u - s(g))$.*
(ii) *For $g \in G(\mathcal{U})$, $lD_s(g) = d(\rho^{g^{-1}})_g(\partial(g) - s(g)) \in L(G)(\mathcal{U})$.*
(iii) *$lD_s : G(\mathcal{U}) \to L(G)(\mathcal{U})$ is surjective.*
(iv) $\mathrm{Ker}\,(lD_s)$ *is precisely $\{g \in G(\mathcal{U}) : \partial(g) = s(g)\}$, and is a Zariski-dense subgroup of $G(\mathcal{U})$.*

Proof. (i) follows from the formula for multiplication in $\tau(G)$, and (ii) is an immediate consequence of (i).

(iii) Let $a \in L(G)(\mathcal{U})$. Again we obtain the right invariant vector field $r_a : G(\mathcal{U}) \to T(G)(\mathcal{U})$. Then $r_a + s : G(\mathcal{U}) \to \tau(G)(\mathcal{U})$ is also a rational section of $\tau(G) \to G$. By [**7**] there is $g \in G(\mathcal{U})$ such that $\partial(g) = r_a(g) + s(g)$, hence $lD_s(g) = a$ by (i).

(iv) $\mathrm{Ker}\,(lD_s)$ is a subgroup of $G(\mathcal{U})$ as lD_s is a crossed homomorphism. As $d(\rho^{g^{-1}})_g$ is an isomorphism between $T(G)_g$ and $T(G)_e$, we see that $\mathrm{Ker}\,(lD_s)$ is as described in (iii). By [**7**] for any proper subvariety X of $G(\mathcal{U})$ there is $g \in G(\mathcal{U})$ such that $\partial(g) = s(g)$. Hence by (i) $\mathrm{Ker}\,(lD_s)$ is Zariski dense in G. □

In the context of Remark 2.5, we denote $\mathrm{Ker}\,(lD_s)$ by $(G, s)^\sharp$. This is a finite-dimensional differential algebraic group, or ∂_0-group in the sense of [**3**], and is an object belonging to Kolchin's differential algebraic geometry. Just for the record, here are some key properties: $(G, s)^\sharp(\mathcal{U})$ is Zariski-dense in G, the ∂_0-subvarieties of $(G, s)^\sharp$ are precisely of the form $X \cap (G, s)^\sharp$ for X a D-subvariety of (G, s), and for any other algebraic D-group (H, t) the ∂_0-homomorphisms between $(G, s)^\sharp$ and $(H, t)^\sharp$ are precisely those induced by algebraic D-group homomorphisms between (G, s) and (H, t).

In any case, we have seen that an algebraic D-group structure (G, s) on an algebraic group G over K determines a generalized logarithmic derivative lD_s on G. We point out now, just for completeness, that conversely any suitable differential rational crossed homomorphism map (over K) from $G(\mathcal{U})$ to $L(G)(\mathcal{U})$ determines an algebraic D-group structure on G.

Some notation: Let X and Y be algebraic varieties defined over K. By a *first-order differential rational map* $h : X(\mathcal{U}) \to Y(\mathcal{U})$, defined over K, we mean a map h from $X(\mathcal{U})$ to $Y(\mathcal{U})$ such that for, for each $x \in X(\mathcal{U})$, $h(x) \in K(x, \partial(x))$.

Lemma 2.6. *Suppose that G is a connected algebraic group over K. Let $l : G(\mathcal{U}) \to L(G)(\mathcal{U})$ be a first-order differential rational crossed homomorphism, defined over K, which is surjective, and is such that* $\mathrm{Ker}\,(l)$ *is Zariski-dense in $G(\mathcal{U})$. Then, there are a unique K-rational automorphism σ of $L(G)$, and a unique algebraic D-group structure (G, s) on G (defined over K), such that $l = \sigma \circ lD_s$.*

Proof. By [**7**], for example, $\nabla(G(\mathcal{U}))$ is Zariski-dense in $\tau(G)(\mathcal{U})$. Define l_1 on $\nabla(G(\mathcal{U}))$ by $l_1(x, \partial(x)) = l(x)$. So by Zariski-denseness, and the properties of l, l_1 extends uniquely to a K-rational surjective crossed homomorphism f from $\tau(G)$ to $L(G)$.

By the Zariski-denseness of $\mathrm{Ker}\,(l)$ in G and the definition of f, we have $\pi(\mathrm{Ker}\,(f)) = G$. On the other hand, $\dim(\tau(G)) = 2\dim(G)$ and $\dim(G) = \dim(L(G))$. Hence $\dim(\mathrm{Ker}\,(f)) = \dim(G)$. Now $\pi|\mathrm{Ker}\,(f) : \mathrm{Ker}\,(f) \to G$ is a group homomorphism so it has finite kernel. But this finite kernel is a subgroup of the vector group $\tau(G)_e$ so has to be trivial. It follows that $f|\tau(G)_e$ is an isomorphism (over K) with $L(G)$. So there is a unique K-automorphism σ of $L(G)$ such that $(\sigma \circ f) \circ i$ is the identity on $L(G)$. Put $f_1 = \sigma \circ f$. Then f_1 gives an algebraic D-group structure (G, s) on G defined over K. For $g \in G(\mathcal{U})$, $\sigma \circ l(g) = f_1(g, \partial(g)) = lD_s(g)$. □

There is a natural notion of a D-morphism between algebraic D-varieties (X, s) and (Y, t). First note that $\tau(-)$ is a functor, so if $f : X \to Y$ is a morphism between the algebraic varieties X and Y with everything defined over K then $\tau(f)$ is a morphism, over K, between $\tau(X)$ and $\tau(Y)$, again defined over f. (If X, Y, f are defined over the constants, then $\tau(f)$ is just the differential of f.) In any case, a morphism between algebraic D-varieties X and Y is by definition a morphism f of algebraic varieties, such that $t \circ f = \tau(f) \circ s$.

In particular we extract the notion of an algebraic D-subvariety of (X, s): so an algebraic subvariety Y of X will be the underlying variety of an algebraic D-subvariety of (X, s) if $s|Y$ maps Y to $\tau(Y)$.

A homomorphism of algebraic D-groups is a D-morphism that is also a homomorphism of algebraic groups.

We call an algebraic D-group (G, s) isotrivial if it is isomorphic over $\mathcal{U}$ to a trivial algebraic D-group (H, s_0), where H is defined over $\mathcal{C}$ and s_0 is the 0-section of $T(G)$.

The interest of the category of algebraic D-groups is that there exist nonisotrivial algebraic D-groups. If A is an abelian variety over $\mathcal{U}$ and $p : G \to A$ is the universal extension of A by a vector group then G has a D-group structure (defined over the field over which A is defined). Moreover any such D-group structure on G is nonisotrivial if A is not isomorphic (as an algebraic group) to an abelian variety defined over $\mathcal{C}$. Given such a D-group structure (G, s) on G, $p((G, s)^\sharp) < A$ is precisely the *Manin kernel* of A. This example is worked out in detail in [**6**].

Let us repeat from [**9**] the discussion of a nonisotrivial algebraic D-group structure on the commutative algebraic group $G_m \times G_a$. So let $G = G_m \times G_a$. Then $T(G) = \tau(G)$ can be identified with $\{(x, y, u, v) : x \neq 0\}$ with group structure $(x_1, y_1, u_1, v_1) \cdot (x_2, y_2, u_2, v_2) = (x_1x_2, y_1 + y_2, u_1x_2 + u_2x_1, v_1 + v_2)$.

Let $s : G \to T(G)$ be the homomorphic section $s(x, y) = (x, y, xy, 0)$. Then (G, s) is an algebraic D-group known to be nonisotrivial.

Note that if $g = (x, y) \in G$, then the differential of multiplication by g^{-1} at g takes $(u, v) \in T(G)_g$ to $(u/x, v) \in L(G)$. Hence by 2.5, the generalized logarithmic derivative lD_s corresponding to s is: $lD_s(x, y) = ((\partial(x) - xy)/x, \partial(y)) = (\partial(x)/x - y, \partial(y))$.

In particular $(G, s)^\sharp = \mathrm{Ker}\,(lD_s) = \{(x, y) \in G : \partial(x)/x = y, \partial(y) = 0\}$, which can be identified with the subgroup of G_m defined by the second-order equation $\partial(\partial(x)/x) = 0$.

3. Differential Galois theory

The conventions of the previous section are in force. In particular (K, ∂) is a differential field of characteristic 0, and we may refer also to the universal differential field extension $(\mathcal{U}, \partial)$ of K.

By a *logarithmic differential equation* over a differential field (K, ∂) we mean something of the form

$$(*) \qquad lD_s(x) = a,$$

where (G, s) is an algebraic D-group defined over K and $a \in L(G)(K)$. (So the indeterminate x ranges over G.) As remarked in the introduction, a special case is the equation $\partial(X) = AX$, where X ranges over GL_n and A is an $n \times n$ matrix over K.

In order to give the right notion of a differential Galois extension of K for $(*)$, we need to place a further restriction on K, G and s.

Definition 3.1. Suppose (G, s) is an algebraic D-group over K. We say that (G, s) is K-large, if for every (maybe reducible) algebraic D-subvariety X of G, which is defined over K, $X(K) \cap (G, s)^\sharp$ is Zariski-dense in X.

Remark 3.2.

(i) The intuitive meaning of (G, s) being K-large is that $(G, s)^\sharp$ has enough points with coordinates in K.

(ii) Suppose that G is defined over C_K and that $s = s_0$, the 0-section of $T(G)$. Then, if C_K is algebraically closed, (G, s) is K-large.

In the next remark, we refer to *differential closures* of K. A differential closure of K is a differential field extension of K that embeds over K into any differentially closed field containing K. The differential closure of K is unique up to K-isomorphism, and is written as $\hat{K}$. Kolchin calls $\hat{K}$ the constrained closure of K.

Remark 3.3. *(G, s) is K-large if and only $(G, s)^\sharp(K) = (G, s)^\sharp(\hat{K})$ for some (any) differential closure $\hat{K}$ of K.*

Proof. Assume first that $(G,s)^\sharp(K) = (G,s)^\sharp(\hat{K})$. Let X be a D-subvariety of (G,s) defined over K. The irreducible components $X_1, \ldots, X_r$ of X are defined over $\overline{K}$ so also $\hat{K}$ and are also D-subvarieties of (G,s). By [**7**] for example for any nonempty Zariski open subset of any X_i defined over $\hat{K}$, there is $a \in U(\hat{K}$ such that $\partial(a) = s(a)$. By our assumptions $a \in (G,s)^\sharp(K)$. So $X \cap (G,s)^\sharp(K)$ is Zariski-dense in X.

Conversely, suppose (G,s) is K-large. Let $a \in (G,s)^\sharp(\hat{K})$, and suppose for a contradiction that $a \notin G(K)$. Now a is *constrained* over K in the sense of Section 10, Chapter III of [**4**]. This means that there is some differential polynomial $P(x)$ over K such that $P(a) \neq 0$, and whenever b is a differential specialization of a over K and $P(b) \neq 0$ then b is a "generic" specialization of a over K (in model-theoretic language $tp(b/K) = tp(a/K)$). Let X be the irreducible K-subvariety of G whose generic point is a. Then (as $\partial(a) = s(a)$), X is a D-subvariety of (G,s). Moreover the differential specializations of a over K are precisely those b such that $b \in X$ and $\partial(b) = s(b)$. Now subject to the conditions "$x \in X$ and $\partial(x) = s(x)$", the condition $P(x) \neq 0$ is clearly equivalent to $x \notin Y$ for Y some proper subvariety of X defined over K. By our assumptions there is $b \in G(K)$ such that $b \in X \setminus Y$ and $\partial(b) = s(b)$. This is a contradiction. □

We call a differential ring (R,∂) *simple* if it has no proper nontrivial differential ideals. We refer to [**12**] for a discussion of simple differential rings.

Lemma 3.4. *Suppose that R is a simple differential ring over K that is finitely generated over K. Then R embeds over K into some differential closure of K.*

Proof. As R has no zero-divisors, R embeds in $\mathcal{U}$ over K. Let $R = K[a]_\partial$ be differentially generated over K by the finite tuple a. Suppose that $\pi(a) = b$ is a differential specialization over K. Then π extends to a surjective ring homomorphism $\pi : R \to K[b]_\partial$. The kernel is a differential ideal, so must be trivial. Thus b is a generic specialization of a over K. It follows that a is constrained over K so lives in some differential closure of K, as does R. □

We can now give the main definition.

Definition 3.5. Let (G,s) be a K-large algebraic D-group defined over K, and $lD_s(x) = a$ be a logarithmic differential equation over K for (G,s). By a differential Galois extension of K for the equation $lD_s = a$ we mean a differential field extension L of K of the form $K(\alpha)$ for some solution α of the equation such that $K[\alpha]$, the (differential) ring generated by K and the coordinates of α, is a simple differential ring.

Lemma 3.6 (Existence and uniqueness of differential Galois extensions)**.** *If (G,s) is a K-large algebraic D-group defined over K, and $a \in L(G)(K)$,*

then there exists a differential Galois extension L of K for the equation $lD_s(x) = a$. Moreover, any two such extensions are isomorphic over K as differential fields.

Proof. By Remark 2.5 (iii) there is a solution $\beta \in G(\mathcal{U})$ of $lD_s(x) = a$. Let α be a maximal differential specialization of β over K. Then $K[\alpha]$ is a simple differential ring and α is also a solution of $lD_s = a$, so we get existence. Let $L = K(\alpha)$.

Suppose L_1 is another differential Galois extension of K for the equation, generated by the solution γ say. By Lemma 3.4 we may assume that both L and L_1 are contained in some differential closure $\hat{K}$ of K. By Remark 3.3, $(G,s)^\sharp(\hat{K}) = (G,s)^\sharp(K)$. As both α and γ are solutions of $lD_s(x) = a$, it follows that $\alpha^{-1} \cdot \gamma \in (G,s)^\sharp(\hat{K}) = (G,s)^\sharp(K)$. Thus clearly $L = L_1$. □

Here are some alternative characterizations of differential Galois extensions:

Lemma 3.7. *Let (G,s) be a K-large algebraic D-group over K, and $L = K(\alpha)$ a differential field extension of K, where α is a solution of $lD_s(x) = a$ (with $a \in L(G)(K)$). Then the following are equivalent:*

(i) *L and α satisfy Definition* 3.5.
(ii) *L is contained in some differential closure of K.*
(iii) *$(G,s)^\sharp(K) = (G,s)^\sharp(L)$ and (G,s) is L-large.*

Proof. (i) implies (ii) is given by Lemmas 3.4 and 3.6 and (ii) implies (iii) follows from Remark 3.3.

(iii) implies (i). Assume L satisfies (iii). Then using Remark 3.3, we have that $(G,s)^\sharp(\hat{L}) = (G,s)^\sharp(K)$. Now $\hat{K}$ embeds in $\hat{L}$ over K, hence by Lemmas 3.4 and 3.6, there is a solution $\beta \in G(\hat{L})$ of $lD_s(x) = a$, such that $K[\beta]$ is a simple differential ring. Now $\alpha = \beta \cdot g$ for some $g \in (G,s)^\sharp(K)$, so clearly $K[\alpha]$ is also a simple differential ring. □

Condition (iii) is the analogue of "no new constants" in the strongly normal case.

Remark 3.8. *Suppose K is algebraically closed. Then the differential Galois extensions of K in the sense of Definition* 3.5 *coincide with the generalized strongly normal extensions of K in the sense of* [**8**]. *In particular, L is a strongly normal extension of K (in the sense of Kolchin) just if L is a differential Galois extension of K for an equation $lD_s(x) = a$, on an algebraic D-group (G,s), where G is defined over C_K and $s = s_0$.*

Proof. This follows from Proposition 3.4 of [**8**]. □

We now point out that given L as above, $\mathrm{Aut}_\partial(L/K)$ has the structure of a differential algebraic group (over K) in two different ways. One corresponds to the usual differential Galois group in the linear case, and is simply of the

form $(H,s)^\sharp(K)$, where H is an algebraic D-subgroup of (G,s). The other, corresponding to the "intrinsic" Galois group introduced by Katz, is of the form $(H_1,s_1)^\sharp(L)$ for s_1 another algebraic D-group structure on G (defined over K), and H_1 a D-subgroup of (G,s_1) defined over K. The algebraic D-groups (H,s) and (H_1,s_1) will be isomorphic, but not necessarily over K, unless they are commutative. (But of course if K is algebraically closed, H and H_1 will be isomorphic over K as *algebraic groups*.)

So fix $L = K(\alpha)$ as in Definition 3.5. By 3.7 we have $L < \hat{K}$ for some fixed copy of the differential closure of K. $\mathrm{Aut}_\partial(L/K)$ denotes the group of differential field automorphisms of L over K. Note that for any $\sigma \in \mathrm{Aut}_\partial(L/K)$, $\sigma(\alpha)$ is also a solution of $lD_s = a$, hence $\sigma(\alpha) = b \cdot c(\sigma)$ for a unique $c(\sigma) \in (G,s)^\sharp(L) = (G,s)^\sharp(K)$ $(= (G,s)^\sharp(\hat{K}))$.

Lemma 3.9. *The map c above is a group isomorphism between $\mathrm{Aut}_\partial(L/K)$ and $(H,s)^\sharp(K)$ for some algebraic D-subgroup H of G defined over K.*

Proof. As an automorphism σ of L over K is determined by its action on α, clearly the map is a group isomorphism with its image. So all that we need is that the image is of the required form. The model-theoretic proof of this goes through showing that the image is a *definable* subgroup of $(G,s)^\sharp(\hat{K})$, and then using quantifier-elimination for differentially closed fields. We will give an algebraic proof, after first discussing the second incarnation of the Galois group.

First, the equation $lD_s(x) = a$ equips G with another structure of an algebraic D-variety (but not in general an algebraic D-group). Let r_a be the right invariant vector field on G determined by a on G. So $s + r_a$ is a K-rational section of $\tau(G) \to G$, which we denote by s'. Note that the equation $lD_s(x) = a$ on G is equivalent to the equation $\partial(x) = s'(x)$.

Now G acts on itself by left translation. Let $S < G$ be the intersection of the stabilizers of the algebraic D-subvarieties of the algebraic D-variety (G,s') that are defined over K: namely (working in some algebraically closed field containing K), $S = \{g \in G : g \cdot X = X$ for all D-subvarieties X of (G,s') defined over $K\}$. S is an algebraic subgroup of G defined over K. S is precisely the analogue in our context of the intrinsic differential Galois group introduced by Katz in the Picard–Vessiot case and discussed by Bertrand in [**1**].

In fact G can be naturally equipped with another structure of an algebraic D-group, (G,s_1). Define the section $s_1 : G \to \tau(G)$ by

$$s_1(g) = s(g) + r_a(g) - l_a(g).$$

Lemma 3.10. *(G,s_1) is an algebraic D-group, and the action of G on itself by left multiplication is an action of the algebraic D-group (G,s_1) on the algebraic D-variety (G,s'). Moreover S is an algebraic D-subgroup of (G,s_1).*

Proof. An easy computation. Note that it follows that $(G, s_1)^\sharp(\mathcal{U})$ acts on $(G, s')^\sharp(\mathcal{U})$. □

Let Z be the set of solutions of $lD_s(x) = a$ in $G(L)$. Note that Z is precisely $(G, s')^\sharp(L)$. In any case, Z is a principal homogeneous space for $(G, s)^\sharp(K)$ (acting on the right). As $(G, s)^\sharp(K) = (G, s)^\sharp(\hat{K})$, and $L < \hat{K}$, X is also the solution set of $lD_s(x) = a$ in $G(\hat{K})$. Clearly $\mathrm{Aut}_\partial(L/K)$ acts on Z and any $\sigma \in \mathrm{Aut}_\partial(L/K)$ is determined by its action on Z. In fact, as $L = K(\beta)$ for any $\beta \in Z$, $\sigma \in \mathrm{Aut}_\partial(L/K)$ is determined by any pair $(\beta, \sigma(\beta))$ for $\beta \in Z$.

Lemma 3.11. *The action of* $\mathrm{Aut}_\partial(L/K)$ *on* Z *is isomorphic to the action (by left multiplication) of* $(S, s_1)^\sharp(L)$ *on* Z*: The isomorphism* d *say takes* σ *to* $\sigma(\beta) \cdot \beta^{-1}$ *for some (any)* $\beta \in Z$.

Proof. We know that $(S, s_1)^\sharp$ acts on Z by left multiplication (by Lemma 3.10). Suppose $\sigma \in \mathrm{Aut}_\partial(L/K)$. Let $d(\sigma) \in G(L)$ be such that $\sigma(\alpha) = d(\sigma) \cdot \alpha$. As $\sigma(\alpha) \in Z$, we have by Lemma 3.10 that $d(\sigma) \in (G, s_1)^\sharp(L)$. As any $\beta \in Z$ is of the form $\alpha \cdot c$ for $c \in (G, s)^\sharp(L) = (G, s)^\sharp(K)$ and σ fixes K pointwise, we see that

$$(*) \qquad \sigma(\beta) = d(\sigma) \cdot \beta \text{ for all } \beta \in Z.$$

We only have to see that $d(\sigma) \in S$. Let X be any D-subvariety of (G, s') defined over K. As $\hat{K}$ is differentially closed and $Z = (G, s')^\sharp(\hat{K})$, it follows that $Z \cap X$ is Zariski-dense in X. For $\beta \in Z \cap X$, $\sigma(\beta) \in Z \cap X$ too. By $(*)$ and Zariski-denseness, $d(\sigma) \cdot X = X$. Thus $d(\sigma) \in S$.

Conversely, suppose that $g \in (S, s_1)^\sharp(L)$. Then $g \cdot \alpha \in Z$. Let X be the algebraic variety defined over K whose K-generic point is α. Then, X is a D-subvariety of (G, s'). Thus $g \cdot X = X$ and so $g \cdot \alpha \in X$, and $\partial(g \cdot \alpha) = s'(g \cdot \alpha)$. Hence $g \cdot \alpha$ is a differential specialization of α over K. By simplicity of $K[\alpha]$, α and $g \cdot \alpha$ satisfy exactly the same differential polynomial equations over K. As both α and $g \cdot \alpha$ generate L it follows that there is an automorphism σ of L over K such that $\sigma(\alpha) = g \cdot \alpha$. So $g = d(\sigma)$. □

Conclusion of proof of Lemma 3.9. Let H be the image of S under conjugation by α^{-1} in G ($g \to \alpha^{-1} \cdot g \cdot \alpha$). Then H is an algebraic D-subgroup of (G, s). By 3.11 and what we already know about 3.9, the image of the embedding $c : \mathrm{Aut}_\partial(L/K) \to (G, s)^\sharp(K)$ is precisely $(H, s)^\sharp(K)$. As the latter is Zariski-dense in H, H is also defined over K. □

Let us fix the isomorphism c between $\mathrm{Aut}_\partial(L/K)$ and $(H, s)^\sharp(K) = (H, s)^\sharp(\hat{K})$.

Lemma 3.12. *There is a Galois correspondence between the set of differential fields in between* K *and* L *and the set of algebraic* D*-subgroups of*

(H, s) defined over K (equivalently defined over $\hat{K}$): given $K < F < L$ the corresponding group is H_F the Zariski closure of the set of $h \in (H, s)^\sharp(K)$ such that $c^{-1}(h)$ is the identity on F.

Proof. This is Theorem 2.12 of [**8**] after making the translation between definable subgroups and D-subgroups. □

Let us complete this section with an example. We will consider the nonisotrivial algebraic D-group structure (G, s) on $G_m \times G_a$ discussed at the end of Section 2, and exhibit a (natural) differential Galois extension $K < L$ whose differential Galois group is (G, s) (or rather $(G, s)^\sharp(K)$). In fact there will be an intermediary differential field $K < F < L$ such that each of $K < F$ and $F < K$ are Picard–Vessiot extensions, but $K < L$ is *not* a Picard–Vessiot extension.

Recall that $s : G \to T(G)$ is given by $s(x, y) = (x, y, xy, 0)$, and $lD_s : G \to L(G)$ is $lD_s(x, y) = (\partial(x)/x - y, \partial(y))$, and so a logarithmic differential equation on (G, s) has the form $\{\partial(x)/x - y = a_1, \partial(y) = a_2\}$. If we restrict our attention to equations where $a_1 = 0$, we obtain equations of the form $\partial(\partial(x)/x) = a$ on G_m.

Also $(G, s)^\sharp$ can be identified with $\{x \in G_m : \partial(\partial(x)/x)) = 0\}$.

Note that the embedding $x \to (x, 0)$ of G_m in G and surjection $(x, y) \to y$ of G onto G_a induces an exact sequence

$$0 \to (G_m, (s_0)_{G_m}) \to (G, s) \to (G_a, (s_0)_{G_a}) \to 0$$

of algebraic D-groups, where s_0 denotes the 0-sections for the corresponding groups. The important fact is that (G, s) is not a product (as a D-group) of the two groups, which is the reason that (G, s) is nonisotrivial.

We will take the ground field of constants to be $\mathbf{C}$. Let $K = \mathbf{C}(e^{ct} : c \in \mathbf{C})$ and $L = K(t, e^{t^2})$. So L is a subfield of a fixed differential closure $\hat{\mathbf{C}} = \hat{K} = \hat{L}$ of $\mathbf{C}$.

Lemma 3.13. *(G, s) is K-large.*

Proof. It is enough to show that all solutions of $\partial(\partial(x)/x) = 0$ in $\hat{K}$ are in K, that is all solutions of $\partial(d)/d = c$ for $c \in \mathbf{C}$ that are in $\hat{K}$ are in K. But this is clear, because e^{ct} is one such such solution and the others are obtained by multiplying by a constant. □

Lemma 3.14. *L is a differential Galois extension of K for the equation $\partial(\partial(x)) = 2$. The Galois group is $(G, s)^\sharp(K)$.*

Proof. Note that L is generated over K as a differential field by e^{t^2}, which is a solution of $\partial(\partial(x)) = 2$. To show that the Galois group is as stated, it is enough to show that $tr.deg(L/K) = 2$, which is well-known. □

Note that $K(t)$ is a Picard–Vessiot extension of K, and L is a Picard–Vessiot extension of $K(t)$, but L is not a Picard–Vessiot extension of K.

4. Isomorphisms of algebraic D-groups and algebraic D-varieties

We will prove:

Proposition 4.1. *Suppose (K, ∂) is an algebraically closed differential field, (G, s) and (H, t) are connected algebraic D-groups over K and there is an isomorphism f between (G, s) and (H, t) defined over some differential field extension of K. Then there is such an isomorphism defined over a Picard–Vessiot extension of K.*

Here is a restatement of the theorem in the language of Kolchin's constrained cohomology (see [**5**]).

Corollary 4.2. *Suppose (K, ∂) has no proper Picard–Vessiot extensions. Then for any connected ∂_0-group G defined over K, $H^1_c(\mathrm{Aut}_\partial(\hat{K}/K), G(\hat{K}))$ is trivial.*

In Proposition 3.11 of [**9**], Corollary 4.2 was stated in the special case that $H(K) = H(\hat{K})$ (which amounts to saying that the corresponding algebraic D-group is K-large).

Kolchin's differential tangent space and its properties (see Chapter 8 of [**5**]) play an important role in the proof of Proposition 4.1. We will summarize the key properties (in the language of algebraic D-groups). Recall first that if V is a finite-dimensional vector space over $\mathcal{U}$, then a ∂-module structure on V is an additive map $D_V : V \to V$ such that $D_V(av) = \partial(a)v + aD_V(v)$ for all $a \in \mathcal{U}$ and $v \in V$. V^∂ denotes $\{v \in V : D_V(v) = 0\}$, a vector space over $\mathcal{C}$ with $\mathcal{C}$-dimension the same as the $\mathcal{U}$-dimension of V

Fact 4.3. Suppose (G, s) is a connected algebraic D-group defined over K, and $V = L(G)$ is its Lie algebra (tangent space at the identity). Then:

(i) s equips V with a canonical ∂-module structure D_V, defined over K.
(ii) For any automorphism f of (G, s), $df_e \in \mathrm{GL}(V)$ restricts to a $\mathcal{C}$-linear automorphism of V^∂, and moreover f is determined by $df_e|V^\partial$.

Proof of Proposition 4.1. First (G, s) and (H, t) will be isomorphic over $\hat{K}$. Let f be such an isomorphism. Let c be a finite tuple from $\hat{K}$ generating the smallest field of definition of f. Write f as f_c. Let (V, D_V) be as in Fact 4.3 for (G, s). Let d be a $\mathcal{C}$-basis for V^∂ contained in $\hat{K}$. Let $L = K\langle c, d\rangle$ be the differential field generated by c and d over K.

Claim I. *L is a strongly normal extension of K.*

Proof. As $L < \hat{K}$, $C_L = C_K$. We have to show that for any automorphism σ of the differential field $\mathcal{U}$ fixing K pointwise, $\sigma(L)$ is contained in the differential field generated by L and $\mathcal{C}$. First $\sigma(d)$ is another basis of V^∂, so with respect to the basis d we may write $\sigma(d)$ as a $n \times n$ nonsingular matrix $B \in \mathrm{GL}_n(\mathcal{C})$. On the other hand $f_\sigma(c) = f_c \cdot h$ for a unique automorphism

h of (G, s). By Fact 4.3(ii), h is determined by $dh_e|V^\partial$, which by in terms of the basis d, is another nonsingular $n \times n$ matrix A over $\mathcal{C}$. It follows that $(\sigma(c), \sigma(d))$ is in the differential field generated by K, c, d, A and B. In particular $\sigma(L) \subseteq L\langle\mathcal{C}\rangle$. □

Claim II. *L is a Picard–Vessiot extension of K.*

Proof. Let $\sigma \in \mathrm{Aut}_\partial(L/K)$, and write the matrices A, B (which will be in $\mathrm{GL}_n(C_K)$) as A_σ, B_σ. Then it is easy to check that $\sigma \mapsto (A_\sigma, B_\sigma)$ is an embedding of $\mathrm{Aut}_\partial(L/K)$ into $\mathrm{GL}_n(C_K) \times \mathrm{GL}_n(C_K)$. From Kolchin's characterizations of Picard–Vessiot extensions, we get Claim II. □

As f is defined over L, we have proved Proposition 4.1. □

Finally let us give some additional results with a similar flavor. The first is really just a remark and is related to the themes of Buium's book [**2**]. Buium calls an algebraic D-variety (X, s) *split* if (X, s) is isomorphic (over $\mathcal{U}$) to a trivial algebraic D-variety, namely one of the form (Y, s_0), where Y is defined over $\mathcal{C}$ and s_0 is the 0-section. He proves that any D-variety (X, s) such that X is a projective variety is split. Moreover assuming Y defined over algebraically closed K, then (Y, s) is split over some strongly normal extension.

Remark 4.4. *Let K be a differential field with algebraically closed constant field. Suppose that (X, s) is an algebraic D-variety over K that is split. Then (X, s) is split over some strongly normal extension K_1 of K.*

Proof. We can find an isomorphism f of (X, s) with some trivial (Y, s_0) defined over $\hat{K}$. Let again $f = f_c$ with c the smallest field of definition of f. Let $K_1 = K\langle c\rangle$. Then, as K_1 is contained in $\hat{K}$ and $C_{\hat{K}} = C_K$, we have that $C_{K_1} = C_K$. For any (differential) automorphism σ of $\mathcal{U}$, Then $f_{\sigma(c)} = f_c \circ g$ for some automorphism g of (Y, s_0). But then g must be defined over $\mathcal{C}$. Hence $\sigma(c)$ is rational over $K(c)(\mathcal{C})$, so $\sigma(K_1) \subseteq K_1\langle\mathcal{C}\rangle$. This shows that K_1 is a strongly normal extension of K. □

The next result is a generalization of Proposition 4.1 in which the higher-dimensional versions of differential tangent spaces from [**10**] enter the picture. We use freely the results from that paper.

Proposition 4.5. *Suppose that K has algebraically closed constant field. Let (X, s) and (Y, t) be algebraic D-varieties defined over K. Suppose that $a \in (X, s)^\sharp(K)$, $b \in (Y, t)^\sharp(K)$ are nonsingular points on X, Y respectively, and that there is some isomorphism f between (X, s) and (Y, t) such that $f(a) = b$. Then there is such an isomorphism defined over a Picard–Vessiot extension of K.*

Proof. For each $m \geq 1$, let V_m the the $\mathcal{U}$-vector space $\mathcal{M}/\mathcal{M}^{m+1}$, where $\mathcal{M}$ is the maximal ideal of the local ring of X at a. V_m is defined over K. For any automorphism h of X such that $h(a) = a$, h induces a linear automorphism h_m say of V_m. Moreover if h' is another automorphism of X fixing a, then $h = h'$ if and only if $h_m = h'_m$ for all m. So far nothing has been said about the D-variety structure. As $a \in (X, s)^\sharp$, ∂ extends to a derivation ∂' on the local ring of X at a that preserves $\mathcal{M}$ and all its powers. This gives V_m the structure of a ∂ module (V_m, D_{V_m}) over $\mathcal{U}$, defined over K (for all m). If h is an automorphism of the algebraic D-variety (X, s) that fixes a then $h_m \in \mathrm{GL}(V_m)$ restricts to a $\mathcal{C}$-linear automorphism h_m^∂ of $\mathrm{GL}(V_m^\partial)$ (where V_m^∂ is the solution space of $D_{V_m} = 0$). Moreover h_m is determined by h_m^∂.

Now we may find an isomorphism f between (X, s) and (Y, t) such that $f(a) = b$ and f is defined over $\hat{K}$. Let c generate the smallest field of definition of f. So c is a finite tuple from $\hat{K}$. Write f as f_c. For any differential automorphism σ of $\mathcal{U}$ that fixes K pointwise, $\sigma(f_c) = f_{\sigma(c)}$ is also an isomorphism of (X, s) with (Y, t) taking a to b. Hence $f_{\sigma(c)} \circ f_c^{-1}$ is an automorphism of (X, s) taking a to itself. Write h^σ for this map.

Claim I. *There is m such that for all $\sigma, \tau \in \mathrm{Aut}_\partial(\mathcal{U}/K)$, $h^\sigma = h^\tau$ if and only if $(h_m^\sigma)^\partial = (h_m^\tau)^\partial$.*

Proof. This follows from compactness and the earlier remarks as the set of h^σ is a uniformly definable family of automorphisms of (X, s). □

Now let m be as in Claim I and let d be a $\mathcal{C}$-basis for $(V_m)^\partial$.

Claim II. *$K_1 = K\langle c, d\rangle$ is a strongly normal extension of K.*

Proof. Let $\sigma \in \mathrm{Aut}_\partial(\mathcal{U}/K)$. As $\sigma(d)$ is also a basis for $(V_m)^\partial$, $\sigma(d) \in K\langle d\rangle\langle\mathcal{C}\rangle$. On the other hand, by virtue of d, $(V_m)^\partial$ can be identified with $\mathcal{C}^r$ for suitable r, and thus $(h_m^\sigma)^\partial$ can be identified with an element of $\mathrm{GL}_r(\mathcal{C})$. By Claim I, it follows that $\sigma(c) \in K\langle c, d\rangle\langle\mathcal{C}\rangle$. Thus $\sigma(K_1) \subseteq K_1\langle\mathcal{C}\rangle$. □

As in the proof of 4.1, we conclude that actually K_1 is a Picard–Vessiot extension of K. As $c \in K_1$, this gives the proposition. □

Acknowledgements. This work was carried out while the author was an Invited Professor at the University of Paris VI in June 2003, and while visiting the University of Naples II in July 2003. I am indebted to Daniel Bertrand at Paris VI for his hospitality, many helpful conversations, and his interest in the work reported here. Thanks also to Zoe Chatzidakis of University Paris VII-CNRS for her suggestions, and to Paola D'Aquino of Naples II for her hospitality.

References

[1] D. Bertrand, *Review of "Lectures on differential Galois theory", by A. Magid*, Bull. Amer. Math. Soc., **33** (1996), 289–294.

[2] A. Buium, *Differential Function Fields and Moduli of Algebraic Varieties*, Lecture Notes in Mathematics, **1226**, Springer, 1986, MR 0874111 (88e:14010), Zbl 0613.12018.

[3] A. Buium, *Differential Algebraic Groups of Finite Dimension*, Lecture Notes in Mathematics, **1506**, Springer, 1992, MR 1176753 (93i:12010), Zbl 0756.14028.

[4] E. Kolchin, *Differential Algebra and Algebraic Groups*, Pure and Applied Mathematics, **54**, Academic Press, New York, 1973, MR 0568864 (58 #27929), Zbl 0264.12102.

[5] E. Kolchin, *Differential Algebraic Groups*, Pure and Applied Mathematics, **114**, Academic Press, Orlando, 1985, MR 0776230 (87i:12016), Zbl 0556.12006.

[6] D. Marker, *Manin kernels*, in 'Connections between Model Theory and Algebraic and Analytic Geometry', Quaderni di Matematica, Univ. Naples II, 2000, MR 1930680 (2003g:12009).

[7] D. Pierce and A. Pillay, *A note on the axioms for differentially closed fields of characteristic zero*, J. Algebra, **204** (1998), 108–115, MR 1623945 (99g:12006), Zbl 0922.12006.

[8] A. Pillay, *Differential Galois theory* I, Illinois J. Math., **42** (1998), 678–699, MR 1649893 (99m:12009), Zbl 0916.03028.

[9] A. Pillay, *Finite-dimensional differential algebraic groups and the Picard–Vessiot theory*, in 'Differential Galois Theory', Banach Center Publications, **58**, Polish Academy of Sciences, 2002, 189–199, MR 1972454 (2004c:12008), Zbl 1036.12006.

[10] A. Pillay and M. Ziegler, *Jet spaces of varieties over differential and difference fields*, Selecta Math., **9**(**4**) (2003), 579–599, MR 2031753.

[11] B. Poizat, *A Course in Model Theory. An Introduction to Contemporary Mathematical Logic*, Universitext, Springer, New York, 2000, MR 1757487 (2001a:03072), Zbl 0951.03002.

[12] M. van der Put and M. Singer, *Galois Theory of Linear Differential Equations*, Grundlehren der Mathematischen Wissenschaften, **328**, Springer, Berlin, 2003, MR 1960772 (2004c:12010), Zbl 1036.12008.

Received January 20, 2004 and revised May 7, 2004. The author was supported by NSF grant DMS 0300639 and NSF Focused Research Grant DMS 0100979.

Department of Mathematics
University of Illinois at Urbana-Champaign
Urbana IL 61801-2975
E-mail address: pillay@math.uiuc.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

METRICAL DIOPHANTINE ANALYSIS FOR TAME PARABOLIC ITERATED FUNCTION SYSTEMS

Bernd O. Stratmann and Mariusz Urbański

We study various aspects of tame finite parabolic iterated function systems that satisfy a certain open set condition. The first goal in our analysis of these systems is a detailed investigation of the conformal measure on the associated limit sets. We derive a formula that describes in a uniform way the scaling of this measure at arbitrary limit points. The second goal is to provide a metrical Diophantine analysis for these parabolic limit sets in the spirit of theorems of Jarník and Khintchine in number theory. Subsequently, we show that this Diophantine analysis gives rise to refinements of the description of the conformal measure in terms of Hausdorff and packing measures with respect to certain gauge functions.

1. Introduction

For a large class of fractal sets the idea of an iterated function system has turned out to be a very convenient and efficient concept. Traditionally, the development of fractal geometry was always very much inspired by various phenomena that appear in conformal analysis and number theory. In this paper we continue this tradition by studying metrical Diophantine aspects of certain tame parabolic iterated function systems. This study generalizes results for geometrically finite Kleinian groups with parabolic elements (obtained in [**S1**], [**S2**], [**S3**], [**SV**], see also [**HV**], [**Su**]) and for parabolic rational rational functions (obtained in [**SU1**], [**SU2**]), which represent complex analytic analogues of Jarník's number theoretical theorem on well-approximable numbers ([**J**], [**B**]) and Khintchine's on a qualitative description of the 'essential support' of the 1-dimensional Lebesgue measure ([**K**]).

The paper is organized as follows: in Section 2 we first define the class of tame finite parabolic iterated function systems that satisfy the Super Strong Open Set Condition (SSOSC). We then recall a few immediate geometrical implications of the bounded distortion properties. In Section 3 we study the h-conformal measures arising from these parabolic systems. (Here, h denotes the Hausdorff dimension of the limit set associated to such a system.) We obtain a formula that describes in a uniform way the scaling of this measure

at arbitrary elements of the limit set. As a by-product we obtain an estimate on the local behaviour of the h-conformal measure at parabolic points. In Section 4 we analyse the limit sets from a Diophantine point of view. Our general approach here follows roughly the analysis given in [**S1**], [**S2**], [**SV**], [**SU1**], [**SU2**]. Nevertheless, the construction of the main tool, namely the measure μ on a Cantor-like subset of the limit set, is different. This construction is simplified and its geometrical and dynamical significance is clarified. Finally, we establish various limit laws leading up to the Khintchine Limit Law for tame parabolic iterated function systems. Subsequently, we show that these laws provide some efficient control on the fluctuations of the h-conformal measure, giving rise to refinements of the description of the h-conformal measure in terms of Hausdorff and packing measures with respect to some gauge functions.

2. Preliminaries

We begin by giving a description of our setting. Let X be a compact subset of some Euclidean space $\mathbb{R}^d$ such that X has nonempty interior and is contained in a bounded connected open set V. Suppose that there are countably many conformal maps $\phi_i : X \to X$, $i \in I$, with I having at least two elements. Then the system $S = \{\phi_i : i \in I\}$ is called a conformal iterated function system if and only if the following eight conditions are satisfied:

(1) (Open Set Condition) $\phi_i(\mathrm{Int}(X)) \cap \phi_j(\mathrm{Int}(X)) = \emptyset$ for all $i \neq j$.
(2) $|\phi_i'(x)| < 1$ everywhere except for finitely many pairs (i, x_i), $i \in I$, for which x_i is the unique fixed-point of ϕ_i and $|\phi_i'(x_i)| = 1$. Such pairs and indices i will be called parabolic and the set of parabolic indices will be denoted by Ω. All other indices will be called hyperbolic.
(3) For all $n \geq 1, \omega = (\omega_1, \ldots, \omega_n) \in I^n$ we have that if ω_n is a hyperbolic index or if $\omega_{n-1} \neq \omega_n$, then ϕ_ω admits a conformal extension to $V \subset \mathbb{R}^d$ that maps V into itself.
(4) If i is a parabolic index, then $\bigcap_{n\geq 0} \phi_{i^n}(X) = \{x_i\}$ (Hence in particular, the diameter of the set $\phi_{i^n}(X)$ tends to 0 for n tending to infinity.)
(5) (Cone Condition) There exist $\alpha, l > 0$ such that for every $x \in \partial X \subset \mathbb{R}^d$ there exists an open cone $\mathrm{Con}(x, u_x, \alpha, l) \subset \mathrm{Int}(X)$ with vertex x, $\|u_x\| = 1$ and central angle α. Here, we have set $\mathrm{Con}(x, u_x, \alpha, l) := \{y : 0 < (y - x, u_x) \leq \cos\alpha \|y - x\| \leq l\}$.
(6) There exists $0 < s < 1$ such that for all $n \geq 1, \omega \in I^n$ we have that if ω_n is a hyperbolic index or if $\omega_{n-1} \neq \omega_n$, then $\|\phi_\omega'\| \leq s$.
(7) (Bounded Distortion Property) There exists $K \geq 1$ such that for all $n \geq 1, \omega = (\omega_1, \ldots, \omega_n) \in I^n$ and $x, y \in V$ we have that if ω_n is a hyperbolic index or if $\omega_{n-1} \neq \omega_n$, then

$$|\phi_\omega'(y)| \leq K\, |\phi_\omega'(x)|.$$

(8) There are constants $L \geq 1, \alpha > 0$ such that

$$\big||\phi_i'(y)| - |\phi_i'(x)|\big| \leq L\|\phi_i'\||y-x|^\alpha \quad \text{for all } i \in I \text{ and } x, y \in V.$$

Note that if $\Omega = \emptyset$, the system S is called hyperbolic, and that if $\Omega \neq \emptyset$, then S is called parabolic. Throughout this paper we shall always assume without further notice that the system S is parabolic and the alphabet I is finite.

We now state a few immediate geometrical consequences of the bounded distortion properties (7), (8) and the cone condition (5). For the proofs of these statements we refer to [**MU1**] and [**MU3**].

For all hyperbolic words $\omega \in I^*$ and all convex subsets C of V we have

$$\operatorname{diam}(\phi_\omega(C)) \leq \|\phi_\omega'\|\operatorname{diam}(C) \tag{2.1}$$

and

$$\operatorname{diam}(\phi_\omega(V)) \leq D\|\phi_\omega'\|. \tag{2.2}$$

Here, the norm $\|\cdot\|$ is the supremum norm on V, and $D \geq 1$ denotes a universal constant. Moreover, for every $x \in X$, $0 < r \leq \operatorname{Dist}(X, \partial V)$, and for every hyperbolic word $\omega \in I^*$ we have

$$\operatorname{diam}(\phi_\omega(X)) \geq D^{-1}\|\phi_\omega'\| \tag{2.3}$$

and

$$\phi_\omega(B(x,r)) \supset B(\phi_\omega(x), K^{-1}\|\phi_\omega'\|r). \tag{2.4}$$

Also, there exists $0 < \beta \leq \alpha$ such that for all $x \in X$ and for all hyperbolic words $\omega \in I^*$

(2.5)

$$\phi_\omega(\operatorname{Int}(X)) \supset \operatorname{Con}\big(\phi_\omega(x), \beta, D^{-1}\|\phi_\omega'\|\big) \supset \operatorname{Con}\big(\phi_\omega(x), \beta, D^{-2}\operatorname{diam}\phi_\omega(V)\big),$$

where $\operatorname{Con}\big(\phi_\omega(x), \beta, D^{-1}\|\phi_\omega'\|\big)$ and $\operatorname{Con}\big(\phi_\omega(x), \beta, D^{-2}\operatorname{diam}(\phi_\omega(V))\big)$ denote some cones with vertices at $\phi_\omega(x)$, angles β, and altitudes $D^{-1}\|\phi_\omega'\|$ and $D^{-2}\operatorname{diam}(\phi_\omega(V))$ respectively. Finally, for every $\omega \in I^*$ (not necessarily hyperbolic) and every $x \in X$, there exists an altitude $l(\omega, x) > 0$ such that

$$\phi_\omega(\operatorname{Int}(X)) \supset \operatorname{Con}\big(\phi_\omega(x), \beta, l(\omega, x)\big). \tag{2.6}$$

We emphasize that for $d \geq 2$ the conditions (7) and (8) with $\alpha = 1$ can be deduced from condition (3). For $d \geq 3$, this has been shown in [**U1**]. For $d = 2$, conditions (7) and (8) follow from Koebe's distortion theorem combined with the observation that complex conjugation in $\mathbb{C}$ is an isometry.

Let I^* denote the set of all finite words in the alphabet I, and let I^∞ be the set of all infinite sequences with entries in I. By condition (3), we have $\phi_\omega(V) \subset V$, for every hyperbolic word ω. For each $\omega \in I^* \cup I^\infty$, we define the length of ω by the uniquely determined relation $\omega \in I^{|\omega|}$. If $\omega \in I^* \cup I^\infty$ and $n \leq |\omega|$, then we write $\omega|_n$ to denote the word $\omega_1\omega_2\ldots\omega_n$. In [**MU1**]

it was shown that $\lim_{n\to\infty}\sup_{|\omega|=n}\{\mathrm{diam}(\phi_\omega(X))\}=0$. Hence, the map $\pi: I^\infty \to X$, given by $\pi(\omega)=\bigcap_{n\geq 0}\phi_{\omega|_n}(X)$, is uniformly continuous. Now, the limit set $J=J_S$ of the system S can be defined as the range of the map π, that is, we define

$$J=\pi(I^\infty).$$

In order to introduce the notion of tameness we define, for every $i\in\Omega$,

$$X_i=\bigcup_{j\in I\setminus\{i\}}\phi_j(X).$$

We call a parabolic conformal iterated function system $S=\{\phi_i : i\in I\}$ tame if $x_i\notin X_i$, for every $i\in\Omega$ and $x_i\neq x_j$ if $i\neq j$. Also, we say that S satisfies the Super Strong Open Set Condition (SSOSC) if the following condition is satisfied:

$$\partial X\cap\bigcup_{i\in I}\phi_i(X)=\{x_i : i\in\Omega\}. \tag{2.7}$$

Unless stated otherwise, for the remaining part of this section we shall assume that S is a tame parabolic finite conformal iterate function system satisfying (SSOSC). The tameness of the system S and formula (2.7) imply

$$B\Big(\bigcup_{i\in I\setminus\Omega}\phi_i(X),2\hat{R}\Big)\subset \mathrm{Int}\,X. \tag{2.8}$$

Also, for each $\omega\in I^*$ and every $A\subset B(x_i,2\hat{R})$ we have that

$$\phi_\omega(A)\cap J=\phi_\omega(A\cap J). \tag{2.9}$$

Note that in order to derive the latter formula, we have to use the fact that the system S is tame. Furthermore, for all $i\in\Omega,\omega\in I^*$ we have

$$\pi^{-1}(\pi(\omega i^\infty))=\omega i^\infty. \tag{2.10}$$

Following [**MU1**], given $t\geq 0$, a Borel probability measure m is called t-conformal for the system S if $m(J)=1$ and if for every Borel set $A\subset X$ and for each $i,j\in I$ with $i\neq j$, we have

$$m(\phi_i(A))=\int_A|\phi_i'|^t dm \tag{2.11}$$

and

$$m(\phi_i(X)\cap\phi_j(X))=0. \tag{2.12}$$

Recall that a parabolic system S is called regular if and only if there exists a t-conformal measure (cf. [**MU1**]). Then $t=h$ is the Hausdorff dimension of the limit set (see [**MU1**]). Combining Theorem 1.4 in [**MU2**] and Corollary 5.8 in [**MU1**], we immediately have the following result:

Theorem 2.1. *A parabolic finite iterated function system is regular.*

Hence, since the systems we consider in this paper are finite, it follows that they are regular. The associated h-conformal measure will always be denoted by m. We shall require the following distortion properties:

Lemma 2.2. *There exists a positive constant $R^* < \hat{R}$ such that the following holds: for each hyperbolic word $\tau \in I^*$ and for every $\omega \in I^\infty$ we have that ϕ_τ is well-defined on $B(\pi(\omega), R^*)$. Moreover*

$$\frac{|\phi'_\tau(y)|}{|\phi'_\tau(x)|} \leq K \quad \textit{for all } \; x, y \in B(\pi(\omega), R^*),$$

and

$$\begin{aligned} K^{-h}|\phi'_\tau(\pi(\omega))|^h m(B(\pi(\omega), r)) &\leq m\big(\phi_\tau(B(\pi(\omega), r))\big) \\ &\leq K^h |\phi'_\tau(\pi(\omega))|^h m(B(\pi(\omega), r)) \end{aligned}$$

for every $r \in [0, R^*]$.

Proof. The statement that $\phi_\tau : B(\pi(\omega), R^*) \to \mathbb{R}^d$ is well-defined and the first distortion property of the lemma are immediate consequences of the fact that $R^* < \hat{R} < \mathrm{Dist}(X, \partial V)$ and property (7) at the beginning of this section. In order to derive the second distortion property of the lemma, choose $0 < R^* < \hat{R}$ sufficiently small such that, for each $i \in \Omega$,

$$B\big(\phi_i(X) \cap (\mathbb{R}^d \setminus B(x_i, \hat{R})), 2R^*\big) \subset \operatorname{Int} X. \tag{2.13}$$

If $\pi(\omega) \in \phi_i(X)$ for some $i \in \Omega$, and if $\|\pi(\omega) - x_i\| \geq \hat{R}$, then $B(\pi(\omega), 2R^*) \subset \operatorname{Int} X$. The proof in this case then follows immediately from a combination of the conformality of the measure m and distortion property (7). In the case that $\pi(\omega) \in \phi_i(X) \cap B(x_i, \hat{R})$, it follows that $B(\pi(\omega), R^*) \subset B(x_i, 2\hat{R})$. Using (2.9) and the conformality of m, we obtain

$$\begin{aligned} m\big(\phi_\tau((B(\pi(\omega), r))\big) &= m\big(\phi_\tau(B(\pi(\omega), r)) \cap J\big) \\ &= m\big(\phi_\tau(B(\pi(\omega), r) \cap J)\big) \\ &= \int_{B(\pi(\omega), r) \cap J} |\phi'_\tau|^h dm = \int_{B(\pi(\omega), r)} |\phi'_\tau|^h dm, \end{aligned}$$

and hence the first distortion property of the lemma gives the proof in this case. Finally, if $\pi(\omega) \notin \bigcup_{i \in \Omega} \phi_i(X)$, then $\pi(\omega) \in \phi_j(X)$ for some $j \in I \backslash \Omega$. In this case (2.8) implies that $B(\pi(\omega), 2R^*) \subset \operatorname{Int} X$, and hence the statement of the lemma follows immediately from (7) and the conformality of m. This proves the lemma. □

The following fact easily follows from the local analysis of parabolic points done in [**MU2**]:

Lemma 2.3. *Assuming that $R^* > 0$ is sufficiently small, there exists a constant $C_1 > 0$ such that for every $i \in \Omega$ and every $r \in (0, R^*]$, the intersection $J \cap B(x_i, r) \setminus \{x_i\}$ is contained in a central cone contained in* $\operatorname{Int} X$ *with vertex x_i and an angle $\le C_1 r^{p_i}$.*

As an immediate consequence of this lemma and (2.9) we get the following:

Lemma 2.4. *There exists a constant $C_1 > 0$ such that for every $i \in \Omega$, every $r \in (0, R^*]$, and every hyperbolic word ω the intersection*

$$J \cap \bar{B}\big(\phi_\omega(x_i), r|\phi'_\omega(x_i)|\big)$$

is contained in a central cone with vertex x_i and an angle $\le C_2 r^{p_i}$.

We are now in a position to prove the following distortion property:

Lemma 2.5. *There exist constants $\rho, R_* > 0$ such that for every $i \in \Omega$, $x \in J \cap B(x_i, R_*)$, and for each $\omega \in I^*$ the map ϕ_ω is well-defined on $B(x, \rho\|x - x_i\|)$ and*

$$\frac{|\phi'_\omega(z)|}{|\phi'_\omega(y)|} \le K \quad \text{for all } \ y, z \in B(x, \rho\|x - x_i\|),$$

and furthermore, for every positive $r \le \rho\|x - x_i\|$ we have

$$K^{-h}|\phi'_\omega(x)|^h m(B(x,r)) \le m\big(\phi_\omega(B(x,r))\big) \le K^h |\phi'_\omega(x)|^h m(B(x,r)).$$

Proof. In view of Lemma 2.3 there exists $R_* > 0$ and $\rho \in (0, 1/2)$ such that $B(x, 2\rho\|x - x_i\|) \subset \operatorname{Int} X$ for all $x \in J \cap B(x_i, R_*)$. Therefore, all the maps $\phi_\omega : B(x, 2\rho\|x - x_i\|) \to \operatorname{Int} X$ are well-defined, and the second part of our lemma follows from the first part. The first part in turn in the case when $d = 1$ is contained in Lemma 2.6 of [**U2**]. In the case $d = 2$ it follows immediately from Koebe's distortion theorem and the observation that the complex conjugation is an isometry. In the case $d \ge 3$ it follows from the inequality following formula (4.9) in the proof of Theorem 4.13 in [**MU2**] that, with $Y = \overline{B(x, \rho\|x - x_i\|)}$, $W = B(x, 2\rho\|x - x_i\|)$, one gets $|\phi'_\omega(z)|/|\phi'_\omega(y)| \le 4$. We are done. □

The constants R_* and R^* of Lemma 2.2 and Lemma 2.5 will be crucial in the sequel. For later use we define

$$R := \min\{R_*, R^*\}.$$

3. The geometry of conformal measures

The main result in this section is the derivation of a 'global formula' for the conformal measure associated with a tame parabolic finite iterated function system. This formula describes in a uniform way the scaling of this measure at arbitrary points in the associated limit set. Our elaboration of this formula follows closely the discussion in [**SV**] and [**SU2**], where we obtained

this type of formula for geometrically finite Kleinian groups with parabolic elements and for parabolic rational maps.

The section is split into two subsections. In the first we give an estimate for the conformal measure around parabolic points. In the second we then derive the global formula. Subsequently, as a first application of this formula, we obtain a first rough description of how the conformal measure relates to the geometric concepts Hausdorff measure and packing measure.

3.1. The conformal measure around parabolic points. We begin this subsection by recalling the following estimates for tame parabolic systems. For $d \geq 2$ a proof can be found in [**MU2**] (Section 4). For $d = 1$ the estimates are obtained immediately from the considerations in [**U2**].

Proposition 3.1. *Let S be a tame parabolic system. Then there exists a constant $Q \geq 1$ and an integer $q \geq 0$ such that for every parabolic index $i \in I$ there exists an integer $p_i \geq 1$ such that for every $j \in I \setminus \{i\}$ and for all $n, k \geq 1$ we have*

$$Q^{-1} n^{-\frac{p_i+1}{p_i}} \leq \inf_X \{|\phi'_{i^n j}|\}, \|\phi'_{i^n j}\|, \operatorname{diam}(\phi_{i^n j}(X)) \leq Q n^{-\frac{p_i+1}{p_i}}, \tag{3.1}$$

$$Q^{-1} n^{-\frac{1}{p_i}} \leq \operatorname{Dist}(x_i, \phi_{i^n}(X_i)) \leq \operatorname{Dist}(x_i, \phi_{i^n}(X_i)) \leq Q n^{-\frac{1}{p_i}}, \tag{3.2}$$

$$\operatorname{Dist}(\phi_{i^n}(X_i), \phi_{i^k}(X_i)) \leq Q \left| n^{-\frac{1}{p_i}} - k^{-\frac{1}{p_i}} \right|. \tag{3.3}$$

Furthermore, for $|n - k| \geq q$ we have

$$\operatorname{Dist}(\phi_{i^n}(X_i), \phi_{i^k}(X_i)) \geq Q \left| n^{-\frac{1}{p_i}} - k^{-\frac{1}{p_i}} \right|. \tag{3.4}$$

The following lemma gives the main result of this section:

Lemma 3.2. *Let m denote the h-conformal measure of the finite parabolic system S. For each $\kappa > 0$ there exists $C_\kappa > 0$ such that for every parabolic index i and for every $x \in \bar{J}$ we have*

$$C_\kappa^{-1} \|x - x_i\|^{h+(h-1)p_i} \leq m(B(x, \kappa\|x - x_i\|)) \leq C_\kappa \|x - x_i\|^{h+(h-1)p_i}.$$

In particular, the constant C_κ depends continuously on κ.

Proof. Since the support of m is equal to $\bar{J}$, we may assume without loss of generality that $\|x - x_i\| \leq \Delta$ for some fixed $0 < \Delta \leq R$. Let $x = \pi(\omega)$ and $\omega \in I^\infty$ be given. Then $\omega = i^n j\tau$, where $j \neq i$, $n \geq 1$, and $\tau \in I^\infty$. Assuming Δ to be chosen sufficiently small, (3.1) implies that

$$n \geq 2Q^2 \kappa^{-1}. \tag{3.5}$$

For the proof of the first inequality in the measure estimate of the lemma, let

$$T := \left\{k : \operatorname{Dist}(\phi_{i^k j}(X), \phi_{i^n j}(X)) \leq \kappa\|x - x_i\| - \operatorname{diam}(\phi_{i^n j}(X))\right\}.$$

Using (3.1), we deduce that

$$m(B(x,\kappa\|x-x_i\|)) \geq \sum_{k\in T} m(\phi_{i^k j}(X)) \sum_{k\in T} \inf\{|\phi'_{i^k j}|\}^h \geq \sum_{k\in T} Q^{-h} k^{-\frac{p_i+1}{p_i}h}.$$

Using (3.2) and (3.1), we have that if

$$Q\big|\, n^{-\frac{1}{p_i}} - k^{-\frac{1}{p_i}}\big| \leq \kappa Q^{-1} n^{-\frac{1}{p_i}} - Q n^{-\frac{p_i+1}{p_i}},$$

then it follows that $k \in T$. Hence in particular, if $k \geq n$ and if

$$Q\big(n^{-\frac{1}{p_i}} - k^{-\frac{1}{p_i}}\big) \leq \kappa Q^{-1} n^{-\frac{1}{p_i}} - Q n^{-\frac{p_i+1}{p_i}}, \tag{3.6}$$

then $k \in T$. Clearly, the statement in (3.6) is equivalent to

$$Q k^{-\frac{1}{p_i}} \geq (Q - \kappa Q^{-1}) n^{-\frac{1}{p_i}} + Q n^{-\frac{p_i+1}{p_i}}.$$

Also, (3.5) implies that $Q n^{-\frac{p_i+1}{p_i}} \leq \kappa (2Q)^{-1} n^{-\frac{1}{p_i}}$. Therefore, if $k \geq n$ and $Q k^{-\frac{1}{p_i}} \geq (Q - \kappa Q^{-1}) n^{-\frac{1}{p_i}} + \kappa (2Q)^{-1} n^{-\frac{1}{p_i}}$, or equivalently if $k \geq n$ and $k \leq \big(1 - \frac{\kappa}{2Q}\big)^{-\frac{1}{p_i}}$, then $k \in T$. It now follows that there exists a constant $\widetilde{C}_\kappa > 0$ (which depends continuously on κ) such that

$$\begin{aligned} m(B(x,\kappa\|x-x_i\|)) &\geq Q^{-h} \sum_{k=n}^{\left(1-\frac{\kappa}{2Q}\right)^{-\frac{1}{p_i}}} k^{-\frac{p_i+1}{p_i}h} \\ &\geq Q^{-h}\Big(1 - \frac{p_i+1}{p_i}h\Big)\left(\Big(1-\frac{\kappa}{2Q}\Big)^{-\frac{1}{p_i}\left(1-\frac{p_i+1}{p_i}h\right)} - 1\right) n^{-\frac{p_i+1}{p_i}h} \\ &\geq \widetilde{C}_\kappa n^{-\frac{h+(h-1)p_i}{p_i}}. \end{aligned}$$

Hence, since by (3.6) we have $\|x - x_i\| \leq Q n^{-\frac{1}{p_i}}$, it follows that

$$m(B(x,\kappa\|x-x_i\|)) \geq \widetilde{C}_\kappa Q^{h+(h-1)p_i} \|x-x_i\|^{h+(h-1)p_i}. \tag{3.7}$$

In order to prove the second inequality in the measure estimate of the lemma, note that $Q k^{-\frac{1}{p_i}} \leq (1+\kappa)\|x-x_i\|$ if and only if $k \geq \big(Q^{-1}(1+\kappa)\|x-x_i\|\big)^{-p_i}$.

Using this observation, (3.2) and (3.1), we obtain

$$\begin{aligned} m(B(x,\kappa\|x-x_i\|)) &\le m(B(x_i,(1+\kappa)\|x-x_i\|) \\ &\le \sum_{j\neq i} \sum_{k=(Q^{-1}(1+\kappa)\|x-x_i\|)^{-p_i}} m(\phi_{i^k j}(X)) \\ &\le \sum_{j\neq i} \sum_{k=(Q^{-1}(1+\kappa)\|x-x_i\|)^{-p_i}} \|\phi'_{i^k j}\|^h \\ &\le \sum_{j\neq i} \sum_{k=(Q^{-1}(1+\kappa)\|x-x_i\|)^{-p_i}} Qk^{-\frac{p_i+1}{p_i}h} \\ &\le 2Q\Big(k^{h\frac{p_i+1}{p_i}-1}\Big)^{-1} \\ &\le \big(Q^{-1}(1+\kappa)\|x-x_i\|\big)^{-p_i\left(1-\frac{p_i+1}{p_i}h\right)} \\ &= \hat{C}_\kappa \|x-x_i\|^{h+(h-1)p_i}, \end{aligned}$$

where $\hat{C}_\kappa < \infty$ denotes a positive constant depending continuously on κ. □

Corollary 3.3. *There exists a constant $C \ge 1$ such that for each $i \in \Omega$ and for all $0 < r \le 2\mathrm{diam}(X)$ we have*

$$C^{-1}\, r^{h+(h-1)p_i} \le m(B(x_i,r)) \le C\, r^{h+(h-1)p_i}.$$

Proof. Let $j \neq i$, and choose $n \ge 1$ to be the least integer such that $Q^{-1}n^{-\frac{1}{p_i}} \le r$. Let $x \in \phi_{i^{n-1}j}(X)$ be fixed. By (3.2) and Lemma 3.2, we have

$$\begin{aligned} m(B(x_i,r)) &\le m(x, 2\|x-x_i\|) \\ &\le C_2\, \|x-x_i\|^{h+(h-1)p_i} \le C_2\, Q(n-1)^{-\frac{h+(h-1)p_i}{p_i}} \\ &\ll n^{-\frac{h+(h-1)p_i}{p_i}} \ll \left(\frac{Q}{2}\right)^{h+(h-1)p_i} r^{h+(h-1)p_i}. \end{aligned}$$

Now, let $k \ge 1$ denote the least integer such that $Qk^{-\frac{1}{p_i}} \le r/2$, and let $y \in \phi_{i^k j}(X)$ be fixed. Similar as above, (3.2) and Lemma 3.2 imply that

$$\begin{aligned} m(B(x_i,r)) &\ge m(B(y, \|y-x_i|)) \\ &\ge C_1\, \|y-x_i\|^{h+(h-1)p_i} \ge C_1\, Q^{-(h+(h-1)p_i)} k^{-\frac{h+(h-1)p_i}{p_i}} \\ &\gg (k-1)^{-\frac{h+(h-1)p_i}{p_i}} \ge 2^{-(h+(h-1)p_i)} r^{h+(h-1)p_i}. \end{aligned}$$

□

Lemma 3.4. *For every $\kappa > 0$ there exists $D_\kappa \ge 1$ such that for each $i \in \Omega$, for every sufficiently small $r > 0$, and for all $x \in J \cap B(x_i, \kappa^{-1}r)$ we have*

$$D_\kappa^{-1} r^{h+(h-1)p_i} \le m(B(x,r)) \le D_\kappa r^{h+(h-1)p_i}.$$

Proof. Since $B(x,r) \subset B(x_i, \|x-x_i\|+r) \subset B(x_i,(1+\kappa^{-1})r)$, it follows from Corollary 3.3 that

$$m(B(x,r)) \le C(1+\kappa^{-1})^{h+(h-1)p_i} r^{h+(h-1)p_i}. \tag{3.8}$$

Now, if $r \le 2\|x-x_i\|$, then $r = \alpha\|x-x_i\|$ for some α such that $\kappa \le \alpha \le 2$. By Lemma 3.2, we have $C_\alpha \le \overline{C} := \sup\{C_t : t \in [\kappa,2]\} < \infty$. Hence, using Lemma 3.2 once again, it follows that

$$\begin{aligned} m(B(x,r)) = m(B(x,\alpha||x-x_i||) &\ge C_\alpha^{-1}||x-x_i||^{h+(h-1)p_i} \\ &\ge \overline{C}^{-1}\left(\frac{r}{\alpha}\right)^{h+(h-1)p_i} \\ &\ge \overline{C}\max\{\kappa^{-(h+(h-1)p_i)}, 2^{-(h+(h-1)p_i)}\} r^{h+(h-1)p_i}. \end{aligned} \tag{3.9}$$

Otherwise if $r \ge 2\|x-x_i\|$, then Corollary 3.3 implies

$$m(B(x,r)) \ge m(B(x_i,r/2)) \ge C\left(\frac{r}{2}\right)^{h+(h-1)p_i} = C2^{-(h+(h-1)p_i)} r^{h+(h-1)p_i}.$$

Combining this last estimate with (3.8) and (3.9), the lemma follows. □

3.2. The global formula for the conformal measure. An element $\omega \in I^\infty$ is called preparabolic if and only if $\sigma^k\omega = i^\infty$ for some $k \ge 0$ and some $i \in \Omega$. The set of all preparabolic elements will be denoted by I_p^∞. Also, a limit point that is not a preparabolic element will be referred to as radial, and we write I_r^∞ to denote the set of all radial points.

For each $\omega \in I^\infty$ we fix an increasing sequence of integers $\{n_j(\omega)\}_{j=1}^{k(\omega)}$ as follows: assume that $n_j(\omega)$ is defined, then we define $n_{j+1}(\omega)$ to be the smallest index greater than $n_j(\omega)$ such that either $\omega_{n_{j+1}(\omega)}$ is hyperbolic or $\omega_{n_{j+1}(\omega)} \ne \omega_{n_{j+1}(\omega)-1}$ (note that $n_1(\omega)$ is well-defined). In case $n_{j+1}(\omega)$ does not exist, then $j = k(\omega)$. Note that if $n_{j+1}(\omega) \ge n_j(\omega)+2$, then there exists a unique parabolic index $i = i(\omega,j)$ such that $\omega_l = i$ for all $n_j(\omega) \le l \le n_{j+1}(\omega)$. Furthermore, if $n_{j+1}(\omega) = n_j(\omega)+1$, then $i(\omega,j)$ denotes some arbitrary element of Ω. Observe that $k(\omega) = \infty$ if and only if $\omega \in I_r^\infty$. For each j, we define

$$r_j(\omega) := R\,\left|\phi'_{\omega|_{n_j(\omega)}}(\pi(\sigma^{n_j(\omega)}\omega))\right|,$$

and we refer to the sequence $\{r_j(\omega)\}_{j=1}^{k(\omega)}$ as the hyperbolic zoom of ω. Note that by the chain rule and by property (6) of Section 2, $\{r_j(\omega)\}_{j=1}^{k(\omega)}$ is a strictly decreasing sequence. Hence, for each $\omega \in I_r^\infty$ and every positive

$$r \le \widetilde{R} = \min\{\inf\{|\phi'_i| : i \notin \Omega\}, \inf\{|\phi'_{ij}| : i \in \Omega, j \ne i\}\},$$

there exists a unique $j \ge 1$ such that $r_{j+1}(\omega) < r \le r_j(\omega)$. For a given ω and r, the neighbours $r_{j+1}(\omega)$ and $r_j(\omega)$ thus determined in the hyperbolic zoom of ω will be denoted by $r_*(\omega)$ and $r^*(\omega)$ respectively. Also, in this situation

we shall write $i(\omega, r)$ to denote the parabolic element $i(\omega, j)$. Finally we define the function ζ, given for $\omega \in I^\infty$ and $r > 0$ by

$$\zeta(\omega, r) := \frac{m(B(x, r))}{r^h}.$$

The following theorem is the main result of this section:

Theorem 3.5 (Global formula for conformal measures)**.** *Let S be a tame parabolic finite iterated function system satisfying the* (SSOSC)*. Then, for each $\omega \in I_r^\infty$ and every $0 < r \leq \widetilde{R}$, and setting $i = i(\omega, r)$, we have*

$$\zeta(\omega, r) \asymp \begin{cases} \left(\frac{r}{r^*(\omega)}\right)^{(h-1)p_i} & \text{for } r^*(\omega) \geq r \geq r^*(\omega)\left(\frac{r_*(\omega)}{r^*(\omega)}\right)^{\frac{1}{p_i+1}} \\ \left(\frac{r_*(\omega)}{r}\right)^{h-1} & \text{for } r_*(\omega) \leq r \leq r^*(\omega)\left(\frac{r_*(\omega)}{r^*(\omega)}\right)^{\frac{1}{p_i+1}}. \end{cases}$$

Proof. Let $\omega \in I_r^\infty$ and $0 < r \leq \widetilde{R}$ be fixed. For ease of notation, throughout the proof we shall suppress the dependence on ω in some of the appearing quantities. Let j be determined by the condition $r^* = r_j$. Hence, $r_* = r_{j+1}$. By (3.2), we have

$$\|\pi(\sigma^{n_j}\omega) - x_i\| = \|\pi(\phi_{i^{n_{j+1}-n_j-1}\omega_{n_{j+1}}}(\pi(\sigma^{n_{j+1}}\omega))) - x_i\| \asymp (n_{j+1} - n_j)^{-\frac{1}{p_i}}.$$

Using the chain rule and (3.1), we obtain

$$\begin{aligned} 1 &= r_{j+1}\left|\phi'_{\omega|_{n_j}}(\pi(\sigma^{n_j}\omega))\right|^{-1}\left|\phi'_{\sigma^{n_j}\omega|_{n_{j+1}-n_j-1}}(\pi(\sigma^{n_{j+1}}\omega))\right|^{-1} \\ &\asymp \left(\frac{r_{j+1}}{r_j}\right)(n_{j+1} - n_j)^{\frac{p_i+1}{p_i}} = \left(\frac{r_*}{r^*}\right)(n_{j+1} - n_j)^{\frac{p_i+1}{p_i}}. \end{aligned}$$

Hence,

$$\left(\frac{r_*}{r^*}\right)^{\frac{1}{p_i}} \asymp \|\pi(\sigma^{n_j}\omega) - x_i\|. \tag{3.10}$$

This implies that if

$$r \geq r^*(\omega)\left(\frac{r_*(\omega)}{r^*(\omega)}\right)^{\frac{1}{p_i+1}},$$

then

$$\left|\phi'_{\omega|_{n_j}}(\pi(\sigma^{n_j}\omega))\right| \cdot \|\pi(\sigma^{n_j}\omega) - x_i\| \leq r \leq \left|\phi'_{\omega|_{n_j}}(\pi(\sigma^{n_j}\omega))\right|,$$

and hence

$$\|\pi(\sigma^{n_j}\omega) - x_i\| \leq \frac{r}{\left|\phi'_{\omega|_{n_j}}(\pi(\sigma^{n_j}\omega))\right|} \leq 1.$$

Now, using property (7) (note that $\phi_{\omega|_{n_j}}$ is hyperbolic) and Lemma 3.4, it follows that

$$\begin{aligned} m(B(\pi(\omega),r)) &\asymp \left|\phi'_{\omega|_{n_j}}(\pi(\sigma^{n_j}\omega))\right|^h m\Big(B\Big(\pi(\sigma^{n_j}\omega), r\left|\phi'_{\omega|_{n_j}}(\pi(\sigma^{n_j}\omega))\right|^{-1}\Big)\Big) \\ &\asymp r_j^h (r r_j^{-1})^{h+(h-1)p_i} = r^h \Big(\frac{r}{r_j}\Big)^{(h-1)p_i} = r^h \Big(\frac{r}{r^*}\Big)^{(h-1)p_i}. \end{aligned}$$

This proves the first case in the theorem. We are now left to consider the case in which

$$r \le r^*(\omega)\left(\frac{r_*(\omega)}{r^*(\omega)}\right)^{\frac{1}{p_i+1}}.$$

Because of (3.10) (after arranging for appropriate constants), this means that

$$\left|\phi'_{\omega|_{n_{j+1}}}(\pi(\sigma^{n_{j+1}}\omega))\right| \le r \le \rho\left|\phi'_{\omega|_{n_j}}(\pi(\sigma^{n_j}\omega))\right| \cdot \|\pi(\sigma^{n_j}\omega) - x_i\|,$$

where $0 < \rho < 1$ is the constant obtained in Lemma 2.5. Therefore, there exists $n_j \le u \le n_{j+1} - 1$ such that

$$\rho\left|\phi'_{\omega|_{u+1}}(\pi(\sigma^{u+1}\omega))\right| \cdot \|\pi(\sigma^{u+1}\omega) - x_i\| \le r \le \rho\left|\phi'_{\omega|_u}(\pi(\sigma^u\omega))\right| \cdot \|\pi(\sigma^u\omega) - x_i\|.$$

In particular, this implies that

$$r \asymp \rho\left|\phi'_{\omega|_u}(\pi(\sigma^u\omega))\right| \cdot \|\pi(\sigma^u\omega) - x_i\|. \tag{3.11}$$

Thus, by using the conformality of m, Lemma 2.5 and Lemma 3.2, it follows that

$$\begin{aligned} m(B(\pi(\omega),r)) &\asymp \left|\phi'_{\omega|_u}(\pi(\sigma^u\omega))\right|^h m\Big(B\Big(\pi(\sigma^u\omega), \rho\|\pi(\sigma^u\omega) - x_i\|\Big)\Big) \\ &\asymp \left|\phi'_{\omega|_u}(\pi(\sigma^u\omega))\right|^h \|\pi(\sigma^u\omega) - x_i\|^{h+(h-1)p_i} \\ &\asymp r^h \|\pi(\sigma^u\omega) - x_i\|^{(h-1)p_i}. \end{aligned} \tag{3.12}$$

On the other hand, the chain rule, (3.1) and (3.2) imply that

$$\begin{aligned} 1 &= r_{j+1}\left|\phi'_{\omega|_u}(\pi(\sigma^u\omega))\right|^{-1}\left|\phi'_{\sigma^u\omega|_{n_{j+1}-u-1}}(\pi(\sigma^{n_{j+1}}\omega))\right|^{-1} \\ &\asymp r_*\left|\phi'_{\omega|_u}(\pi(\sigma^u\omega))\right|^{-1}(n_{j+1}-u)^{\frac{p_i+1}{p_i}}, \end{aligned}$$

as well as

$$\|\pi(\sigma^u\omega) - x_i\|^{-(p_i+1)} \asymp (n_{j+1}-u)^{\frac{p_i+1}{p_i}}.$$

These two latter comparabilities together with (3.11) show that

$$r \asymp r_*\|\pi(\sigma^u\omega) - x_i\|^{-(p_i+1)}\|\pi(\sigma^u\omega) - x_i\| = r_*\|\pi(\sigma^u\omega) - x_i\|^{-p_i}.$$

Hence, $\|\pi(\sigma^u\omega) - x_i\| \asymp (r_*/r)^{1/p_i}$, which together with (3.12) implies that

$$m(B(\pi(\omega),r)) \asymp r^h\left(\frac{r_*}{r}\right)^{h-1}.$$

This proves the second case in the theorem. □

The following corollaries are immediate consequences of the previous theorem.

Corollary 3.6. *If $\omega \in I_r^\infty$, then for each $j \geq 1$ we have*

$$m(B(\pi(\omega), r_j(\omega))) \asymp r_j(\omega)^h.$$

Corollary 3.7. *The conformal measure m is a doubling measure. This means that for every $c > 0$ there exists $B > 0$ such that for each $z \in J$ and every $r > 0$ we have*

$$m(B(z, cr)) \leq Bm(B(z, r)).$$

Finally, as a first nontrivial application of Theorem 3.5 we derive an alternative proof of the following geometrical fact, which was obtained under slightly weaker assumptions in [**MU2**]. For this let H^t and P^t denote the t-dimensional Hausdorff and packing measure respectively.

Theorem 3.8. *If S is a tame finite parabolic system satisfying the* (SSOSC)*, then the following hold:*

(a) *If $h > 1$, then $0 < \mathrm{H}^h(J) < \infty$ and $\mathrm{P}^h(J) = \infty$.*
(b) *If $h = 1$, then $0 < \mathrm{H}^h(J), \mathrm{P}^h(J) < \infty$.*
(c) *If $h < 1$, then $0 < \mathrm{P}^h(J) < \infty$ and $\mathrm{H}^h(J) = 0$.*

Additionally, if either measure H^h or P^h is finite and positive, then its normalized version is equal to the conformal measure m.

Proof. In [**MU1**] (Lemma 5.6 and Theorem 5.7) it was shown that for a tame finite parabolic system satisfying the (SSOSC) the h-conformal measure m is atomless. This combined with Corollary 3.6 and the inverse Frostmann lemma (see [**MU3**]) implies that $\mathrm{H}^h(J) < \infty$ and $\mathrm{P}^h(J) > 0$. Now, if $h \geq 1$, then Theorem 3.5 immediately gives that, for every $x \in \pi(I_r^\infty)$,

$$\limsup_{r \to 0} \frac{m(B(x, r))}{r^h} \ll 1,$$

which implies that $\mathrm{H}^h(J) > 0$. If in addition $x = \pi(\omega)$, for $\omega \in I_r^\infty$ containing arbitrarily long blocks of i's for some $i \in \Omega$, then

$$\begin{aligned} \liminf_{r \to 0} \frac{m(B(x, r))}{r^h} &\leq \liminf_{r \to 0} \zeta \left(\omega, \rho r^*(\omega) \left(\frac{r_*(\omega)}{r^*(\omega)} \right)^{\frac{1}{p_i+1}} \right) \\ &= \liminf_{r \to 0} \left(\frac{r_*(\omega)}{r^*(\omega)} \right)^{\frac{(h-1)p_i}{p_i+1}} = 0. \end{aligned}$$

Now, by ergodicity of the measure m (see [**MU2**], Corollary 5.11) and since m is positive on open sets, it follows that m-almost everywhere we have

$$\liminf_{r \to 0} \frac{m(B(x, r))}{r^h} = 0.$$

We conclude that $\mathrm{P}^h(J) = \infty$, which proves case (a) of the theorem. Case (b) is an immediate consequence of Theorem 3.5. The proof of case (c) is analogous to the proof of case (a), and we omit it. □

4. Metrical Diophantine analysis

In this section we give a metrical Diophantine analysis for tame parabolic finite iterated function systems. In the first subsection we calculate the Hausdorff dimensions of certain subsets of the limit set that are of zero h-conformal measure. These sets are comprised of radial elements that under the system have a rather rapid approach to the parabolic points. In particular, these sets are the natural analogues of the sets of well-approximable numbers. In the second subsection we derive various limit laws that give useful approximations of the 'essential support' of the h-conformal measure associated with a tame finite parabolic iterated function system. Subsequently, we show that these laws lead to good estimates on the growth of the function ζ in the global formula (Theorem 3.5), which in turn give rise to a refined description of the conformal measure in terms of Hausdorff measures and packing measures with respect to some explicit gauge functions.

4.1. Iterated function systems in the spirit of Jarník. We first have to introduce the notion of a canonical ball. For $i \in \Omega$, $\delta > 0$ and a hyperbolic word $\omega \in I^*$, we define

$$B_\omega(i) = B_\omega = \overline{B}\big(\phi_\omega(x_i), R|\phi'_\omega(x_i)|\big),$$
$$B^\delta_\omega(i) = B^\delta_\omega = \overline{B}\big(\phi_\omega(x_i), (R|\phi'_\omega(x_i)|)^{1+\delta}\big).$$

The closed ball B_ω will be referred to as the canonical ball associated with the hyperbolic word ω.

Our main interest in this section will be focused on the sets

$$J^\delta_i := \bigcap_{q\geq 1} \bigcup_{n\geq q} \bigcup_{|\omega|=n} B^\delta_\omega(i)$$

and

$$J^\delta := \bigcup_{i\in\Omega} J^\delta_i.$$

The main result in this section is stated in the following theorem. The proof of this theorem will occupy the remaining part of this section. It will be given in several steps, some of which are formulated in separate lemmata.

Theorem 4.1. *Let $S = \{\phi_i : i \in I\}$ be a tame parabolic finite iterated function system satisfying* (SSOSC). *Then, for every $i \in \Omega$ the following hold:*

(a) *If $h \leq 1$, then*

$$\mathrm{HD}(J^\delta) = \frac{h}{1+\delta}.$$

(b) *If $h \geq 1$, then*

$$\mathrm{HD}(J_i^\delta) = \begin{cases} \dfrac{h}{1+\delta} & \text{if } \ \delta \geq h-1, \\ \dfrac{h+\delta p_i}{1+\delta(1+p_i)} & \text{if } \ \delta \leq h-1. \end{cases}$$

In particular, with $p_{\min} := \min\{p_i : i \in \Omega\}$, we have

$$\mathrm{HD}(J^\delta) = \begin{cases} \dfrac{h}{1+\delta} & \text{if } \delta \geq h-1, \\ \dfrac{h+\delta p_{\min}}{1+\delta(1+p_{\min})} & \text{if } \delta \leq h-1. \end{cases}$$

The first step in the proof is to give an upper bound for $\mathrm{HD}(J_i^\delta)$.

Lemma 4.2. *For each $i \in \Omega$ and every $\delta > 0$ we have*

$$\mathrm{HD}(J_i^\delta) \leq \min\left\{\frac{h}{1+\delta}, \frac{h+\delta p_i}{1+\delta(1+p_i)}\right\}.$$

Proof. For $n \geq 1$, let H_n denote the family of all hyperbolic words of length n. For every $\epsilon > 0$ we have

$$\begin{aligned} \mathrm{H}^{\frac{h}{1+\delta}+\epsilon}(J_i^\delta) &\leq \liminf_{q\to\infty} \sum_{n\geq q} \sum_{\omega\in H_n} \Big((R|\phi'_\omega(x_i)|)^{1+\delta}\Big)^{\frac{h}{1+\delta}+\epsilon} \\ &\leq R^{h+\epsilon(1+\delta)} \liminf_{q\to\infty} \sum_{n\geq q} \sum_{\omega\in H_n} |\phi'_\omega(x_i)|^{h+\epsilon(1+\delta)}. \end{aligned}$$

From Lemma 4.3 and Theorem 4.6 in [**MU1**] we deduce that there exists an $(h+\epsilon(1+\delta))$-semiconformal measure ν. We then apply Theorem 5.1 in [**MU1**], which gives that ν is in fact $(h+\epsilon(1+\delta))$-conformal, and that $\nu(x_j) > 0$ for some $j \in \Omega$. From the definition of the limit set J it follows that there exists a hyperbolic word $\tau \in I^*$ such that $\phi_\tau(x_j) \in B(x_i, R)$. Hence, by Lemma 2.2, we have

$$|\phi'_\omega(x_i)| \leq K|\phi'_\omega(\phi_\tau(x_j))| = K|\phi'_\tau(x_j)|^{-1}|\phi'_{\omega\tau}(x_j)|$$

for every hyperbolic word $\omega \in I^*$. Combining this estimate with the conformality of ν, it follows that for each $q \geq 0$ and every $n \geq q$ we have

$$\begin{aligned} \sum_{n\geq q} \sum_{\omega\in H_n} |\phi'_\omega(x_i)|^{h+\epsilon(1+\delta)} &\leq \big(K|\phi'_\tau(x_j)|^{-1}\big)^{h+\epsilon(1+\delta)} \sum_{n\geq q} \sum_{\omega\in H_n} |\phi'_{\omega\tau}(x_j)|^{h+\epsilon(1+\delta)} \\ &\leq \sum_{n\geq q} \sum_{|\omega|=n} \nu(\phi_{\omega\tau}(x_j))\nu(x_j)^{-1} \\ &\leq \nu(x_j)^{-1}\nu\big(\{\phi_\gamma(x_j) : |\gamma| \geq q + |\tau|\}\big) \leq \nu(x_j)^{-1}. \end{aligned}$$

Hence, $\mathrm{H}^{\frac{h}{1+\delta}+\epsilon}(J_i^\delta) \le \nu(x_j)^{-1}$, and consequently $\mathrm{HD}(J_i^\delta) \le \frac{h}{1+\delta} + \epsilon$. By letting ϵ tend to 0, we derive that $\mathrm{HD}(J_i^\delta) \le \frac{h}{1+\delta}$.

In order to obtain the second upper bound, note that by Lemma 2.4 the intersection $J \cap B(\phi_\omega(x_i), (R|\phi'_\omega(x_i)|)^{1+\delta})$ is contained in a central cone with vertex $\phi_\omega(x_i)$ and angle $\le C_2 R^{1+\delta}|\phi'_\omega(x_i)|)^{\delta p_i} \asymp (R|\phi'_\omega(x_i)|)^{\delta p_i}$, or equivalently the radius of the base $\preceq (R|\,\phi'_\omega(x_i)\,|)^{1+\delta(p_i+1)}$. Thus $J \cap B\,(\phi_\omega(x_i),\, (R|\phi'_\omega(x_i)|)^{1+\delta})$ can be covered by at most const $(R|\phi'_\omega(x_i)|)^{-\delta p_i}$ balls of radii $(R|\phi'_\omega(x_i)|)^{1+\delta(p_i+1)}$. Therefore, for every $\epsilon > 0$ we have

$$\begin{aligned}
&\mathrm{H}^{\frac{h+\delta p_i}{1+\delta(1+p_i)}+\epsilon}(J_i^\delta) \\
&\qquad \le \liminf_{q\to\infty} \sum_{n\ge q} \sum_{\omega\in H_n} \big((R|\phi'_\omega(x_i)|)^{1+\delta(1+p_i)}\big)^{\frac{h+\delta p_i}{1+\delta(1+p_i)}+\epsilon} (R|\phi'_\omega(x_i)|)^{-\delta p_i} \\
&\qquad \asymp \liminf_{q\to\infty} \sum_{n\ge q} \sum_{\omega\in H_n} |\phi'_\omega(x_i)|^{h+\epsilon(1+\delta(1+p_i))}.
\end{aligned}$$

Now the proof follows exactly in the same way as in the first part. □

As a first step towards the proof of the lower bound in Theorem 4.1, we obtain the following lemma:

Lemma 4.3. *There exists a universal constant $b(d) \ge 1$ such that the following holds: for every open set $G \subset \operatorname{Int} X$ and each $n \ge 1$, there exists a finite set $I_{G,n} \subset \bigcup_{j\ge n} I^j$ of mutually incomparable hyperbolic words, which has the properties that $m\big(\bigcup_{\omega\in I_{G,n}} B_\omega\big) \ge b(d)^{-1} m(G)$ and that the balls in $\{B_\omega : \omega \in I_{G,n}\}$ are pairwise disjoint subsets of G.*

Proof. Fix $i \in \Omega$. We define

$$J_\infty := \pi\big(\{\omega \in I^\infty \setminus \{\tau i^\infty : \tau \in I^*\} : \omega \text{ contains arbitrarily long blocks of } i\text{'s}\}\big).$$

Then, since the conformal measure m is positive on nonempty open subsets of J, Corollary 5.11 in [**MU1**] implies that $m(J_\infty) = 1$. Now, let $q \ge 1$ be sufficiently large such that $\phi_{i^q}(X) \subset B(x_i, K^{-1}R)$. It follows from the definition of J_∞ that if $x \in J_\infty$, then there exists an increasing infinite sequence $\{l_j\}_j$ with $l_j \ge n$ for all $j \ge 1$, a sequence $\{q_j\}_j$ with $q_j \ge q+1$ for all $j \ge 1$, and words $\omega^{(j)} \in I^{l_j+q_j}$ such that for all $j \ge 1$ we have $x \in \phi_{\omega^{(j)}}(X)$, $\omega^{(j)}_{l_j} \ne i$ and $\sigma^{l_j}\omega|_{q_j} = i^{q_j}$. It now follows that

$$x \in \phi_{\omega^{(j)}|_{l_j+1}}(B(x, K^{-1}R)) \subset B_{\omega^{(j)}},$$

that $\omega^{(j)}|_{l_j+1}$ is a hyperbolic word, and that $\lim_{j\to\infty} \operatorname{diam}(B_{\omega^{(j)}}) = 0$. Hence, the set $G \cap J_\infty$ can be covered by canonical balls B_ω for which $|\omega| \ge n$. Let Γ denote such a cover of $G \cap J_\infty$. By the Besicovitch Covering Theorem, there exists a universal constant $b(d) \ge 2$ such that Γ contains $b(d)/2$ subfamilies, each consisting of pairwise disjoint elements, such that G is

contained in the union of all balls in these subfamilies. It follows that for at least one of these subfamilies, say Γ_0, we have $m\big(\bigcup_{B_\omega\in\Gamma_0} B_\omega\big) \geq 2/b(d)m(G\cap J_\infty) = 2/b(d)m(G)$. Since there clearly exists a finite subset Γ_f of Γ_0 having the property that $m\big(\bigcup_{B_\omega\in\Gamma_f} B_\omega\big) \geq \frac{1}{2}m\big(\bigcup_{B_\omega\in\Gamma_0} B_\omega\big)$, the conclusion of the lemma follows. □

Proof of Theorem 4.1. Our next step in the proof of the theorem is the construction of a Cantor set contained in J_i^δ. Crucial for this will be a certain increasing sequence $\{n_l\}_{l\geq 0}$ of nonnegative integers, and it will become clear during the construction how one has to choose this sequence. We begin by defining for $l \geq 0$ the sets $I_l \subset I^*$ by induction as follows: let $B_\emptyset := \overline{B}(x_i, R)$ and $I_0 := \{\emptyset\}$. Suppose that I_l has been defined, and let $\omega \in I_l$ be fixed. By Lemma 4.3 there exists a finite set ω^* consisting of hyperbolic words such that

$$\omega^* \subset \bigcup_{k\geq \max\{|\omega|+l,\, n_{l+1}\}} I^k$$

and the family $\{B_\tau\}_{\tau\in\omega^*}$ consists of pairwise disjoint balls such that $B_\tau \subset \operatorname{Int} B_\omega^\delta$ for every $\tau \in \omega^*$ (note that $\tau|_{|\omega|} = \omega$). In addition

$$m\Big(\bigcup_{\tau\in\omega^*} B_\tau\Big) \geq \frac{1}{b(d)} m(\operatorname{Int} B_\omega^\delta) \gg m(B_\omega^\delta). \tag{4.1}$$

Here, the latter inequality follows from the conformality of m, Lemma 2.2 and Corollary 3.3. Put

$$I_{l+1} = \bigcup_{\omega\in I_l} \omega^*.$$

Now, let $\{F_l\}_{l\geq 1}$ denote the family of nested nonempty compact subsets of $B_\emptyset$ given by

$$F_l := \bigcap_{\omega\in I_l} B_\omega.$$

Note that we have in particular that

$$F = \bigcap_{l\geq 1} F_l \neq \emptyset.$$

Next, for each $l \geq 1$ we construct a Borel probability measure μ_l supported on the set F_{l-1} as follows: let $\mu_1 := \frac{1}{m(B_\emptyset)} m|_{B_\emptyset}$, and assume that the measure μ_l has already been defined for some $l \geq 1$. Recall that

$$\omega^* := \{\tau \in I_{l+1} : \tau|_{n_l} = \omega\}$$

for $\omega \in I_l$. Now, for each $\omega \in I_l$ and every Borel set $A \subset B_\omega$ we put

$$\mu_{l+1}(A) := \frac{\sum_{\tau\in\omega^*} m(A\cap B_\tau)}{\sum_{\tau\in\omega^*} m(B_\tau)}. \tag{4.2}$$

This defines a Borel probability measure μ_{l+1} on F_l having the property that $\mu_{l+1}(B_\omega) = \mu_l(B_\omega)$ for every $\omega \in I_l$. A straightforward inductive argument gives that $\mu_q(B_\omega) = \mu_l(B_\omega)$ for every $q \geq l$. Also, since for each $\omega \in \bigcup_{l\geq 0} I_l$ the set $B_\omega \cap F$ is an open subset of F , we conclude that the weak limit $\mu := \lim_{l\to\infty} \mu_l$ exists and is supported on F, and that $\mu(B_\omega) = \mu_l(B_\omega)$ for each $l \geq 1$ and every $\omega \in I_l$. For $\omega \in I_l$ and $j \leq l$, let $k_j = k_j(\omega) \leq |\omega|$ denote the unique integer determined by $\omega|_{k_j} \in I_j$. Using (4.1) and (4.2), a straightforward inductive argument gives that for every $l \geq 1$ and every $\omega \in I_l$ we have

$$
\begin{aligned}
\mu(B_\omega) = \mu_l(B_\omega) &= \prod_{j=1}^{l} \frac{m(B_{\omega|_{k_j}})}{\sum_{\tau\in\omega|^*_{k_{j-1}}} m(B_\tau)} \\
&= m(B_\omega) \prod_{j=1}^{l-1} \frac{m(B_{\omega|_{k_j}})}{\sum_{\tau\in\omega|^*_{k_j}} m(B_\tau)} \frac{1}{m(B(x_i,R))} \\
&= m(B_\omega) \prod_{j=1}^{l-1} \frac{m(B_{\omega|_{k_j}})}{m(B^\delta_{\omega|_{k_j}})} \ \exp(O(l)) \\
&= m(B_\omega) \prod_{j=1}^{l-1} \frac{|\phi'_{\omega|_{k_j}}(x_i)|^h}{|\phi'_{\omega|_{k_j}}(x_i)|^h m\big(B(x_i,(R|\phi'_{\omega|_{k_j}}(x_i)|)^\delta)\big)} \ \exp(O(l)) \\
&= m(B_\omega) \prod_{j=1}^{l-1} m\big(B(x_i,(R|\phi'_{\omega|_{k_j}}(x_i)|)^\delta)\big)^{-1} \ \exp(O(l)).
\end{aligned} \tag{4.3}
$$

For every $\eta \in \bigcup_{j\geq l-1} I_j$ define

$$
\prod_{l-1}(\eta) := \prod_{j=1}^{l-1} m\big(B(x_i,(R|\phi'_{\eta|_{k_j}}(x_i)|)^\delta)\big)^{-1}.
$$

Since

$$
\lim_{n\to\infty} \sup\big\{|\phi'_\omega(x_i)| : \omega \in I^n\big\} = 0, \tag{4.4}
$$

it follows that there exists $n_0 \geq 1$ such that for each ω with $|\omega| \geq n_0$ we have

$$
(R|\phi'_\omega(x_i)|)^{1+\delta} \leq \tfrac{1}{3} R|\phi'_\omega(x_i)|. \tag{4.5}
$$

Since the set I_l is finite, it follows from (4.4) that there exists a positive number $\widetilde{R} \leq R$ such that if $\omega \in I_l$ and if $R|\phi'_\tau(x_i)| \leq \widetilde{R}$ for some $\tau \in \omega^*$, then $|\omega| \geq n_0$. For fixed $z \in F$ and $0 < r \leq \widetilde{R}/3$, consider the family $\mathcal{F}$ of all words $\omega \in \bigcup_{l\geq 0} I_{l+1}$ for which

$$
B_\omega \cap B(z,r) \cap J \neq \emptyset, \quad R\,|\phi'_\omega(x_i)| < 3r, \quad R\,|\phi'_{\omega|_{k_l}}(x_i)| \geq 3r. \tag{4.6}
$$

We shall now see that the family $\mathcal{F}_* = \{\omega|_{k_l} : \omega \in \mathcal{F}\}$ is a singleton, and that if this is the case with $\{\gamma\} = \mathcal{F}_*$, then it follows that

$$B(z, r) \subset B_\gamma. \tag{4.7}$$

For this, fix some element $\gamma \in \mathcal{F}_*$ and $\omega \in \mathcal{F}$ such that $\gamma = \omega|_{k_l}$ and such that $y \in B_\omega \cap B(z, r) \cap J$. Clearly, by construction of the set J, we have $y \in B_\gamma^\delta$. From (4.5) and (4.6) we deduce that if $x \in B(z, r)$, then

$$\begin{aligned}\|x - \phi_\gamma(x_i)\| &\le \|x - z\| + \|z - y\| + \|y - \phi_\gamma(x_i)\| \\ &< r + r + (R|\phi'_\omega(x_i)|)^{1+\delta} \\ &\le 2r + \tfrac{1}{3}R|\phi'_\omega(x_i)| \le \tfrac{2}{3}R|\phi'_{\omega|_{k_l}}(x_i)| + \tfrac{1}{3}R|\phi'_{\omega|_{k_l}}(x_i)| \\ &= R|\phi'_\omega(x_i)| = R|\phi'_\gamma(x_i)|.\end{aligned}$$

Hence we have proved (4.7); in particular, using (4.6) and the construction of the set F, we obtain $\mathcal{F}_* = \{\gamma\}$.

Let $\epsilon > 0$ be fixed. Since $T_0 := \sup\{\prod(\tau) : \tau \in I_{l-1}\} < \infty$, we obtain for n_l sufficiently large and for all $\eta \in I_l$ that

$$T_0 \exp(O(l)) \le |\phi'_\eta(x_i)|^{-\epsilon}.$$

Combining this estimate and (4.6), it follows that

$$\prod_{l-1}(\gamma) \exp(O(l)) \le r^{-\epsilon}. \tag{4.8}$$

To complete the proof of Theorem 4.1, it now suffices to show that $\mu(B(z, r))$ can essentially be estimated from above by $r^{-2\epsilon} r^\theta$, for

$$\theta = \frac{h}{1+\delta} \quad \text{or} \quad \theta = \frac{h + \delta p_i}{1 + \delta(1 + p_i)}.$$

We split this estimate into three cases:

Case 1: $r \ge \left(R\,|\phi'_\gamma(x_i)|\right)^{1+\delta}$. Using (4.7), (4.8) and the conformality of m, we obtain

$$\begin{aligned}\mu(B(z, r)) &\le \mu(B_\gamma) = \mu_l(B_\gamma) \\ &\le m(B_\gamma) \prod_{l-1}(\gamma) \exp(O(l)) = |\phi'_\gamma(x_i)|^h \prod_{l-1}(\gamma) \exp(O(l)) \\ &\ll |\phi'_\gamma(x_i)|^h r^{-\epsilon} \ll r^{\frac{h}{1+\delta} - \epsilon},\end{aligned}$$

which completes the discussion for this case.

Before dealing with the remaining cases, note that, using (4.3), (4.6), (4.8) and Corollary 3.3, we have

$$\begin{aligned}(4.9)\quad \mu(B(z,r)) &\le \sum_{\omega\in\mathcal{F}} \mu(B_\omega) = \sum_{\omega\in\mathcal{F}} m(B_\omega) \prod_l (\omega)\exp(O(l))\\ &\le m(B(z,7r)) \prod_l (\omega)\exp(O(l))\\ &= m(B(z,7r)) \prod_{l-1} (\omega)\exp\big(O(l)\,(m(B(x_i,(R|\phi'_\gamma(x_i)|)^\delta))^{-1}\big)\\ &\ll r^{-\epsilon}\,|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)}\, m(B(z,7r)).\end{aligned}$$

Case 2: $r \le \big(R\,|\phi'_\gamma(x_i)|\big)^{1+\delta}$ and $r \ge K^2RQ^{p_i+2}\big(R\,|\phi'_\gamma(x_i)|\big)^{1+\delta+\delta p_i}$. From (4.7) and Lemma 2.2 we deduce that

$$r \le R\,|\phi'_\gamma(x_i)| \le KR\,|\phi'_{\tau|_n}(\pi(\sigma^n\tau))|,$$

where $z=\pi(\tau)$ and $\tau|_n=\gamma$. This implies that

$$(4.10)\qquad r/K \le R\,|\phi'_{\tau|_n}(\pi(\sigma^n\tau))| = R\,|\phi'_\gamma(\pi(\sigma^n\tau))|.$$

Now, since $z\in B_\gamma\cap J$, we have $z\in B^\delta_\gamma$, and therefore

$$\pi(\sigma^n\tau)\in B\big(x_i, K(R\,|\phi'_\gamma(x_i)|)^\delta\big).$$

Let $\sigma^n\tau = i^q\omega$, with $\omega_1\ne i$. By Proposition 3.1 (formula (3.2)), we have $\|\pi(\sigma^n\tau)-x_i\| \ge Q^{-1}q^{-\frac{1}{p_i}}$. Hence, using the fact that $Q^{-1}q^{-\frac{1}{p_i}} \le K(R\,|\phi'_\gamma(x_i)|)^\delta$ and Proposition 3.1 (formula (3.1)), we obtain

$$\begin{aligned}(4.11)\qquad R\,\big|\phi'_{\tau|_{n+q+1}}(\pi(\sigma^{n+q+1}\tau))\big| &= R\,\big|\phi'_\gamma(\pi(\sigma^n\tau))\big|\cdot\big|\phi'_{i^q\omega_1}(\pi(\sigma\omega))\big|\\ &\le R\,\big|\phi'_\gamma(\pi(\sigma^n\tau))\big|Qq^{-\frac{1}{p_i}}\\ &\le KRQ^{p_i+2}\,\big(R|\phi'_\gamma(x_i)|\big)^{1+\delta(p_i+1)}\\ &\le r/K.\end{aligned}$$

It follows that

$$(r/K)^* = R\,\big|\phi'_\gamma(\pi(\sigma^n\tau))\big|,\quad (r/K)_* = R\,\big|\phi'_{\tau|_{n+q+1}}(\pi(\sigma^{n+q+1}\tau))\big|.$$

Choose a small $\kappa>0$ to be specified in the course of the proof. Without loss of generality we may assume that $z\notin J_i^{\delta+\kappa}$. Thus by choosing $r>0$ to be sufficiently small, we can assume that $z\notin B_\gamma^{\delta+\kappa}$, and hence in particular that $\pi(\sigma^n\tau)\notin B\big(x_i, K^{-1}(R|\phi'_\gamma(x_i)|)^{\delta+\kappa}\big)$. Since

$$\|\pi(\sigma^n\tau)-x_i\| \le Qq^{-\frac{1}{p_i}}$$

by Proposition 3.1 (formula (3.2)), we have $Qq^{-\frac{1}{p_i}} \geq K^{-1}(R|\phi'_\gamma(x_i)|)^{\delta+\kappa}$. Hence, using Proposition 3.1 (formula (3.1)), we obtain

(4.12)

$$\begin{aligned} R\left|\phi'_{\tau|_{n+q+1}}(\pi(\sigma^{n+q+1}\tau))\right| &= R\,|\phi'_\gamma(\pi(\sigma^n\tau))|\cdot|\phi'_{i^{q}\omega_1}(\pi(\sigma\omega))| \\ &\geq K^{-1}R\,|\phi'_\gamma(x_i)|Q^{-1}q^{-\frac{1}{p_i}} \\ &\geq (R(KQ)^{p_i+2})^{-1}(R\,|\phi'_\gamma(x_i)|)^{1+(\delta+\kappa)(p_i+1)}. \end{aligned}$$

Write $r = c(R|\phi'_\gamma(x_i)|)^{1+\delta+\eta}$, for $0 \leq \eta \leq \delta p_i$ and $1 \leq c \leq K^2RQ^{p_i+2}$. Suppose first that the first part of the global formula (Theorem 3.5) holds for the centre z and radius r/K. Using (4.12), we obtain

$$\begin{aligned} &cK^{-1}(R|\phi'_\gamma(x_i)|)^{1+\delta+\eta} \\ &\geq R|\phi'_\gamma(\pi(\sigma^n\tau))|\left(\frac{R|\phi'_{\tau|_{n+q+1}}(\pi(\sigma^{n+q+1}\tau))|}{R|\phi'_\gamma(\pi(\sigma^n\tau))|}\right)^{\frac{1}{p_i+1}} \\ &\geq RK^{-1}|\phi'_\gamma(x_i)|\left((R(KQ)^{p_i+2})^{-1}\frac{(R|\phi'_\gamma(x_i)|)^{1+(\delta+\kappa)(p_i+1)}}{R|\phi'_\gamma(\pi(\sigma^n\tau))|}\right)^{\frac{1}{p_i+1}} \\ &\asymp |\phi'_\gamma(x_i)|^{1+\delta+\kappa}. \end{aligned}$$

Note that if $r > 0$ is chosen to be sufficiently small (and hence the word length of γ is large), we have $\eta \leq 2\kappa$. Then, applying Theorem 3.5, (4.9) and Corollary 3.7, we obtain

$$\begin{aligned} \mu(B(z,r)) &\ll r^{-\epsilon}|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)}m(B(z,r/K)) \\ &\asymp r^{-\epsilon}r^h|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)}\left(\frac{r/K}{(r/K)^*}\right)^{h-1} \\ &\asymp r^{-\epsilon}r^h|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)}\left(\frac{r}{|\phi'_\gamma(x_i)|}\right)^{(h-1)p_i} \\ &= r^{-\epsilon}r^{h+(h-1)p_i}|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)-(h-1)p_i} \\ &\asymp r^{-\epsilon}r^{h+(h-1)p_i}r^{\frac{-\delta(h+(h-1)p_i)-(h-1)p_i}{1+\delta+\eta}} = r^{-\epsilon}r^{\frac{hp_i\eta-p\eta+h+h\eta}{1+\delta+\eta}}. \end{aligned}$$

Note that we have

$$\frac{hp_i\eta - p\eta + h + h\eta}{1+\delta+\eta} \geq \frac{h}{1+\delta} - \epsilon \tag{4.13}$$

if and only if

$$\eta(hp_i - p_i + hp_i\delta - p_i\delta + h\delta) \geq -\epsilon(1+\delta+\eta).$$

Clearly, since $\eta \leq 2\kappa$, the latter inequality is satisfied if we choose $\kappa > 0$ to be sufficiently small. Hence, we can assume without loss of generality that

(4.13) holds. It then follows that

$$\mu(B(z,r)) \le r^{\frac{h}{1+\delta}-2\epsilon},$$

which gives the Case 2 assuming the first part of the global formula.

Now suppose that the second part of the global formula (Theorem 3.5) holds for the centre $z=\pi(\tau)$ and radius r/K. Then (4.9), Corollary 3.7 and Theorem 3.5 imply that

$$\begin{aligned}(4.14)\quad \mu(B(z,r)) &\ll r^{-\epsilon}|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)}m(B(z,r/K))\\ &\le r^{-\epsilon}|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)}r^h\left(\frac{(r/K)_*}{r}\right)^{h-1}\\ &\asymp r^{-\epsilon}r|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)p}|\phi'_{\tau|_{n+q+1}}(\pi(\sigma^{n+q+1}\tau))|^{h-1}.\end{aligned}$$

If $h\le 1$, then using (4.12), we can continue the estimate in this case as follows:

$$\begin{aligned}\mu(B(z,r)) &\ll r^{-\epsilon}r|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)p_i}|\phi'_\gamma(x_i)|^{(h-1)(1+(\delta+\kappa)(p+1))}\\ &= r^{-\epsilon}r|\phi'_\gamma(x_i)|^{h+h\kappa p_i+h\kappa-\kappa p_i-1-\delta-\kappa}\\ &= r^{-\epsilon}r|\phi'_\gamma(x_i)|^{h-1-\delta+a\kappa},\end{aligned}$$

where we have set $a:=hp_i+h-p_i-1\le 0$. Hence,

$$\mu(B(z,r)) \ll r^{-\epsilon}rr^{\frac{h-1-\delta+a\kappa}{1+\delta+\eta}} = r^{-\epsilon}r^{\frac{h+\eta+a\kappa}{1+\delta+\eta}} \le r^{-\epsilon}r^{\frac{h+a\kappa}{1+\delta}},$$

where in the last inequality we used the assumption that $h\le 1$. Now, by choosing $\kappa>0$ to be sufficiently small, it follows that

$$\mu(B(z,r)) \le r^{\frac{h}{1+\delta}-2\epsilon}.$$

This completes Case 2 for $h\le 1$.

If $h>1$, then using (4.11), we can continue the estimate in (4.14) as follows:

$$\begin{aligned}\mu(B(z,r)) &\ll r^{-\epsilon}r|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)}|\phi'_\gamma(x_i)|^{(h-1)(1+\delta(p+1))}\\ &= r^{-\epsilon}r|\phi'_\gamma(x_i)|^{h-1-\delta} = r^{-\epsilon}rr^{\frac{h-1-\delta}{1+\delta+\eta}} = r^{-\epsilon}r^{\frac{h+\eta}{1+\delta+\eta}}\\ &\le \begin{cases} r^{\frac{h}{1+\delta}-\epsilon} & \text{if } \delta\ge h-1\\ r^{\frac{h+\delta p_i}{1+\delta(1+p_i)}} & \text{if } \delta\le h-1.\end{cases}\end{aligned}$$

Here, the latter inequality is obtained by using the facts that $\eta\le\delta p_i$ and that for $\delta\le h-1$ it holds that $\frac{h+\eta}{1+\delta+\eta}$ decreases if η increases.

Hence, the proof of Case 2 is complete.

Case 3: $r \le K^2 R Q^{p_i+2}\big(R|\phi'_\gamma(x_i)|\big)^{1+\delta+\delta p_i}$. From (4.9) and Corollary 3.7 we deduce that

$$\begin{aligned}\mu(B(z,r)) &\ll r^{-\epsilon}|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)} m(B(z,r/K)) \\ &= r^{-\epsilon}|\phi'_\gamma(x_i)|^{-\delta(h+(h-1)p_i)} r^h \zeta(z,r/K) \\ &\ll r^{-\epsilon} r^h r^{-\frac{\delta(h+(h-1)p_i)}{1+\delta+\delta p_i}} \zeta(z,r/K) \\ &= r^{\frac{h+\delta p_i}{1+\delta+\delta p_i}-\epsilon}\zeta(z,r/K).\end{aligned} \tag{4.15}$$

If $h \ge 1$, then we can apply Theorem 3.5, and we obtain

$$\mu(B(z,r)) \ll r^{\frac{h+\delta p_i}{1+\delta+\delta p_i}-\epsilon}.$$

If $h \le 1$, we can assume $i = i_{\max}$, which means $p_i = \max\{p_j : j \in \Omega\}$. Let $k \ge 1$ be the index in the hyperbolic zoom associated with the point z and with the radius r/K. If $n_{k+1} = n_k+1$, then we can proceed as in the previous case to obtain the desired result. Hence, suppose that $n_{k+1} \ne n_k + 1$. It follows that $\sigma^{n_k}\tau = j^u \tau_{n_k+1}$ for some $j \in \Omega$, $u \ge 1$ and $\tau_{n_k+1} \ne j$. Now, for $t \in [(r/K)_*, (r/K)^*]$ we write $\zeta(z,t) = t^{\alpha(t)}$. Then

$$\begin{aligned}\alpha(t) &= \frac{\log \zeta(z,t)}{\log t} \\ &= \begin{cases} p_j(h-1) + \frac{p_j(1-h)\log((r/K)^*)}{\log t} & \text{for } (r/K)^* \ge r \ge (r/K)^*\left(\frac{(r/K)_*}{(r/K)^*}\right)^{\frac{1}{p_j+1}}, \\ 1-h+\frac{(h-1)\log((r/K)_*)}{\log t} & \text{for } (r/K)_* \le r \le (r/K)^*\left(\frac{(r/K)_*}{(r/K)^*}\right)^{\frac{1}{p_j+1}}. \end{cases}\end{aligned}$$

From this we deduce that α has its minimum at $t = (r/K)^*\left(\frac{(r/K)_*}{(r/K)^*}\right)^{\frac{1}{p_j+1}}$. Therefore, we can assume without loss of generality that

$$(r/K) = (r/K)^*\left(\frac{(r/K)_*}{(r/K)^*}\right)^{\frac{1}{p_j+1}}. \tag{4.16}$$

Also, by choosing $\kappa > 0$ sufficiently small, we can assume that $z \notin J^{\delta+\kappa}$. For $r > 0$ small, we then have $z \notin B^{\delta+\kappa}_{\tau|_k}(p_j)$. Now, by the same arguments as those leading to formula (4.12) in Case 2, we have

$$(r/K)_* \ge (R(KQ)^{p_i+2})^{-1}((r/K)^*)^{1+(\delta+\kappa)(p_j+1)}. \tag{4.17}$$

Hence, Theorem 3.5 and (4.16) imply that

$$\zeta(z,r/K) \le \left(\frac{(r/K)_*}{(r/K)^*}\right)^{\frac{(h-1)p_j}{p_j+1}}. \tag{4.18}$$

Now write

$$\left(\frac{(r/K)_*}{(r/K)^*}\right)^{\frac{(h-1)p_j}{p_j+1}} = (r/K)^\alpha = (r/K)^*\left(\left(\frac{(r/K)_*}{(r/K)^*}\right)^{\frac{1}{p_j+1}}\right)^\alpha$$

and for every $t \in (0,1)$ consider the number $\alpha(t)$ determined by the equation

$$\left(\frac{t}{(r/K)^*}\right)^{\frac{(h-1)p_j}{p_j+1}} = (r/K)^*\left(\left(\frac{t}{(r/K)^*}\right)^{\frac{1}{p_j+1}}\right)^{\alpha(t)}. \tag{4.19}$$

We are interested in a sufficiently good lower bound on $\alpha(r/K)$. And indeed, solving Equation (4.19) for $\alpha(t)$, one easily deduces that the function $t \mapsto \alpha(t)$ is increasing throughout the entire interval $(0,1)$. Therefore, invoking (4.17), we may assume that

$$\begin{aligned}(r/K)_* &= R(KQ)^{p_i+2})^{-1}((r/K)^*)^{1+(\delta+\kappa)(p_j+1)}\\ &\le R(KQ)^{p_i+2})^{-1}((r/K)^*)^{1+\delta(p_j+1)}.\end{aligned}$$

Combining this and (4.16), we obtain

$$(r/K) \ll (r/K)^*\big((r/K)^*\big)^\delta = \big((r/K)^*\big)^{1+\delta}.$$

Then by combining this, (4.18) and (4.17), we get

$$\zeta(z, r/K) \ll \big((r/K)^*\big)^{(\delta+\kappa)p_j(h-1)} \ll (r/K)^{\frac{p_j(h-1)(\delta+\kappa)}{1+\delta}}.$$

Substituting this latter inequality in (4.15), we obtain

$$\mu(B(z,r)) \ll r^{\frac{h+\delta p_i}{1+\delta+\delta p_i}+\frac{\delta p_j(h-1)}{1+\delta}} r^{-\epsilon+\frac{\kappa p_j(1-h)}{1+\delta}}.$$

A straightforward calculation, using the facts that $p_i \ge p_j$ and $h \ge \dfrac{p_j}{p_j+1}$, shows that

$$\frac{h+\delta p_i}{1+\delta+\delta p_i} + \frac{\delta p_j(h-1)}{1+\delta} \ge \frac{h}{1+\delta}.$$

Hence, if κ is chosen sufficiently small, we finally obtain

$$\mu(B(z,r)) \ll r^{\frac{h}{1+\delta}-\epsilon}. \qquad \square$$

4.2. Limit laws for iterated function systems. Define the set

$$I_* := \{i^n j : i \in \Omega,\, j \ne i,\, n \ge 1\} \cup (I \setminus \Omega).$$

A word $\omega \in I^\infty$ can be written uniquely as an infinite word in elements from I_* if and only if ω is not of the form τi^∞ for any $i \in \Omega$ and $\tau \in I^*$. Let

$$\sigma^* : I_*^\infty \to I_*^\infty$$

denote the shift map on I_*^∞. Also, for $i \in \Omega$ and $\omega \in I_*^\infty$ define

$$Q_i(\omega) := \begin{cases} n & \text{if } \omega_1 = i^n j \text{ for some } n \ge 1 \text{ and } j \ne i,\\ 0 & \text{otherwise.}\end{cases}$$

In [**MU1**] we proved that the iterated function system $S^* = \{\phi_\omega : \omega \in I^*\}$ is hyperbolic, and that S^* is regular if and only if S is regular. The shift map σ^* can be interpreted as the symbolic representation of the system S^*. As in the previous section, in this section we shall always assume that S is a tame parabolic finite iterated function system satisfying (SSOSC), and that m is the associated conformal measure for S. Clearly, m is also conformal for S^*. Hence, there exist Borel probability measures $\widetilde{m}$ and μ^* on I_*^∞ that are equivalent to each other (with uniformly bounded Radon–Nikodým derivatives) such that $m = \widetilde{m} \circ \pi^{-1}$ and $\mu^* \circ (\sigma^*)^{-1} = \mu^*$ (see [**MU1**]). For $\epsilon \in \mathbb{R}$, $i \in \Omega$ and $n \geq 1$, we define

$$A_{i,n}(\epsilon) := \left\{\omega \in I_*^\infty : Q_i(\omega) \geq n^{\frac{p_i}{h+(h-1)p_i}-\epsilon}\right\}$$

and

$$A_{i,\infty}(\epsilon) := \left\{\omega \in I_*^\infty : \sigma^{*n}(\omega) \in A_{i,n}(\epsilon) \text{ for infinitely many } n\right\}.$$

Lemma 4.4. *For $i \in \Omega$ and $\epsilon \in \mathbb{R}$ we have $\widetilde{m}(A_{i,\infty}(\epsilon)) > 0$ if and only if $\epsilon \geq 0$.*

Proof. Using the definition of $\widetilde{m}$ and the conformality of m, we obtain

$$\begin{aligned}\sum_{n\geq 1} \mu^*((\sigma^*)^{-n}(A_{i,n}(\epsilon))) &= \sum_{n\geq 1} \mu^*(A_{i,n}(\epsilon)) = \sum_{n\geq 1} \widetilde{m}(A_{i,n}(\epsilon)) \\ &= \sum_{n\geq 1} \sum_{k\geq n^{\frac{p_i}{h+(h-1)p_i}-\epsilon}} k^{-\frac{p_i+1}{p_i}h} \\ &\asymp \sum_{n\geq 1} n^{-1+\epsilon\frac{h+(h-1)p_i}{p_i}}.\end{aligned} \tag{4.20}$$

Since $h + (h-1)p_i > 0$ (see [**MU2**]), it follows that the series

$$\sum_{n\geq 1} \mu^*((\sigma^*)^{-n}(A_{i,n}(\epsilon)))$$

converges for $\epsilon < 0$. Thus, the "weaker part" of the Borel–Canteli lemma gives that $\mu_*(A_{i,\infty}(\epsilon)) = 0$, which then implies that $\widetilde{m}(A_{i,\infty}(\epsilon)) = 0$. This proves one direction of the equivalence in the lemma.

In order to prove the remaining part of the lemma, recall the following well-known result from elementary analysis:

- Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of events in a probability space (X, P). If $\sum_{n\in\mathbb{N}} P(X_n) = \infty$ and if $P(X_n \cap X_k) \ll P(X_n)\,P(X_k)$ for all distinct $n, k \in \mathbb{N}$, then $P\left(\limsup_{n\to\infty} X_n\right) \gg 1$.

By again using formula (4.20), the 'if-part' of the lemma follows from this general result once we have shown that for all $n, k \in \mathbb{N}$ with $n > k$ we have

$$\widetilde{m}\big((\sigma^*)^{-k}(A_{i,k}(\epsilon)) \cap (\sigma^*)^{-n}(A_{i,n}(\epsilon))\big) \ll \widetilde{m}\big((\sigma^*)^{-k}(A_{i,k}(\epsilon))\big)\, \widetilde{m}\big((\sigma^*)^{-n}(A_{i,n}(\epsilon))\big).$$

Since μ^* and $\widetilde{m}$ are equivalent, and since μ^* is σ^*-invariant, it follows that in order to obtain this latter inequality it is sufficient to show that

$$\widetilde{m}\big(A_{i,k}(\epsilon) \cap (\sigma^*)^{-(n-k)}(A_{i,n}(\epsilon))\big) \ll \widetilde{m}(A_{i,k}(\epsilon))\, \widetilde{m}(A_{i,n}(\epsilon)).$$

Since the set $A_{i,k}(\epsilon)$ can be written as a union of S^*-cylinders of length 1, it can be written also as a union of cylinders of length $(n-k)$. If $A_{i,k}(\epsilon) = \bigcup B_k(\epsilon)$ denotes such a representation by cylinders of length $(n-k)$, then by the σ^*-invariance of μ^* and by the Bounded Distortion Property (7) and the conformality of m, we have for each $\omega \in A_{i,n}(\epsilon)$ and $\tau \in B_k(\epsilon)$ that

$$\begin{aligned}\widetilde{m}\big((\sigma^*)^{-k}(A_{i,k}(\epsilon)) \cap (\sigma^*)^{-n}(A_{i,n}(\epsilon))\big) &\asymp \widetilde{m}\big((\sigma^*)^{-(n-k)}(A_{i,n}(\epsilon)) \cap B_k(\epsilon)\big) \\ &\asymp |\phi'_\tau(\pi(\omega))|^h \widetilde{m}(A_{i,n}(\epsilon)) \cap (\sigma^*)^{n-k}(B_k(\epsilon))).\end{aligned}$$

This implies that

$$\frac{\widetilde{m}((\sigma^*)^{-(n-k)}(A_{i,n}(\epsilon)) \cap B_k(\epsilon))}{\widetilde{m}(B_k(\epsilon))} \asymp \frac{|\phi'_\tau(\pi(\omega))|^h \widetilde{m}(A_{i,n}(\epsilon))}{|\phi'_\tau(\pi(\omega))|^h} = \widetilde{m}(A_{i,n}(\epsilon)),$$

or equivalently that

$$\widetilde{m}((\sigma^*)^{-(n-k)}(A_{i,n}(\epsilon)) \cap B_k(\epsilon)) \asymp \widetilde{m}(A_{i,n}(\epsilon))\, \widetilde{m}(B_k(\epsilon)).$$

If in this latter inequality we sum up over all sets $B_k(\epsilon)$, we obtain

$$\widetilde{m}((\sigma^*)^{-(n-k)}(A_{i,n}(\epsilon)) \cap A_{i,k}(\epsilon)) \asymp \widetilde{m}(A_{i,n}(\epsilon))\, \widetilde{m}(A_{i,k}(\epsilon)),$$

which in particular gives the desired inequality. □

Lemma 4.5. *For $i \in \Omega$ and $\epsilon \geq 0$ we have $\widetilde{m}(A_{i,\infty}(\epsilon)) = 1$.*

Proof. Let $i \in \Omega$ and $\epsilon > 0$ be fixed. Clearly, $\sigma^*(A_{i,\infty}(\epsilon)) \subset A_{i,\infty}(\epsilon)$. Hence, using the ergodicity of the map σ^* and the previous lemma, the statement of the lemma follows. □

Theorem 4.6 (Limit Law I)**.** *For $\widetilde{m}$-almost every $\omega \in I^\infty_*$ and for all $i \in \Omega$ we have*

$$\limsup_{n\to\infty} \frac{\log Q_i((\sigma^*)^n(\omega))}{\log n} = \frac{p_i}{h + (h-1)p_i}.$$

Proof. In order to obtain the lower bound for the 'lim sup' in the lemma, fix some $i \in \Omega$ and note that by Lemma 4.5 we have $\widetilde{m}(A_{i,\infty}(0)) = 1$. If

$\omega \in A_{i,\infty}(0)$, there exists by definition a sequence $(k_j)_{j\in\mathbb{N}}$ of natural numbers k_j, such that $(\sigma^*)^{k_j}(\omega) \in A_{i,k_j}(0)$ for all $j \in \mathbb{N}$. This implies for all j that

$$Q_i((\sigma^*)^{k_j}(\omega)) \geq k_j^{p_i/(h+(h-1)p_i)},$$

and hence that

$$\limsup_{n\to\infty} \frac{\log Q_i((\sigma^*)^n(\omega))}{\log n} \geq \frac{p_i}{h+(h-1)p_i}.$$

In order to obtain the upper bound for the 'lim sup' in the lemma, let $\epsilon < 0$ and $i \in \Omega$. By Lemma 4.4, there exists a set $F_i(\epsilon)$ such that $\widetilde{m}(F_i(\epsilon)) = 1$, and such that if $\omega \in F_i(\epsilon)$ then there exists a number $n_0 = n_0(\omega) \in \mathbb{N}$ with the property that $(\sigma^*)^n(\omega) \notin A_{i,n}(\epsilon)$ for all $n \geq n_0$. Hence, for $\omega \in F_i(\epsilon)$ we have for all $n \geq n_0$ that

$$\limsup_{n\to\infty} \frac{\log Q_i((\sigma^*)^n(\omega))}{\log n} \leq \frac{p_i}{h+(h-1)p_i} - \epsilon.$$

If we put $F_i = \bigcap_{n\geq 1} F_i(-\frac{1}{n})$, then $\widetilde{m}(F_i) = 1$ and for each $\omega \in F_i$ we have

$$\limsup_{n\to\infty} \frac{\log Q_i((\sigma^*)^n(\omega))}{\log n} \leq \frac{p_i}{h+(h-1)p_i}.$$

Hence, for $\omega \in A_{i,\infty}(0) \cap B_i$ we obtain the equality stated in the theorem. □

Note that if $Q_i(\omega) = n$, then it follows from (3.3) that $|x_i - \pi(\omega)| \asymp (n+1)^{-1/p_i}$. This now leads to our second limit law.

Theorem 4.7 (Limit Law II)**.** *For $\widetilde{m}$-almost every $\omega \in I_*^\infty$ we have for all $i \in \Omega$ that*

$$\limsup_{n\to\infty} \frac{-\log|\pi((\sigma^*)^n(\omega)) - x_i|}{\log n} = \frac{1}{h+(h-1)p_i}.$$

Proof. Fix $\omega \in I_*^\infty$ and $i \in \Omega$. By definition of Q_i and using (3.3), we have for $n \in \mathbb{N}$ that

$$|\pi((\sigma^*)^n(\omega)) - x_i| \asymp (Q_i((\sigma^*)^n(\omega)) + 1)^{-1/p_i}.$$

Hence, it follows that

$$\lim_{n\to\infty} \left| \frac{-\log|\pi((\sigma^*)^n(\omega)) - x_i|}{\log n} - \frac{\log Q_i((\sigma^*)^n(\omega))}{p_i \log n} \right| = 0.$$

Using Limit Law I, we find that, for $\widetilde{m}$-almost all $\omega \in I_*^\infty$,

$$\begin{aligned}\limsup_{n\to\infty} \frac{-\log|\pi((\sigma^*)^n(\omega)) - x_i|}{\log n} &= \frac{1}{p_i} \limsup_{n\to\infty} \frac{\log Q_i((\sigma^*)^n(\omega))}{\log n} \\ &= \frac{1}{h+(h-1)p_i}.\end{aligned}$$

□

Since $\widetilde{m}$ is ergodic and positive on nonempty open sets, we have that $\widetilde{m}$-almost every point in I_*^∞ has arbitrarily long blocks with parabolic entries only. Taking this observation into account, we now modify on a set of full measure the definition of the hyperbolic zoom $(r_j(\omega))_j$ as follows: for a given $i \in \Omega$ we include only those elements in the hyperbolic zoom for which $n_j(\omega) \geq n_{j-1}(\omega) + 2$ and $i(\omega, j) = i$. In other words, we consider subsequences $(r_{j_k}(\omega))_k$ and $(n_{j_k}(\omega))_k$ such that $n_{j_k}(\omega) \geq n_{j_k-1}(\omega) + 2$ and $\omega_{n_{j_k-1}(\omega)} = i$. Such subsequences will be referred to as the i-restricted hyperbolic zoom and the i-restricted optimal sequence, respectively.

Theorem 4.8 (Limit Law III)**.** *For each $i \in \Omega$ the i-restricted optimal sequence at $\widetilde{m}$-almost every $\omega \in I_*^\infty$ has the property that*

$$\limsup_{k\to\infty} \frac{\log(n_{j_k+1}(\omega) - n_{j_k}(\omega))}{\log j_k} = \frac{p_i}{h + (h-1)p_i}.$$

Proof. Let $i \in \Omega$ and $\omega \in I_*^\infty$. Define the function $N_n : I_*^\infty \to \mathbb{N}$ by $(\sigma^*)^n(\omega) = \sigma^{N_n(\omega)}(\omega)$, for every $n \geq 1$. Then we see by induction that $N_j(\omega) = n_j(\omega)$, for all $j \in \mathbb{N}$ (this follows, since $n_1(\omega) = N_1(\omega)$ and, assuming that $n_j(\omega) = N_j(\omega)$, since $n_{j+1}(\omega) = n_j(\omega) + N_1(\omega)(\sigma^{n_j(\omega)}(\omega)) = N_{j+1}(\omega)$).

Using Limit Law II and the fact that $|\pi(\sigma^{N_{j_k}(\omega)}(\omega)) - x_i| \asymp (N_{j_k+1}(\omega) - N_{j_k}(\omega))^{-1/p_i}$, it follows that for $\widetilde{m}$-almost all ω we have

$$\begin{aligned}
\limsup_{k\to\infty} \frac{\log(n_{j_k+1}(\omega) - n_{j_k}(\omega))}{\log j_k} &= \limsup_{k\to\infty} \frac{\log(N_{j_k+1}(\omega) - N_{j_k}(\omega))}{\log j_k} \\
&= \limsup_{k\to\infty} \frac{-p_i \ \log|\pi(\sigma^{N_{j_k}(\omega)}(\omega)) - x_i|}{\log j_k} \\
&= \limsup_{k\to\infty} \frac{-p_i \ \log|\pi((\sigma^*)^{j_k}(\omega)) - x_i|}{\log j_k} \\
&= \frac{p_i}{h + (h-1)p_i}.
\end{aligned}$$

□

Theorem 4.9 (Limit Law IV)**.** *For each $i \in \Omega$ the i-restricted hyperbolic zoom at $\widetilde{m}$-almost every $\omega \in I_*^\infty$ has the property that*

$$\limsup_{k\to\infty} \frac{\log\left(r_{j_k}(\omega) \,/\, r_{j_k+1}(\omega)\right)}{\log j_k} = \frac{1 + p_i}{h + (h-1)p_i}.$$

Proof. For $i \in \Omega$ and $\omega \in I_*^\infty$ we saw in the proof of Theorem 3.5 that

$$\frac{r_{j_k}(\omega)}{r_{j_k+1}(\omega)} \asymp (n_{j_k+1}(\omega) - n_{j_k}(\omega))^{(1+p_i)/p_i}$$

for $k \in \mathbb{N}$. Combining this estimate with Limit Law III, it follows for $\widetilde{m}$-almost all $\omega \in I_*^\infty$ that

$$\limsup_{k\to\infty} \frac{\log\left(r_{j_k}(\omega)\,/\,r_{j_k}(\omega)\right)}{\log j_k} = \limsup_{k\to\infty} \frac{1+p_i}{p_i}\ \frac{\log(n_{j_k+1}(\omega)-n_{j_k}(\omega))}{\log j_k}$$
$$= \frac{1+p_i}{h+(h-1)p_i}. \qquad \square$$

The following theorem presents the main result in this section:

Theorem 4.10. (The Khintchine Limit Law for parabolic iterated function systems)**.** *The hyperbolic zoom at $\widetilde{m}$-almost every $\omega \in I_*^\infty$ satisfies*

$$\limsup_{j\to\infty} \frac{\log\left(r_j(\omega)\,/\,r_{j+1}(\omega)\right)}{\log\log \frac{1}{r_j(\omega)}} = \frac{1+p_{\max}}{h+(h-1)p_{\max}},$$

where we have set $p_{\max} := \max\{p_i : i \in \Omega\}$.

Proof. For $\widetilde{m}$-almost all $\omega \in I_*^\infty$ we have

$$\lim_{j\to\infty} \frac{\log r_j(\omega)}{j} = \lim_{j\to\infty} \frac{\log\left|\phi'_{n_j(\omega)}(\pi(\sigma^{n_j(\omega)}(\omega)))\right|}{j}$$
$$= \lim_{j\to\infty} \frac{\log\left|\phi'_{N_j(\omega)}(\pi(\sigma^{N_j(\omega)}(\omega)))\right|}{j}$$
$$= \lim_{j\to\infty} \frac{\log\left|\phi'_{N_j(\omega)}(\pi(\sigma^*)^j(\omega))\right|}{j} = \chi,$$

where the last equality follows from the Birkhoff Ergodic Theorem, using the ergodicity of the system $(I_*^\infty, \sigma^*, \mu^*)$ and the definition

$$\chi := \int_{I_*^\infty} \log\left|\phi'_{\omega_1}(\pi(\sigma^*)(\omega))\right| \, dm^*(\omega) > -\infty.$$

Hence,

$$\lim_{j\to\infty} \frac{\log\log \frac{1}{r_j(\omega)}}{\log j} = 1.$$

The theorem follows by combining this equality with Limit Law IV and noting that

$$\max_{i\in\Omega} \frac{1+p_i}{h+(h-1)p_i} = \frac{1+p_{\max}}{h+(h-1)p_{\max}}. \qquad \square$$

Corollary 4.11. *For the function ζ of the h-conformal measure m (see Theorem* 3.5*) associated with a tame parabolic finite iterated function system satisfying* (SSOSC) *the following hold:*

(i) *For $h = 1$, we have for all $\omega \in I_*^\infty$ and $0 < r < \operatorname{diam}(I_*^\infty)$ that*

$$\zeta(\omega, r) \asymp 1.$$

(ii) *For $h < 1$, we have for $\widetilde{m}$-almost every $\omega \in I_*^\infty$ that*

$$\limsup_{r\to 0} \frac{\log \zeta(\omega, r)}{\log\log \frac{1}{r}} = \frac{(1-h)p_{\max}}{h + (h-1)p_{\max}}.$$

(iii) *For $h > 1$, we have for $\widetilde{m}$-almost every $\omega \in I_*^\infty$ that*

$$\liminf_{r\to 0} \frac{\log \zeta(\omega, r)}{\log\log \frac{1}{r}} = \frac{(1-h)p_{\max}}{h + (h-1)p_{\max}}.$$

Proof. Statement (i) is an immediate consequence of Theorem 3.5. In order to prove statement (ii), let $\omega \in I_*^\infty$ and $r > 0$ sufficiently small be given. Without loss of generality we may assume that $r_{j+1}(\omega) \le r < r_j(\omega)$ and that $\omega_{n_j(\omega)+1} = i$, for some $i \in \Omega$. For r in this range, an elementary calculation shows that the maximal value of $\zeta(\omega, r)$ is achieved if r is comparable to

$$r_{j,\max}(\omega) := r_j(\omega) \left(\frac{r_{j+1}(\omega)}{r_j(\omega)} \right)^{1/(1+p_i)}.$$

For this value of r we have

$$\zeta(\omega, r_{j,\max}(\omega)) \asymp \left(\frac{r_j(\omega)}{r_{j+1}(\omega)} \right)^{(1-h)p_i/(1+p_i)}.$$

As we have seen above in the proof of the Khintchine law, for $\widetilde{m}$-almost all $\omega \in I_*^\infty$ it is sufficient to restrict the discussion to those indices j for which $\omega_{n_j(\omega)} = i$, with $p_i = p_{\max}$. It follows that for all $\epsilon > 0$ and for m-almost all $\omega \in I_*^\infty$ we eventually have

$$\frac{(1-\epsilon)\,(1+p_i)}{h + (h-1)p_i} \log\log \frac{1}{r_j(\omega)} \le_{\text{i.o.}} \log \frac{r_j(\omega)}{r_{j+1}(\omega)} \le \frac{(1+\epsilon)(1+p_i)}{h + (h-1)p_i} \log\log \frac{1}{r_j(\omega)}$$

(where '$\le_{\text{i.o.}}$' indicates that the inequality holds 'infinitely often', i.e., for some infinite sequence of values of j). Hence, the estimate above implies that

$$\left(\log \frac{1}{r_j(\omega)} \right)^{\frac{(1-\epsilon)(1-h)p_{\max}}{h+(h-1)p_{\max}}} \ll_{\text{i.o.}} \zeta(\omega, r_{j,\max}(\omega)) \ll \left(\log \frac{1}{r_j(\omega)} \right)^{\frac{(1+\epsilon)(1-h)p_{\max}}{h+(h-1)p_{\max}}}.$$

This proves statement (ii) in the corollary. Statement (iii) follows from a similar argument, and we omit its proof. □

We are now in the position to derive a refinement of the description of the geometric nature of the h-conformal measure given in Theorem 3.8 . Namely, using the latter corollary, we have the following statements concerning its relationship to the packing measure $\mathrm{P}_{\psi_\lambda}$ and Hausdorff measure $\mathrm{H}_{\psi_\lambda}$ with respect to the dimension function ψ_λ. Here, the function ψ_λ is given for $\lambda \in \mathbb{R}$ and positive r by

$$\psi_\lambda(r) := r^h \left(\log \frac{1}{r} \right)^{(1+\lambda)(1-h)p_{\max}/(h+(h-1)p_{\max})}.$$

Corollary 4.12. *If S is a tame finite parabolic iterated function system satisfying* (SSOSC), *we have the following table:*

λ vs. h	$h<1$	$h>1$
$\lambda>0$	$m \ll \mathcal{H}_{\psi_\lambda}$ *and* $\mathcal{H}_{\psi_\lambda}(J)=\infty$	$\exists E_\lambda$ *s.t.* $m(E_\lambda)=1$, $\mathcal{P}_{\psi_\lambda}(E_\lambda)=0$
$\lambda\le 0$	$\exists F_\lambda$ *s.t.* $m(F_\lambda)=1$, $\mathcal{H}_{\psi_\lambda}(F_\lambda)=0$	$m \ll \mathcal{P}_{\psi_\lambda}$ *and* $\mathcal{P}_{\psi_\lambda}(J)=\infty$

The symbol '$\ll$' indicates absolute continuity between two measures.

Acknowledgements. We thank the Stochastic Institute at the University of Göttingen, the Mathematics Department at the University of Warwick and IMPA in Rio de Janeiro for their warm hospitality and excellent working conditions when writing this paper.

References

[B] A.S. Besicovitch, *Sets of fractional dimension* (IV): *On rational approximation to real numbers*, Jour. London Math. Soc., **9** (1934), 126–131, Zbl 0009.05301.

[HV] R. Hill and S.L. Velani, *The Jarník–Besicovitch Theorem for geometrically finite Kleinian groups*, Proc. London Math. Soc. (3), **77** (1998), 524–550, MR 1643409 (99f:11100), Zbl 0924.11063.

[J] V. Jarník, *Diophantische Approximationen und Hausdorffsches Maß*, Mathematicheskii Sbornik, **36** (1929), 371–382, Zbl 55.0719.01.

[K] A. Khintchine, *Continued Fractions*, Noordhoff, Groningen, 1963, MR 0161834 (28 #5038). Translation of *Tspenye dobni*, 3rd ed., Nauka, Moscow, 1961.

[MU1] R. D. Mauldin and M. Urbański, *Parabolic iterated function systems*, Ergod. Th. & Dynam. Sys., **20** (2000), 1423–1447, MR 1786722 (2001m:37047), Zbl 0982.37045.

[MU2] R. D. Mauldin, M. Urbański, *Fractal measures for parabolic* IFS, Adv. in Math., **168** (2002), 225–253, MR 1912133 (2003e:28023), Zbl 1013.28007.

[MU3] R. D. Mauldin, M. Urbański, *Graph Directed Markov Systems: Geometry and Dynamics of Limit Sets*, Cambridge Tracts in Mathematics, **148** Cambridge University Press, Cambridge, 2003, MR 2003772, Zbl 1033.37025.

[S1] B.O. Stratmann, *Fractal dimensions for Jarník limit sets of geometrically finite Kleinian groups; the semi-classical approach*, Ark. Math., **33** (1995), 385–403, MR 1373031 (97a:30056), Zbl 0851.30027.

[S2] B.O. Stratmann, *Weak singularity spectra of the Patterson measure for geometrically finite Kleinian groups with parabolic elements*, Michigan Math. J., **46** (1999), 573–587, MR 1721571 (2001a:37059), Zbl 0961.30033.

[S3] B.O. Stratmann, *Multiple fractal aspects of conformal measures; a survey*, in 'Workshop on Fractals and Dynamics', eds. M. Denker, S.-M. Heinemann and B.O. Stratmann, Math. Gottingensis, Heft **5** (1997), 65–71.

[SU1] B.O. Stratmann and M. Urbański, *Jarník and Julia; a Diophantine analysis for parabolic rational maps for geometrically finite Kleinian groups with parabolic elements*, Math. Scan., **91** (2002), 27–54, MR 1917680, Zbl 1019.37028.

[SU2] B.O. Stratmann, M. Urbański, *The geometry of conformal measures for parabolic rational maps*, Math. Proc. Cambridge Phil. Soc., **128** (2000), 141–156, MR 1724435 (2000i:37066), Zbl 0974.37029.

[SV] B.O. Stratmann and S. Velani, *The Patterson measure for geometrically finite groups with parabolic elements, new and old*, Proc. London Math. Soc. (2), **71** (1995), 197–220, MR 1327939 (97f:58023), Zbl 0821.58026.

[Su] D. Sullivan, *Disjoint spheres, approximation by imaginary quadratic numbers, and the logarithm law for geodesics*, Acta Math., **149** (1983), 215–237, MR 0688349 (84j:58097), Zbl 0517.58028.

[U1] M. Urbański, *Rigidity of multi-dimensional conformal iterated function systems*, Nonlinearity, **14** (2001), 1593–1610, MR 1867094 (2003c:37031), Zbl 1001.37023.

[U2] M Urbański, *Parabolic Cantor sets*, Fund. Math., **151** (1996), 241–277, MR 1424576 (98f:58076), Zbl 0895.58036.

Received January 22, 2002. The research of the second author was supported in part by the NSF Grant DMS 9801583.

Mathematical Institute
University of St Andrews
St Andrews KY16 9SS
Scotland
E-mail address: bos@maths.st-and.ac.uk

Department of Mathematics
University of North Texas
Denton TX 76203-1430
E-mail address: urbanski@dynamics.math.unt.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 216, No. 2, 2004

A NOTE ON A FOURTH ORDER PDE WITH CRITICAL NONLINEARITY

Changyou Wang

We consider the Euler–Lagrange equation of a functional arising from conformal geometry in four dimensions, a fourth order equation with borderline nonlinearity. We present a short proof of the fact that any $W^{2,2}$-solution is smooth.

1. Introduction

Let (M, g) be a four-dimensional compact Riemannian manifold. Motivated by problems in four-dimensional spectral theory and conformal geometry, Chang and Yang [**CY**] (cf. also Chang [**C**]) introduced the functional $F : W^{2,2}(M) \to \mathbf{R}$:

(1.1)
$$F(w) = \int_M \left\{(\Delta w)^2 + (\alpha\Delta w + \beta|Dw|^2)^2 + T(Dw, Dw) + E(w - \overline{w})\right\} dv,$$

where $\alpha, \beta \in \mathbf{R}$, $\overline{w} = \frac{1}{\operatorname{vol} M}\int w$; $E : \mathbf{R} \to \mathbf{R}$ and $T \in \operatorname{sym}^2(T^*M)$ satisfy:

$$\max\left\{|E(x)|, |E'(x)|\right\} \le c_1 e^{c_2|x|}, \quad |T(v, v)| \le c_3|v|^2. \tag{1.2}$$

Direct computations show that the Euler–Lagrange equation associated with critical points of F on $W^{2,2}(M)$ is

$$\begin{aligned} 2(1+\alpha^2)\Delta^2 w + 2\beta \operatorname{div}\left(\alpha D(|Dw|^2) - (\alpha\Delta w + \beta|Dw|^2)Dw\right) \\ = \operatorname{div}(T(Dw, \cdot)) - \left(E'(w - \overline{w}) - \overline{E}'(w - \overline{w})\right) \end{aligned} \tag{1.3}$$

where $\overline{E}'(w - \overline{w}) = \frac{1}{\operatorname{vol} M}\int E'(w - \overline{w})$. Chang, Gursky and Yang [**CGY**] proved that any F-minimizing solution $u \in W^{2,2}(M)$ to (1.3) is actually smooth. It was asked in [**CGY**] whether any weak solution $u \in W^{2,2}(M)$ is smooth. Indeed, Uhlenbeck and Viaclovsky [**UV**] confirmed this recently and proved the smoothness for any weak solution $u \in W^{2,2}(M)$ to (1.3). The proof in [**CGY**] relied on F-minimality. The idea in [**UV**] is based on some uniqueness properties for small perturbations of Δ^2 in various Sobolev spaces and seems to be an indirect argument. Here we provide an alternative and direct proof of the smoothness for weak solutions to (1.3); namely, we show that under a smallness assumption on the $W^{2,2}$ norm, the normalized L^p-norm of the gradient of u on a ball decays like a positive power of the radius

of the ball. This, combined with Morrey's decay lemma and the conformal invariance of the $W^{2,2}$ norm in dimension four, implies the Hölder continuity of u. Higher-order regularity then follows from [**CGY**]. This type of so-called ϵ_0-decay lemma is very common in the context of regularity theory for harmonic maps (cf. Schoen–Uhlenbeck [**SU**]). In fact, this kind of idea was also employed by Chang, Wang and Yang in their study of the regularity problem of biharmonic maps into spheres [**CWY**].

Since regularity is a local result, we assume, for simplicity, that $M = \Omega \subset \mathbf{R}^4$ is a bounded smooth domain, with the Euclidean metric g. Now we state the decay lemma:

Lemma A. *There exist $\epsilon_0 > 0$ and $\theta_0 \in (0, \frac{1}{2})$ such that if $u \in W^{2,2}(\Omega)$ is a weak solution to* (1.3) *and if for $B_r(x) \subset \Omega$ we have*

$$\int_{B_r(x)} |Du|^4 + |D^2u|^2 \le \epsilon_0^2 \tag{1.4}$$

then, for any $2 < p < 4$,

$$(\theta_0 r)^{p-4} \int_{B_{\theta_0 r}(x)} |Du|^p \le \tfrac{1}{2} r^{p-4} \textstyle\int_{B_r(x)} |Du|^p + C\big(p, \|D^2u\|_{L^2(\Omega)}\big) r^p. \tag{1.5}$$

Since $u \in W^{2,2}(\Omega)$, the absolute continuity of $\int |Du|^4 + |D^2u|^2$ implies that there exists an $r_0 > 0$ such that (1.4) holds for u over any ball $B_r(x) \subset \Omega$ with $0 < r \le r_0$. Therefore, we can apply the lemma repeatedly and conclude that there exists a $\delta_0 \in (0, 1)$ such that $r^{p-4} \int_{B_r(x)} |Du|^p$ behaves like $r^{p\delta_0}$ for all $0 < r < r_0$ and $x \in \Omega$. This, combined with Morrey's lemma, implies that $u \in C^{\delta_0}(\Omega)$ and hence $u \in C^\infty(\Omega)$, via [**CGY**]. In particular, one has (cf. also [**UV**]):

Theorem B. *If $u \in W^{2,2}(M)$ is a weak solution to* (1.3)*, then $u \in C^\infty(M)$.*

2. Proof of Lemma A

It follows from Fubini's theorem that there is an $s \in [\frac{r}{2}, r]$ such that

$$\int_{\partial B_s(x)} |Du|^4 + |D^2u|^2 \le 2r^{-1} \int_{B_r(x)} |Du|^4 + |D^2u|^2. \tag{2.1}$$

Let $u_1 \in W^{2,2}(B_s(x))$ satisfy

$$\Delta^2 u_1 = -\frac{\beta}{1+\alpha^2} \operatorname{div}\big(\alpha D(|Du|^2) - (\alpha \Delta u + \beta |Du|^2)\, Du\big), \tag{2.2}$$

$$u_1 = \frac{\partial u_1}{\partial r} = 0 \quad \text{on } \partial B_s(x). \tag{2.3}$$

Let $u_2 \in W^{2,2}(B_s(x))$ satisfy

$$\Delta^2 u_2 = \frac{1}{2(1+\alpha^2)}\big(\operatorname{div}(T(Du,\cdot)) - (E'(u-\overline{u}) - \overline{E}'(u-\overline{u}))\big), \tag{2.4}$$

$$u_2 = \frac{\partial u_2}{\partial r} = 0 \quad \text{on } \partial B_s(x). \tag{2.5}$$

Let $u_3 = u - u_1 - u_2 \in W^{2,2}(B_s(x))$. Then we have

$$\begin{aligned} \Delta^2 u_3 &= 0 && \text{in } B_s(x), \\ u_3 &= u \quad \text{and} \quad \frac{\partial u_3}{\partial r} = \frac{\partial u}{\partial r} && \text{on } \partial B_s(x). \end{aligned} \tag{2.6}$$

For u_1, it follows (see, e.g., Lemma 2.2 of [**CWY**]) that for any $q \in (1, \frac{4}{3})$

$$\begin{aligned} \|D^3 u_1\|_{L^q(B_s(x))} &\le C\||Du||D^2u| + |Du|^2|Du|\|_{L^q(B_s(x))} \\ &\le C\big(\|D^2u\|_{L^2(B_s(x))} + \|Du\|^2_{L^4(B_s(x))}\big)\|Du\|_{L^{\frac{2q}{2-q}}(B_s(x))} \\ &\le C\epsilon_0\|Du\|_{L^{\frac{2q}{2-q}}(B_s(x))}. \end{aligned}$$

This, combined with the Sobolev embedding theorem, implies

$$\begin{aligned} \|Du_1\|_{L^{\frac{2q}{2-q}}(B_{\frac{r}{2}}(x))} \le \|Du_1\|_{L^{\frac{2q}{2-q}}(B_s(x))} &\le C\epsilon_0\|Du\|_{L^{\frac{2q}{2-q}}(B_s(x))} \\ &\le C\epsilon_0\|Du\|_{L^{\frac{2q}{2-q}}(B_r(x))}. \end{aligned} \tag{2.7}$$

Here we have used the fact that $Du_1 = 0$ on $\partial B_s(x)$. To estimate u_2, observe that (1.2) implies that $|T(Du,\cdot)| \le C|Du| \in L^4(\Omega)$ and

$$\|T(Du,\cdot)\|_{L^4(\Omega)} \le C\|u\|_{W^{2,2}(\Omega)}. \tag{2.8}$$

The Moser–Trudinger inequality and (1.2) imply $E'(u-\overline{u}) - \overline{E}'(u-\overline{u}) \in L^p(\Omega)$ for any $1 < p < \infty$ and

$$\big\|E'(u-\overline{u}) - \overline{E}'(u-\overline{u})\big\|_{L^4(\Omega)} \le C\|u\|_{W^{2,2}(\Omega)}. \tag{2.9}$$

Multiplying (2.4) by u_2 and integrating it over $B_s(x)$, we get

$$\begin{aligned} \int_{B_s(x)} |D^2u_2|^2 &= \int_{B_s(x)} |\Delta u_2|^2 \\ &\le C\int_{B_s(x)} \big(|Du|\,|Du_2| + |E'(u-\overline{u}) - \overline{E}'(u-\overline{u})|\,|u_2|\big) \\ &\le C\|u\|_{W^{2,2}(\Omega)}\left(\int_{B_s(x)} (|u_2|^2 + |Du_2|^2)\right)^{\frac{1}{2}} \\ &\le C\|u\|_{W^{2,2}(\Omega)} r\left(\int_{B_s(x)} |D^2u_2|^2\right)^{\frac{1}{2}}. \end{aligned}$$

Here we have applied the Poincaré inequality for u_2 in the last step. Thus

$$\int_{B_s(x)} |D^2 u_2|^2 \le C\|u\|^2_{W^{2,2}(\Omega)} r^2. \tag{2.10}$$

This, combined with the Sobolev embedding theorem, gives

$$\int_{B_s(x)} |Du_2|^4 \le C\bigg(\int_{B_s(x)} |D^2 u_2|^2\bigg)^2 \le Cr^4\|u\|^2_{W^{2,2}(\Omega)}. \tag{2.11}$$

In particular, for any $q \in (1, \frac{4}{3})$, we have

$$\Big(\frac{r}{2}\Big)^{\frac{2q}{2-q}-4} \int_{B_{\frac{r}{2}}(x)} |Du_2|^{\frac{2q}{2-q}} \le C(\|u\|_{W^{2,2}(\Omega)}) r^{\frac{2q}{2-q}}. \tag{2.12}$$

Since u_3 is a biharmonic function on $B_s(x)$, we know that

$$\int_{B_s(x)} |D^2 u_3|^2 \le \int_{B_s(x)} |D^2 u|^2.$$

A standard Caccipolli-type argument implies that

$$\int_{B_{\frac{r}{3}}(x)} |D^2 u_3|^2 \le Cr^{-2} \int_{B_{\frac{r}{2}}(x)} |Du_3|^2. \tag{2.13}$$

This, combined with the subharmonicity of $|\Delta u_3|^2$, implies

$$r^2\|Du_3\|^2_{L^\infty(B_{\frac{r}{4}}(x))} \le C \int_{B_{\frac{r}{3}}(x)} |D^2 u_3|^2 \le Cr^{-2} \int_{B_{\frac{r}{2}}(x)} |Du_3|^2. \tag{2.14}$$

In particular, for any $\theta \in (0, \frac{1}{4})$ and $q \in (1, \frac{4}{3})$,

$$(\theta r)^{\frac{2q}{2-q}-4} \int_{B_{\theta r}(x)} |Du_3|^{\frac{2q}{2-q}} \le C\theta^{\frac{2q}{2-q}} r^{\frac{2q}{2-q}-4} \int_{B_{\frac{r}{2}}(x)} |Du_3|^{\frac{2q}{2-q}}. \tag{2.15}$$

Putting (2.7), (2.13), (2.15) together, we obtain, for any $q \in (1, \frac{4}{3})$ and $\theta \in (0, \frac{1}{4})$,

$$\begin{aligned}(\theta r)^{\frac{2q}{2-q}-4} \int_{B_{\theta r}(x)} |Du|^{\frac{2q}{2-q}} \le (C\epsilon_0\theta^{\frac{2q}{2-q}-4} + C\theta^{\frac{2q}{2-q}}) r^{\frac{2q}{2-q}-4} \int_{B_r(x)} |Du|^{\frac{2q}{2-q}} \\ + C\big(\theta, q, \|u\|_{W^{2,2}(\Omega)}\big) r^{\frac{2q}{2-q}}.\end{aligned} \tag{2.16}$$

Therefore, by choosing $\theta_0 = (4C)^{\frac{q-2}{2q}}$ and then choosing ϵ_0 sufficiently small, we have

$$\begin{aligned}(\theta_0 r)^{\frac{2q}{2-q}-4} \int_{B_{\theta_0 r}(x)} |Du|^{\frac{2q}{2-q}} \le \frac{1}{2} r^{\frac{2q}{2-q}-4} \int_{B_r(x)} |Du|^{\frac{2q}{2-q}} \\ + C\big(q, \|u\|_{W^{2,2}(\Omega)}\big) r^{\frac{2q}{2-q}}.\end{aligned} \tag{2.17}$$

Set $p = \frac{2q}{2-q}$. Observe that $p \in (2,4)$ for $q \in (1, \frac{4}{3})$. This completes the proof of Lemma A. □

Acknowledgement. I thank Professor S.Y. Alice Chang of the Princeton University Department of Mathematics for informing me of this problem and for helpful discussions on it.

References

[C] S.Y. Chang, *On a fourth-order partial differential equation in conformal geometry*, Essays in honor of Alberto P. Calderón. Papers from the International Conference held at the University of Chicago, Chicago, IL, February 1996, (Michael Christ, Carlos E. Kenig and Cora Sadosky, eds.), University of Chicago Press, Chicago, IL, 1999, 127–150, MR 1743859 (2001g:58059), Zbl 0982.53036.

[CGY] S.Y. Chang, M. Gursky and P. Yang, *Regularity of a fourth order nonlinear* PDE *with critical exponent*, Amer. J. Math., **121**(**2**) (1999), 215–257, MR 1680337 (2000b:49066), Zbl 0921.35032.

[CWY] S.Y. Chang, L. Wang and P. Yang, *A regularity theory of biharmonic maps*, Comm. Pure Appl. Math., **52**(**9**) (1999), 1113–1137, MR 1692148 (2000j:58025), Zbl 0953.58013.

[CY] S.Y. Chang and P. Yang, *Extremal metrics of zeta function determinants on* 4*-manifolds*, Ann. of Math., **142** (1995), 171–212, MR 1338677 (96e:58034), Zbl 0842.58011.

[SU] R. Schoen and K. Uhlenbeck, *A regularity theory for harmonic maps*, J. Differential Geom., **17** (1982), 307–335, MR 0664498 (84b:58037a), Zbl 0521.58021.

[UV] K. Uhlenbeck and J. Viaclovsky, *Regularity of weak solutions to critical exponent variational equations*, Math. Res. Lett., **7**(**5-6**) (2000), 651–656, MR 1809291 (2001j:35059), Zbl 0977.58020.

Received June 20, 2002

Department of Mathematics
University of Kentucky
Lexington KY 40506
E-mail address: cywang@ms.uky.edu

The author was partially supported by NSF grant DMS 9970549.

CONTENTS

Volume 216, no. 1 and no. 2

Wayne **Aitken**, Michael D. Fried and Linda M. Holt: *Davenport pairs over finite fields* 1

Al **Boggess** and Roman Dwilewicz: *CR approximation on a nonrigid hypersurface graph in* $\mathbf{C}^n$ 201

M.A. **Chebotar**, Wen-Fong Ke and Pjek-Hwee Lee: *Maps characterized by action on zero products* 217

Roman **Dwilewicz** with Al Boggess 201

B. **El Wahbi**, M. Rachidi and E.H. Zerouali: *Recursive relations, Jacobi matrices, moment problems and continued fractions* 39

Thomas J. **Enright** and Nolan R. Wallach: *A Pieri rule for Hermitian symmetric pairs II* 51

Xiaoming **Fan**: *Random attractor for a damped sine-Gordon equation with white noise* 63

Michael D. **Fried** with Wayne Aitken and Linda M. Holt 1

Joel **Friedman** and Jean-Pierre Tillich: *Wave equations for graphs and the edge-based Laplacian* 229

Bo **Guan** and Huai-Yu Jian: *The Monge–Ampère equation with infinite boundary value* 77

Paul Alton **Hagelstein**: *Rearrangements and the local integrablility of maximal functions* 111

Paul Alton **Hagelstein**: *Rearrangements and the local integrablility of maximal functions* 111

Kentaro **Hamachi**: *Differentiability of quantum moment maps and G-invariant star products* 127

Linda M. **Holt** with Wayne Aitken and Michael D. Fried 1

Huai-Yu **Jian** with Bo Guan 77

Elias G. **Katsoulis**: *Geometry of the unit ball and representation theory for operator algebras* 267

Wen-Fong **Ke** with M.A. Chebotar and Pjek-Hwee Lee 217

Masatoshi **Kokubu**, Masaaki Umehara and Kotaro Yamada: *Flat fronts in hyperbolic 3-space* 149

Tsiu-Kwen **Lee**: *Generalized skew derivations characterized by acting on zero products* 293

Tsiu-Kwen **Lee**: *Generalized skew derivations characterized by acting on zero products* 293

Stephen R. **McDowall**: *An inverse problem for the transport equation in the presence of a Riemannian metric* 303

C.A. **Morales** and M.J. Pacifico: *Sufficient conditions for robustness of attractors* 327

M.J. **Pacifico** with C.A. Morales 327

Anand **Pillay**: *Algebraic D-groups and differential Galois theory* 343

M. **Rachidi** with B. El Wahbi and E.H. Zerouali 39

Giovanni **Stegel**: *Inductive algebras for trees* 177

Bernd O. **Stratmann** and Mariusz Urbański: *Metrical diophantine analysis for tame parabolic iterated function systems* 361

Jean-Pierre **Tillich** with Joel Friedman 229

Masaaki **Umehara** with Masatoshi Kokubu and Kotaro Yamada 149

Mariusz **Urbański** with Bernd O. Stratmann 361

Nolan R. **Wallach** with Thomas J. Enright 51

Changyou **Wang**: *A note on a fourth order PDE with critical nonlinearity* 393

Kotaro **Yamada** with Masatoshi Kokubu and Masaaki Umehara 149

E.H. **Zerouali** with B. El Wahbi and M. Rachidi 39

Guidelines for Authors

Authors may submit manuscripts at pjm.math.berkeley.edu/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use LaTeX, but papers in other varieties of TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PACIFIC JOURNAL OF MATHEMATICS

Volume 216 No. 2 October 2004

CR approximation on a nonrigid hypersurface graph in $\mathbf{C}^n$ 201
AL BOGGESS AND ROMAN DWILEWICZ

Maps characterized by action on zero products 217
M.A. CHEBOTAR, WEN-FONG KE AND PJEK-HWEE LEE

Wave equations for graphs and the edge-based Laplacian 229
JOEL FRIEDMAN AND JEAN-PIERRE TILLICH

Geometry of the unit ball and representation theory for operator algebras 267
ELIAS G. KATSOULIS

Generalized skew derivations characterized by acting on zero products 293
TSIU-KWEN LEE

An inverse problem for the transport equation in the presence of a Riemannian metric 303
STEPHEN R. MCDOWALL

Sufficient conditions for robustness of attractors 327
C.A. MORALES AND M.J. PACIFICO

Algebraic D-groups and differential Galois theory 343
ANAND PILLAY

Metrical diophantine analysis for tame parabolic iterated function systems 361
BERND O. STRATMANN AND MARIUSZ URBAŃSKI

A note on a fourth order PDE with critical nonlinearity 393
CHANGYOU WANG

0030-8730(200410)216:2;1-M